J. theor. Biol. (1994) **171**, 29–39

Will Neuroscience Explain Consciousness?

GERMUND HESSLOW

Department of Physiology and Biophysics, University of Lund, Sölvegatan 19, *S*-223 62 *Lund, Sweden*

(*Received on* 1 *November* 1993, *Accepted in revised form on* 20 *April* 1994)

This paper is a defence of a pragmatic version of mind–brain reductionism from a neuroscientist's point of view. It is claimed that there are good reasons to believe that future neuroscience will be able to explain (in a weak and pragmatic sense) the puzzling aspects of mind and consciousness. Opposition to reductionism comes from both philosophical and empirical quarters. It is argued here that philosophical arguments, such as semantic problems with the concept of identity, are unconvincing and should be regarded with the greatest suspicion. The most influential empirical result that has been claimed to constitute a problem for reductionism is the temporal delay and mental antedating of consciousness found by Benjamin Libet. It is argued that these results, far from being a problem for reductionism, constitute evidence for a particular view of the physiological origins of consciousness. Finally, it is argued that many subjective aspects of experience can already be given satisfactory scientific explanations and that scientific progress is likely to rob the mind and subjective experience of their mystery.

1. Introduction

Many neuroscientists work on details of brain function and are not much concerned with the question of whether neuroscience will explain consciousness, but among those who have thought about the issue, the majority are reductionists: that is, they believe that neuroscience (physiology, anatomy, chemistry, etc) will eventually explain consciousness. This probably reflects historical experience. The reductionist strategy in biomedical research has been so spectacularly successful, even in areas in which philosophers and theologians had declared it to be impossible, that continued success is taken for granted. To put it simply and bluntly, microscopes and biochemical assays have created modern medicine, while various "holistic" approaches have only resulted in quackery and obscurantism. This view may be unfair, but I think that it is how medical history appears to most medical scientists. If one looks at the truly enormous growth of knowledge in biochemistry, cell biology and microbiology and at our dramatically improved ability to alleviate human suffering, one will need very good reasons and a very strong confidence to question the underlying scientific strategy.

Philosophical considerations, on the other hand, have only had a minor role in shaping the attitudes of scientists, and although some of them would be delighted to read a philosophical defence of reductionism, most regard philosophical arguments with suspicion and consider them to be essentially irrelevant. Watson, Crick and their successors never wasted any time to rebut vitalistic objections to the idea that the basic processes of life might be explicable in molecular terms. When this did occur, as in Jacques Monod's defence of reductionism in *Chance and Necessity* (1971), it received a lot of attention from theologians, philosophers and members of the general public, but most scientists regarded Monod's argument as superfluous.

Even among those neuroscientists who are less convinced about the possibility of explaining the mind in physiological terms, there is probably a greater willingness to consider empirical arguments than philosophical ones. "This is not a question that can be answered at the desk—let us get on with the

0022–5193/94/210029 + 11 $08.00/0

experimental work and time will tell who is right", is the dominating attitude. Naturally, there are exceptions. One of the greatest neuroscientists of this century, Sir John Eccles, has always been an ardent dualist and antireductionist. But Eccles' writings about the mind–body problem have been much more influential outside neuroscience than among his colleagues.

To summarize, one could say that neuroscientists hold a pragmatic and rather unsophisticated form of reductionism: we cannot decide on philosophical grounds whether neuroscience will ever explain consciousness; this is an empirical question that can only be resolved by trying; historical experience from other areas of biology, however, is ground for optimism.

In this paper I will try to defend this attitude. Although I have certain reservations about some particular forms of reductionism, I believe on the whole that the sceptical attitude towards philosophy is well founded and that the empirical arguments which have been advanced against reductionism are untenable.

2. Pragmatic Reductionism

Let me first make a few comments on the meaning of "reductionism". This term can be interpreted in many ways and has been given different definitions by philosophers (see e.g. Churchland, 1982). Basically, it means that psychological facts will eventually be explained by facts about the nervous system, but the meaning of this doctrine will depend on what one means by "explanation" and also on how one decides which facts should be called "psychological" and which "neuroscientific".

It would be easy to discredit reductionism simply by formulating it in unreasonably strong terms. For instance, if one believes, as many philosophers do, that *A* "explains" *B* only if *B* can be logically derived from *A*, reductionism will require that *every* true psychological statement, without exception, will be logically derivable from a set of neuroscientific statements in combination with statements, that bridge the gap between mind and brain by identifying mental events with physiological ones. In my view, this is a far too strong and quite unnecessary requirement of explanations and there are several theories of explanation which do not make the strong demand of logical derivability (see e.g. Hesslow, 1983).

On the other hand, if one thinks that the mind is "really" only an aspect of the brain, one could argue that psychological facts are "really" neuroscientific facts couched in psychological terminology. This would make reductionism trivially true. I will assume here that there are different groups of facts, some representing our daily experience of mental phenomena and another group containing the findings of neuroscience.

Tying an argument to one specific definition of "reduction" can lead to much unproductive hairsplitting. I will therefore here use the term in a rather unsophisticated sense. I do not wish to claim that *every* possible statement that will ever be made about the mind will get an exact and faithful translation into neuroscientific terms. There will always be exceptions, and it is not even likely that this will ever seem a worthwhile goal of science. What I do believe, is that science will dispel so much of the veils and the mystique surrounding consciousness and other mental phenomena, that these will appear as unproblematic as the nature of "life" now appears to biology.

It is not being suggested, of course, that we understand all the processes of life. But it is fair to say that nothing remains (among biologists) of the mystique that previously surrounded life and led to the belief that this was an area that must for ever remain out of reach for natural science or that we must assume the existence of a special "vital force" to explain the difference between living and dead matter. Analogously, we may never be able to answer *every* question that will ever be asked about the mind, but the impenetrable mystery, which has led to the belief in a special mental substance, for ever outside the domain of science, will disappear.

There are forms of reductionist strategy in neuroscience that may be dangerous. For instance, there is a certain naïve type of reductionism which holds that scientific advance only lies in studying ever smaller constituents of nature. There is a real risk that the present focus on molecular biology will lead to a neglect of "system properties" which are essential for understanding how the brain works. This kind of bias is not a necessary part of the kind of reductionism defended here.

Another question is what practical value a reduction might have. If one wanted to explain, say, why a chess playing computer moved a pawn, it might be a reasonable answer that it did so in order to protect the queen. It is possible that a correct answer could be given in terms of the computer's electronic circuitry, or in a "low level" programming language, but such answers would appear pointless to most of us. It would be an answer at the wrong "level". An interesting answer would have to refer to the purpose of the particular move and this will involve certain "high level" features of the program. There are probably many mental phenomena which it would be pointless to "reduce" to physiology, even if it was

possible. For the same reason, many psychological terms, although maybe redefined and integrated with neuroscientific concepts, will maintain a central role in explaining human behaviour.

To summarize, I do now wish to claim that every possible mental statement can or should be translated into a logically equivalent neurobiological statement. The pragmatic reductionism I want to defend is that all the important questions about mental phenomena will eventually receive scientific answers, that is more or less complete scientific explanations, and that most of these answers and explanations will involve neuroscience. Consciousness is not outside the reach of natural science; it will be understood when we understand the brain.

3. Philosophical Antireductionism

The philosophical literature purporting to refute reductionism is vast and it is impossible to deal with more than a fraction of it here (for reviews see e.g. Churchland, 1982; Dennett, 1978; or Churchland, 1988). Nevertheless, it may be possible to defend a sceptical attitude towards this literature by discussing an example. I have chosen to illustrate my scepticism with philosophical refutations of the identity theory.

It is usually held that a prerequisite for explaining consciousness by (or reducing it to) neuroscience is that some version of the identity theory is true. This means that it must be possible to defend statements of the kind:

$$M \text{ is identical to } N,$$

where M is some mental state and N is a neurophysiological state of the brain. The word "state" should here be taken in a very wide sense. Mental states include, for instance, thoughts, emotions, experiences and beliefs, and neurophysiological states include anatomical facts as well as chemical and electrical states of individual nerve cells.

If such identity statements were true, the psychological law "M_1 causes M_2", for instance, could be derived from the physiological law "N_1 causes N_2"together with the identity statements "$M_1 = N_1$" and "$M_2 = N_2$".

But such identity statements cannot possibly be true, it is claimed by antireductionist philosophers, because it contradicts a basic logical principle called Leibniz' law. This principle says that a is identical to b, if and only if everything that can be truthfully said of a can also be said of b and conversely. Another way of putting it is to say that for a and b to be identical, they must have the same properties. But, according to the antireductionists, mental and physiological states cannot have the same properties, and cannot therefore be identical.

Suppose, for instance, that I now have a thought which is profound, obscene or comical. According to the identity theory, it then follows that there is a neuronal state in my brain which is profound, obscene or comical. But this is absurd. A certain state in a collection of nerve cells cannot be comical any more than a thought can be hyperpolarized. It follows that the identity theory is absurd and consequently, it is impossible to reduce consciousness to neuroscience.

My response to arguments of this type is that they should be regarded as *reductio ad absurdum* arguments. Since they purport to prove something which clearly cannot be proved in this way, it follows that they rest on erroneous premises. If the reader finds this reply too flippant, I would remind him that I am in extremely good company, namely with Kant, who rejected Anselm's "ontological" proof of the existence of God in essentially this way.

This famous argument goes roughly as follows. God must be understood as the most perfect being that we can imagine. Now, the concept of perfection must include existence. For if God did not exist, we would be able to imagine something more perfect, namely a God which, in addition to all his other perfections, also existed. God must exist, because the thought of a God who does not exist is self-contradictory. Kant rejected this argument by pointing out that we could similarly prove that there must be a perfect island, which is plainly absurd. But note that the absurdity is not in the conclusions, that there are perfect Gods or islands. Neither is it obvious where the semantical or logical error lies. Indeed, this is still a subject of debate among philosophers. Very few theologians or philosophers take the ontological proof seriously, and the reason is clearly that they reject the idea that semantical sophistry could ever prove conclusions of this sort.

The argument against the identity theory can be shown to lead to many similar absurdities. For instance, Leibniz's law can prove that a famous author never wrote a certain book. For "the author of *Alice in Wonderland*" is known to most children whereas Charles Lutwidge Dodgson is not. It follows that Dodgson cannot be identical to the author of *Alice* . . . Note that, when we reject this argument, it is not because we easily spot the logical error. But since we know that it is unreasonable, we are entitled to reject the principles on which it is based.

The lesson is that logical and semantical subtleties of the kind exemplified by the argument from Leibniz's law are unreliable; they are too difficult for

human reason to handle safely. It is therefore unwise to let them determine our opinions on important issues.

There are many other arguments against the identity theory. A famous one was formulated by the logician Saul Kripke (1980) and rests on his concept "rigid designator". A rigid designator is a term which denotes the same object in every "possible world". "The president of France", for instance, is not a rigid designator, because different possible worlds will have different presidents. "Francois Mitterand", however, is a rigid designator in Kripke's terminology. According to Kripke, an identity statement of the type "$a = b$" can be true only if a and b are both rigid designators. If an identity statement is true at all, it must therefore be true in all possible worlds, and thus, in logical terminology, necessarily true. Now expressions like "my present pain" and "my present neurophysiological state" are rigid designators. But identity statements like "my present pain is identical to my present neurophysiological state" are not necessarily true. They are therefore not true at all and the identity theory must accordingly be false.

Again, I will not try to pinpoint the precise error of this argument. I merely wish to point out that the sheer difficulty of Kripke's reasoning should be sufficient to distrust it. Remember that we are not discussing philosophy here, but the proper future direction of neuroscience.

The above reflections are not intended as a defence of the identity theory. The latter has been used here only as an example to illustrate how unreasonable it would be to base a scientific strategy on semantical arguments. Indeed, even if the identity theory were false, it would not necessarily have much bearing on reductionism. The identity theory is very strong and the reader may wonder why reductionists have made themselves vulnerable to attack by defending such a bold view. The reason, I think, lies in the popularity of the strong view of explanation mentioned above. If we believe that mental facts can be explained by neural facts and we then let ourselves be convinced that explanations require derivability, then we will be forced to require identity. Unless mental states are identical to neural states, it would be strictly impossible to derive statements about the former from statements about the latter. If we do not subscribe to this theory of explanation, there is really no need to worry about the identity theory in the first place.

4. Empirical Antireductionism

The opponents of reductionism sometimes also employ empirical arguments. For instance, they refer to results from modern neuroscience which are claimed to show that there are so vast and fundamental differences between what goes on in the mind and what goes on in the brain, that the gap cannot possibly be bridged and no form of psycho-neural identity can therefore be true. Another tactic is to refer to some difficult problem which science cannot now solve and then claim that it will never be able to do this.

4.1. THE TIMING OF CONSCIOUS EXPERIENCE: LIBET'S EXPERIMENTS

An example of arguments of the first kind (the second kind will be illustrated later) is based on the sequence of experiments which has been carried out by Benjamin Libet and his associates (Libet, 1973, 1978, 1981; Libet *et al.*, 1979) and which allegedly show that a conscious experience is delayed in time relative to its physiological correlate. This, it has been claimed, constitutes an empirical refutation or at least a serious difficulty for the identity theory. Let me give a brief (and of course much simplified) summary of these studies.

If an electrical stimulus, just strong enough to elicit a conscious experience, is applied to the skin of the hand, an electrical potential change can be recorded in a specific part of the cerebral cortex. The stimulus gives rise to nerve impulses in the skin which are transmitted to the spinal cord, the brain stem, thalamus and the primary sensory cortex in the brain. The whole sequence takes about 10–20 milliseconds. In the cerebral cortex, a cascade of activity is initiated in a large number of cells, which activate or inhibit other nerve cells, etc. This activity, which in large part is electrical in nature, can be recorded by electrodes placed on the skull. Such a sequence of electrical activity which is evoked by a brief peripheral stimulus is called an "evoked potential".

An evoked potential can occur even if the stimulus strength is too small to generate a conscious experience. However, the potential looks different when the strength is increased so that the subject becomes conscious of the stimulus. In particular, later components of the potential only occur with stronger stimuli. A conscious experience seems to occur only when the stimulus is strong enough to cause an evoked potential which lasts for about 0·5 seconds.

A very simple interpretation of this finding is that it is this later part of the potential which is involved in generating the conscious experience, and that the experience thus does not occur until about 0·5 seconds after the stimulus. This interpretation may appear somewhat naïve and improbable. Half a second is a fairly long time in this context and we can start to give

a verbal report of the event much earlier than after 0·5 seconds. A more natural interpretation of the physiological facts is that the stronger stimulus causes both some late electrical brain activity *and* a conscious experience, but the former does not have to cause the latter. In his later experiments, however, Libet found evidence in support of the naïve interpretation. It actually seems as if a conscious experience does not arise until about half a second after the skin stimulus and almost as long after the first nerve impulses have reached the cerebral cortex.

In connection with brain surgery, it is sometimes possible to place stimulation electrodes directly on the exposed surface of the brain and to pass a current through them. Stimulation of that part of the brain's surface, known as the somatosensory cortex, can give rise to various sensations. If one stimulates the area which receives nerve impulses from the back of the hand, for instance, the subject may experience a sensation resembling a touch of the hand.

When Libet stimulated the cerebral cortex in this way, he found that it was usually necessary to apply a whole series of impulses (a stimulation train). Even when stimuli which were so weak that a single pulse was not noticed at all by the subject were used, a train of stimuli could still result in a conscious experience (usually a stimulation frequency of 60 impulses per second was used). Normally, a pulse train needs to be maintained for about 0·5 seconds in order to cause a conscious experience. Since stimulation for 0·4 seconds is not reported at all by the subject, the conclusion seems unavoidable that the conscious experience with this type of stimulation does not arise until about 0·5 seconds after the onset of stimulation.

Thus far, there is nothing really remarkable about Libet's findings. Presumably, consciousness requires a highly specific and complicated pattern of nerve cell activity in the brain, something that even a brief skin stimulus can give rise to. The effects of a crude and artificial direct stimulation of the cortex can obviously never come close to resembling the natural activity pattern, and it is not very surprising that the direct stimulation must be maintained for a while before the nerve cell activity is sufficient for generating a conscious experience. But this does not imply that consciousness under normal circumstances is delayed by half a second.

There is evidence for this, however. If the cortical stimulation is paired with skin stimulation, the latter can often be modified. In some cases, cortical stimulation can completely prevent the experience of a skin stimulus which would otherwise have occurred. Now, what is truly remarkable is that this can happen even if the cortical stimulation is started long—even up to 0·5 seconds—after the skin stimulus. Thus, if we stimulate the skin at time 0 and begin cortical stimulation 0·4 seconds later, the subject will not report any experience at all of the skin stimulus.

The most straightforward interpretation of this "retroactive masking", as Libet calls the phenomenon, is that consciousness did not arise during the 0·4 seconds. For if it had, the cortical stimulation, which had not yet started, could not interfere with experience. It is of course conceivable that a conscious experience had actually occurred earlier and that the cortical stimulus only prevented the formation of a *memory* of the event. Libet rejects this explanation on the ground that cortical stimulation in some cases actually enhances the experience of an earlier skin stimulus ("retroactive enhancement").

Libet's conclusion that the conscious experience arises 0·5 seconds after the skin stimulus depends on how the subject's experience is verified. Since the reaction time is considerably shorter than 0·5 seconds, one could imagine an experiment where the subject was instructed to, say, move his index finger at the instant when he experiences the stimulus. In that case, the latency of consciousness would probably be shorter than 0·5 seconds. Libet rejects this method, because it would enable the subject to react "reflexively" without being truly conscious. Instead he asks the subject after the experiment.

Libet has made another very important observation in this connection. Suppose that we start delivering a cortical stimulus train to the left side at time 0, and then give a skin stimulus to the left hand (from which the nerve impulses go to the right cortical hemisphere) after 0·2 seconds. One would, perhaps, expect that this would result in a conscious experience of the cortical stimulus at 0·5 seconds and of the skin stimulus at 0·7 seconds, since the delay of consciousness in both cases seems to be about 0·5 seconds (cf. Fig. 1). But this is not how it appears to the subject. If asked which stimulus came first, he will answer that the skin stimulus preceded the cortical stimulus. In spite of the fact that we have good reason to believe that the experience of the skin stimulus in reality occurs later than the experience of the cortical stimulation, the subjective view is that the skin stimulus came first.

To explain this, Libet suggests that the experience of the skin stimulation is "antedated" in consciousness, so that it becomes temporally correct. There is a "subjective referral" of the experience backwards in time (Libet, 1978; Libet *et al.*, 1979). This antedating does not work for the more artificial cortical stimulation which is therefore experienced as occurring

when it actually does. Although this conclusion has been challenged (e.g. in Churchland, 1981), there are many other experiments which support the antedating hypothesis (Libet, 1981), and it will be assumed here that it is essentially correct.

As illustrated in Fig. 1, just as the cortical stimulation is experienced half a second after the beginning of the pulse train: that is, at 0·5 seconds, the skin stimulus at 0·2 seconds will be experienced at 0·7 seconds. Since this latter is antedated, however, it will seem to be occurring half a second earlier, at 0·4 seconds, that is *before* the cortical stimulation.

To summarize Libet's results, it seems as if the conscious experience and the processes in the brain which give rise to it occur about half a second after the first nerve impulses from the skin have reached the cerebral cortex. The fact that we do not notice any such delay suggests that the brain "moves" the experience backwards in time, so that it will seem to us to be simultaneous with the stimulation.

Libet has claimed that these results demonstrate a temporal "dissociation" between mental and physiological events and that this causes difficulties for the identity theory. "On the face of it, an apparent lack of synchrony between the 'mental' and the 'physical' would appear to provide an experimentally based argument against 'identity theory' " (Libet, 1978: 80). This conclusion is far from self-evident, but Libet has here been supported by other antireductionists like Eccles (Popper & Eccles, 1977) and the physicist Roger Penrose (1989). The former has claimed that "This antedating procedure does not seem to be explicable by any neurophysiological process". Instead, the antedating must depend on "the self-conscious mind" which can "play tricks with time" (Popper & Eccles, 1977: 364).

4.2. COMMENTS ON LIBET

I want to make three comments on Libet's experiments and conclusions. The first one is that the alleged difficulty for the identity theory rests on an extremely implausible idea, namely that the antedating of an experience to 0·5 seconds before the neuronal correlate occurs, must mean that the *experience itself* has been moved backwards in time. Had this been the case, it would obviously have been a difficulty not only for the identity theory but for any theory of consciousness, because it seems to entail backwards causation. However, it seems far more natural to assume that the antedating is something that occurs after the experience has been terminated. It is not the experience itself that has been moved. Rather, it is the *retrospective judgement* of when the experience occurred that has been influenced. Libet is aware that this interpretation is possible but rejects it on rather loose grounds (Libet, 1978: 80).

The second comment is that Libet's results, although scientifically very important, are not nearly as remarkable as they are often made out to be. The existence of a mechanism which arranges the temporal relationships between various incoming stimuli is obvious from many other well-known facts. For instance, a stimulus can reach the cerebral cortex via routes with different conduction velocities. If I prick my finger with a needle, I will activate both pain and touch receptors in the skin. The information about touch reaches the cortex after some tens of milliseconds, while the pain impulses take a different, slower, route and reach the cortex much later. The difference is of the order of a second, but we do not usually experience two stimuli. We may also have a visual impression of the needle prick which will reach the cortex at a third point in time. In other cases the

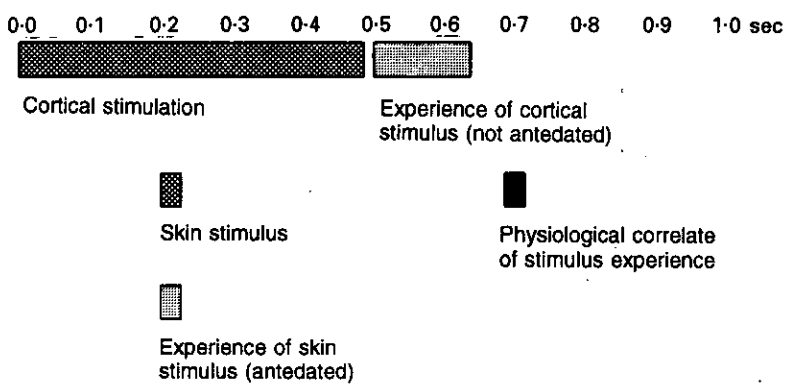

FIG. 1. Temporal relationships in Libet's experiments. Train stimulation of the left cerebral cortex begins at time 0 and results in a conscious experience of 0·5 sec. A stimulus to the skin of the left hand is delivered at 0·2 sec. Because of the delay of consciousness, this results in a physiological correlate of the experience at 0·7 sec. Contrary to expectation, the subject does not experience the skin stimulus as coming after the cortical stimulus. The skin stimulation is "antedated" in consciousness to the correct time. This process does not work for the artificial cortical stimulus which therefore appears to come after the skin stimulus.

cortex can get information about the same event at different times via vision and hearing, but we are usually able to interpret the impressions as being of the same event. Libet's antedating may well be a special case of this quite general phenomenon.

My third and final comment is that Libet's work, far from supporting antireductionism, is an excellent demonstration of how empirical research can throw light on consciousness. It is possible that Libet's two central hypotheses (interpreted with the above mentioned reservations), (i) conscious experience is delayed by about 0·5 seconds and (ii) the brain corrects this temporal delay, will not stand up to further testing, but at present it must be admitted that they are well supported and that they are of a kind that bridges the gap between physiology and consciousness, that is, precisely the kind of findings a reductionist expects.

5. Sketch of an Explanation of Consciousness

In order to justify further my optimism about the outlook for scientific attempts to understand consciousness, I would like to show, with the help of a couple of examples, what I think future scientific explanations of consciousness may look like. Note that I am not claiming that we can provide such explanations now. My object is rather to give a indication of what we can expect. The first example, which is rather speculative, concerns the origin and function of consciousness. The second example is grounded in solid neuroscience and concerns subjective aspects of experience.

5.1. THE ORIGIN OF CONSCIOUSNESS

There is an aspect of Libet's experiments which both he and those who cite him in support of antireductionism seem to have overlooked, and that is that his results tend to suggest that consciousness may be considerably less important than it is usually made out to be.

Most of us carry with us a naïve view of consciousness as a *mediator* between stimulus and response. If I burn a finger and then withdraw the hand, I will almost always say that I withdrew the hand *because* of the pain, and "the pain" usually refers to the conscious experience. It is obvious, however, that a withdrawal reflex does not require any consciousness. It is mediated by the spinal cord and works in roughly the same way even if the brain is disconnected. The idea that consciousness mediates stimulus and response is an illusion, but a very powerful one and it takes some effort to get used to the true facts.

Could it be that the belief that reactions to stimuli must be mediated by consciousness is just as illusory when applied to "higher" cognitive, functions? We like to imagine that our actions, i.e. voluntary actions in contrast to reflexes, spring from some kind of conscious deliberation or that they are formed by thought processes. For instance, when we carry on a conversation, we understand our partner's words with our consciousness, whereafter we, consciously, construct our reply. But if Libet is right, so that consciousness is always delayed by about 0·5 seconds, this cannot be correct. We would have been very much slower than we actually are. Our replies during the conversation must therefore be formed before we are conscious of the utterance to which we are replying.

Introspection confirms this conclusion. For instance, when someone puts a question to us and we deliver a rapid answer, it is striking how inaccessible the process is by which the answer is formed. A sentence appears "ready made" so to speak in our mind. We can notice it or reflect on it, but we have no knowledge whatsoever about how it actually arose. When we have an idea, it is often characteristic that it surprises us and how completely hidden the process is that created it.

But if consciousness is not needed to create thoughts, if it does not do anything but take note of those thoughts which arise out of a hidden process, it may seem puzzling that we would need consciousness at all. If we can respond and carry on a conversation without participation of consciousness, why has it evolved?

I think that the answer is roughly as follows. Let us assume a simple brain without consciousness, a brain which lacks an "inner world" and in which no conscious activities such as thinking, deliberation, planning, etc, occur. Those structures which generate ideas and verbal sentences always require an external stimulus before doing anything. This brain can answer a question, carry out an instruction, describe a sense impression or visualize a verbal description, but it does so only if prompted by a question, an instruction, a sense impression or a description. Now such a simple brain does not permit anything resembling "thinking", because it can only generate one response at a time. There can be no train of thought, no awareness of current thoughts, nothing like that sequence of ideas or inner verbal responses, which is what we normally associate with thinking or with an inner world of consciousness.

Nevertheless, there is a way in which more advanced results than the single responses just described could arise in this simple brain, namely in

a conversation with another, equally simple brain. Imagine a conversation between two simple-minded (i.e. simple-brained) individuals, where a person's utterance generates a response from the other person, which in turn generates a new response from the first speaker, and so on [cf. Fig. 2(a)]. Although this may be an unusual way of looking at it, I submit that such a conversation without any consciousness involved is entirely possible. Recall again that consciousness is *not* involved when we generate a sentence in response to a question. There is nothing outlandish about this scene.

But if this is possible, then a single person should be able to "simulate" this process alone, simply by listening to his own utterances and responding to them as if they had come from someone else [cf. Fig. 2(b)].

There are several limitations in this kind of "thinking". The process is time consuming, it cannot be kept secret (which is much more important than it might seem at first glance) and it cannot be applied to thoughts with non-verbal components. One way of solving these problems would be to equip the brain with the ability to record and react to its own thoughts, that is to its utterances, before "letting them out" in the form of speech. Output from the motor areas of the language centre of the brain could be fed directly into the sensory language area without taking the unnecessary route via speech and hearing [Fig. 2(c)]. A thought/utterance could then serve as a stimulus for a new thought, which is a stimulus for the next one, etc.

When we react to external stimuli, such as visual impressions or other people's speech, we do so relatively automatically and do not notice much of our consciousness. The latter does not play any great role in this kind of situation which is confirmed by Libet's findings. The experience of being conscious arises when our behaviour is elicited by internally generated stimuli. When I say something silently to myself and respond to this by saying something else, etc, it may seem natural to interpret the sequence of thoughts as components of an "inner world", especially since at each step, I can describe the previous step, that is I can think of myself thinking.

An important part of this is probably our ability to simulate sense impressions. If part of the brain can activate another part which normally receives or

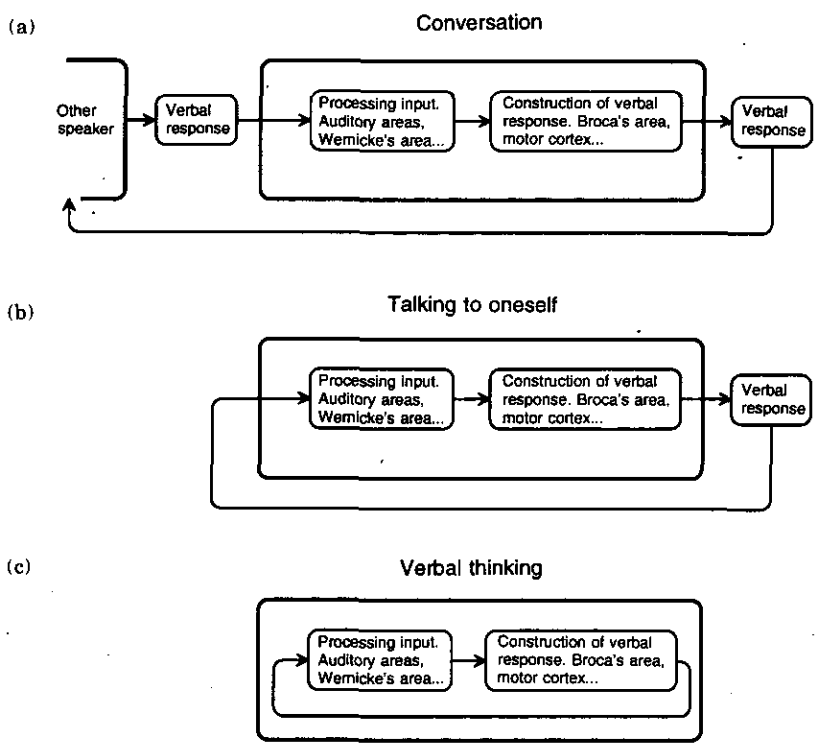

FIG. 2. The origin of verbal thinking. In (a) (conversation) one person makes verbal responses to the statements of another person, who then responds to the first, etc. In (b) (talking to oneself), a person listens to his own statements and responds to them, as if they were made by someone else. In (c) (verbal thinking), instead of being loud and audible, the speaker's statements are so weak that they do not cause the mouth to move, but the neural output from the brain's speech areas are fed into the areas normally receiving audible speech and cause perceptions which are similar to those normally caused by other speakers. They therefore create the impression that there is speech "in the head" (i.e. in consciousness) to which the person can respond with new weak statements. The process is basically the same as talking to oneself, except that it does not utilize sound and hearing but takes a short cut within the brain.

processes sensory information, such as the visual cortex or visual association areas, we should be able to "see" even if there is nothing there outside us which causes the seeing. This simulation is usually not very realistic, and it does not have to be in order to serve a function, but it can be sufficiently real to contribute to our experience of an inner world. Under certain circumstances, however, these impressions can be very realistic. For instance, one of the symptoms of schizophrenia is that the patient cannot distinguish between stimuli with an external origin and internally generated stimuli. Another example is drug-induced hallucinations.

It should perhaps be added that this is purely speculative. There is no empirical evidence for the suggestion that visualization or imaging occurs in this way, but it is an attractive hypothesis because it would explain these phenomena and the main mechanisms it requires, other than those involved in normal seeing, hearing and feeling, are certain anatomical pathways, for instance from a memory area to the visual association areas.

When I become conscious of a stimulus to my skin, it is tempting to say that the experience at some definite point in time "reaches" consciousness, as if the nerve impulses, after having travelled up to the brain, suddenly fall into a container called consciousness. A more accurate description, admittedly also crude, would be to say that the skin stimulus can give rise to many reactions without involvement of consciousness, but sometimes these reactions can also elicit other reactions, for instance a verbal description of the first reaction, which can itself be described verbally, and it is at this moment that we consider ourselves conscious of the stimulation.

These reflections do not explain anything about the underlying physiological processes in the brain, how a thought or covert verbal response is formed in response to another thought, but it does explain why thinking in several steps should be accompanied by the experience of an "inner world". There is no difference in principle between "unconsciously" replying to an utterance by another person and responding to an utterance which has arisen in my own brain and which I can "hear" already before it has become audible speech. In the latter case, where a long sequence of thoughts can occur without any interaction with the external world, and where such a sequence can elicit a thought about an earlier part of the sequence, the idea of an inner world becomes very natural. If to this we add that the brain, as a step in a sequence of thoughts, can simulate sensory input, perhaps by activating those parts of the brain which normally receive and process signals from the sense organs, the inner world becomes unavoidable. I would like to suggest that it would be fairly simple to create an inner world, and thus a kind rudimentary consciousness, in a computer in this way.

Maybe we here have an embryo to an explanation of the time delay of conscious experience. This delay appears puzzling, because we are misled by erroneous metaphors to regard consciousness as a sort of passive "container" of various experiences. We say, for instance, that a sensory impression "reaches" consciousness, as if it was some kind of final destination which every impression must sooner or later reach. There is, of course, no reason why it would have to take time for the impulse from the hand in Libet's experiment, once it has reached the brain, to also enter the final station, consciousness. But if consciousness consists in the brain's ability to respond to the response caused by the original stimulus, the time delay would appear to be quite natural. What makes us conscious of the skin stimulus is the fact that it gives rise to a thought about itself, for instance the overt or covert statement that there was a pricking of the skin and the statement that I just noticed this pricking of the skin by responding to it. It is the process that produces this statement, a process which is completely hidden to us and which does not itself involve any consciousness, which takes time.

The purpose of the previous reflections has not been to advance a physiological theory of consciousness, they are still too sketchy, but to point out a possible future direction for the kind of empirical research on consciousness which Libet has begun and which I think has a good chance of eventually explaining where, how and why consciousness arises in the human brain.

5.2. HOW TO EXPLAIN QUALITIES OF EXPERIENCE

When I look at the summer sky, I notice a number of things about the reality outside my own body. The sky is deep blue and a number of small white clouds slowly pass by. One could imagine a physiological explanation of this phenomenon which among other things includes an increased activity in neurons which signal blueness. But, the critics of reductionism object, my experience of the summer sky cannot be identical to this activity. There is a private experiential quality, *how the sky appears to me*, the serenity, peacefulness and beauty, for instance, that science will never be able to explain. An even more difficult challenge for science was expounded by the philosopher Thomas Nagel in an essay entitled "What is it like to be a bat?" (Nagel 1974). Will neuroscience ever

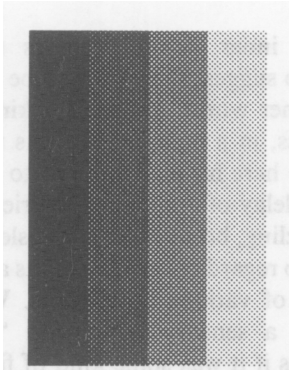

FIG. 3. Contrast enhancement. Although each grey area in the figure is of uniform luminance, the lighter areas appear to have light bands, and the darker areas dark bands, at the boundaries.

be able to explain what it is like, *for the bat*, to be a bat? Not surprisingly, Nagel's answer is no.

A problem with this type of objection is that it so vaguely identifies the problem. Precisely what is it that should be explained here and that science cannot handle? Those who put such general and all-encompassing questions to science (or for that matter to religion or art) will never see them answered. Neither science nor anything else can explain "my experience" just as it will never explain "reality", "the universe" or "nature". These things are simply not the sort of phenomena that can have explanations.

But if we lower our ambitions slightly and put questions about specific *aspects* of an experience, physiology can often provide quite enlightening answers. Let me give an example. If one looks at Fig. 3, one will see that the two darker fields are particularly dark just where they border on the lighter fields. These are in turn particularly light where they border on the darker fields. These thin bands on the borders (they are called "Mach bands" after the Austrian physicist and philosopher Ernst Mach) do not exist in reality. The fact that we "see" them is a consequence of a mechanism in the retina which is designed to sharpen the perception of contrast by exaggerating the difference between light and dark.

When light falls on the retina, a long sequence of nerve cells are activated. Firstly, the rods and cones are stimulated, but these initiate processes in other cells and it is only after a rather complicated sequence of events that the ganglionic cells are activated and send the information onto the brain. The ganglionic cells are subjected to two kinds of influence. They are stimulated by light which falls on the retina above them and they are inhibited by light which falls on the retina surrounding the cell. A simplified diagram illustrating this is shown in Fig. 4. The "+" signs illustrate how cells in a strongly illuminated area to the left and cells in a weakly illuminated area to the right are stimulated in different degrees. At the same time, the cells exert a mutual inhibition of each other ("−" signs). Notice that strong light in a certain site causes a stronger inhibition of the surrounding ganglion cells.

This arrangement, which is called lateral inhibition, has an interesting consequence. Notice the two cells in the middle, located on different sides of the border between light and dark. The left cell is subjected to the same amount of excitation as the two cells to the left, but because the cells to the right are more weakly illuminated is also subjected to weaker inhibition from the right. The net result is that it is more strongly activated than its neighbours to the left. The cell to the right of the midline is stimulated as strongly as the neighbours to the right. The cells at the far right are

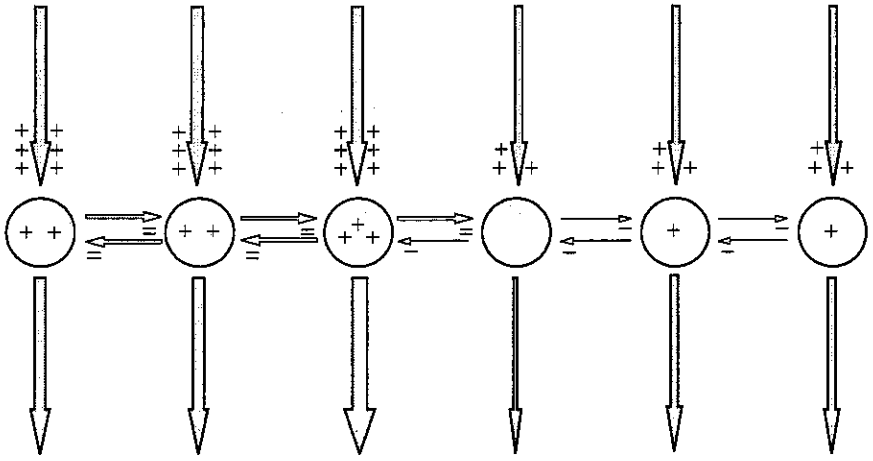

FIG. 4. Lateral inhibition in the retina. Strong (left) and weak (right) light produces strong and weak excitation, respectively, of the ganglion cells. Strongly and weakly excited cells produce strong (=) and weak (−) inhibition of neighbouring cells. Strongly excited cells which lie close to weakly excited cells will thus be more weakly inhibited and send out more intense signals. The brain will perceive the area as more strongly illuminated.

also weakly inhibited, because they are surrounded by weakly illuminated cells. But the cell closest to the mid-line has a strongly illuminated neighbour on one side and will therefore be strongly inhibited. It will thus be less activated than the neighbours to the right. The result is that the retina, at the border between light and dark, exaggerates the difference. The purpose is of course to sharpen contrasts and improve our ability to discern the borders between different objects.

We have not explained the "total" experience that someone has when looking at Fig. 3, but we have explained one important aspect of it. Many other aspects can be explained by similar considerations of what is known about the structure and function of the nervous system. The processing of visual signals in the retina is only the first step in a highly intricate series of successive processing which occurs on the way from the retina to the cortex and from one part of the cortex to another. In this case too, I think that it may be instructive to make a comparison with the concept of life. Science cannot answer the big, all-encompassing question: What is life? But as we have noticed before, this is not a serious limitation. We can explain so many aspects of what we call life that there is no longer any great mystery. Even if many questions remain, they are specific questions of cell biology, mainly molecular genetics. There is no longer any great puzzle of life. In a similar vein, I do not think that we will ever have a single explanation of the total subjective experience, but when a sufficient number of partial questions about specific aspects of experience are answered, this will seem sufficient.

There are probably aspects of the self, of consciousness and of our private experiences which are better described by poetry and art than by science. But this is a truth that can be applied to all parts of reality. As long as we stick to a reasonable level of ambition, there are no convincing reasons for neuroscientists to despair: eventually, the questions about mind and consciousness will receive scientific answers.

We should not forget, however, that there are many people who do not want a solution to the mystery and who wish to keep the enigma intact. I think that they will find, as many before them have found in other areas, that the worlds opened up by science are far more exciting and fascinating than the kind of mystery that results from ignorance.

Acknowledgements: This work was supported by grants from the Medical Faculty, University of Lund and from the Swedish Medical Research Council (project no. 09899).

REFERENCES

CHURCHLAND, P. M. (1988). *Matter and Consciousness*. Cambridge, MA: MIT Press.

CHURCHLAND, P. S. (1981). On the alleged backwards referral of experiences and its relevance to the mind–body problem. *Phil. Sci.* **48**, 165–181.

CHURCHLAND, P. S. (1982). Mind–brain reduction: new light from the philosophy of science. *Neuroscience* **7**, 1041–1047.

DENNETT, D. C. (1978). Current issues in the philosophy of mind. *Am. Phil. Q.* **15**, 249–261.

HESSLOW, G. (1983). Explaining differences and weighting causes. *Theoria* **49**, 87–111.

KRIPKE, S. A. (1980). *Naming and Necessity*. Oxford: Basil Blackwell.

LIBET, B. (1973). Electrical stimulation of the cortex in human subjects and conscious sensory aspects. In: *Handbook of Sensory Physiology*, Vol. 2 (Iggo, A., ed.) pp. 743–790. Berlin: Springer-Verlag.

LIBET, B. (1978). Neuronal vs. subjective timing for a conscious sensory experience. In: *Cerebral Correlates of Conscious Experience* (Buser, P. A. & Rogeul-Buser, A., eds) pp. 69–82. Amsterdam: Elsevier/North-Holland Biomedical Press.

LIBET, B. (1981). The experimental evidence for subjective referral of a sensory experience backwards in time: reply to P.S. Churchland. *Phil. Sci.* **48**, 182–197.

LIBET, B., WRIGHT, E. W. JR., FEINSTEIN, B. & PEARL, D. K. (1979). Subjective referral of the timing for a conscious sensory experience: a functional role for the somato-sensory specific projection system in man. *Brain* **102**, 191–222.

MONOD, J. (1971). *Chance and Necessity*. New York: Alfred Knopf.

NAGEL, T. (1974). What is it like to be a bat? *Phil. Rev.* **83**, 435–450.

PENROSE, R. (1989). *The Emperor's New Mind*. Oxford: Oxford University Press.

POPPER, K. R. & ECCLES, J. C. (1977). *The Self and Its Brain*. Berlin: Springer-Verlag.