MACHINES, BRAINS, AND PERSONS

by Donald M. MacKay

Abstract. This paper explores the suggestion that our conscious experience is embodied in, rather than interactive with, our brain activity, and that the distinctive brain correlate of conscious experience lies at the level of global functional organization. To speak of either brains or computers as thinking is categorically inept, but whether stochastic mechanisms using internal experimentation rather than rule-following to determine behavior could embody conscious agency is argued to be an open question, even in light of the Christian doctrine of man. Mechanistic brain science does nothing to discredit Christian experience in dialogue with God or the Christian hope of eternal life.

This paper is about persons in the sense in which you, the reader, and I are persons—members of the community of personal knowers, and potentially "meetables"—terminals of dialogue. Its purpose is to ask in what ways, if any, our concept of the human person is affected by current developments in brain science on the one hand and in artificial intelligence on the other. It falls into three unequal sections. In the first we consider the relation between what we know of ourselves in immediate experience and what we are beginning to learn from brain science about our physical embodiment. In the second we look at the growing capacities of machines to behave in ways we regard as intelligent and ask what limits, if any, can be set to such developments. Finally we ask what implications there may be in all this for theological—especially Christian—ideas of human personhood and human destiny.

PERSONS AND THEIR EMBODIMENT

Our knowledge of what it means to be a person has diverse sources. To begin with, we each have an ongoing story to tell of facts known to us in conscious experience, such as “I see-a-sheet-of-paper-before-me” or “I
feel-sorry-for-Joe" or "I believe-that-the-cat-is-outside." Let us call these collectively the "I-story." To use again a visual aid which I have found helpful, we might picture the I-story as a great list of statements in one column of a (very long) blackboard, each bearing witness to a fact about us that we would be lying to deny (MacKay 1962). (There may, of course, be some facts of our experience—inarticulate feelings, for example—which we do not know how to put into words; these obviously form part of the data to which the I-story bears witness, however imperfectly.)

Brain science is concerned to elaborate a very different story about us in terms of patterns of cell firings and the like. It makes the working assumption (and it is no more) that for each fact of our conscious experience some corresponding story—the "brain-story"—exists to be told in neural terms. There is no suggestion that the correspondence is necessarily one to one. Several alternative states of neural activity, for example, might be possible correlates of exactly the same experience of seeing-the-sheet-of-paper; but at least if I am having that experience, brain science assumes that something must be true about my nervous system which would not be the case if I were not having it.

The mysterious character of this correlation, for which there is slowly but steadily accumulating evidence (Buser & Rougeul-Buser 1978), must not be overlooked. To dramatize it, let us imagine (science-fiction-wise) that advances in technology could tomorrow allow a neurosurgeon to remove your brain from its casing and keep it alive and well in a portable container with a suitable extension cord from your body to maintain all its connections and supplies in functional order (MacKay 1975). Assuming that no significant time lags or extraneous signals were introduced, we may suppose (for the sake of argument) that you would continue to have a more or less normal flux of conscious experience.

Suppose however that the container with your brain at the end of its cord were placed in front of you, among a variety of other physical objects (perhaps including other brains) which you are free to explore with physical probes, electrical stimulators, and the like. You (sitting in your chair) can watch the ongoing results of your explorations on various instruments. As long as you explore the other objects, or indeed anything else in the physical universe, all goes well. But if once you start probing and stimulating your own brain tissue... ouch! You begin to have direct experience of bizarre events—seeing flashes of light, feeling sick or elated, thinking strange thoughts.

That, you may say, is not so surprising, seeing there is this cable running from the brain out there into my body. Suppose then that the relevant cable connections (though not, of course, the blood supply)
were temporarily removed or blocked. What would you now expect to happen if the stimulating current were applied to the same brain area? According to the working assumption whose consequences we are exploring, you would *continue to experience* the same bizarre events, provided that the same patterns of neural activity were elicited in that strange two handfuls of neural tissue out there in front of you. Furthermore, whatever ideas went through your mind (as we conventionally say) would continue to be reflected in the changing pattern of physical activity in the depths of that same small mass of tissue.

Why should the physical state of that little sample, out of all the vast contents of the physical universe, have this immediate relationship with the events of your conscious experience? This is not a question to which any conceivable advances in our understanding of the details or principles of brain activity can give an answer. In that sense it highlights not a mere scientific puzzle, but an ontological mystery. For each of us it points to a brute fact—as far as brain science is concerned an irreducibly given and mysterious fact—about our embodiment as persons, with which we must simply learn to live. Our imaginary thought-experiment also makes clear the sense in which it is meaningful (even operationally meaningful) to ask whereabouts in the brain the direct correlate of our conscious experiencing (a better term here than *consciousness*) is located: of which more anon.

**What is Special About Our Embodiment?**

It is natural to ask whether we can identify anything scientifically distinctive in the grey-pink mass before us as the correlate of its remarkable associations with our experience. Some would hypothesize (in the tradition of René Descartes) that it must contain a special region equipped to sustain two-way "interaction" with another "world"—the "world of the mind" (Eccles 1980). To date, however, no evidence has emerged to suggest that any part of the human brain differs in its physiological principles from other parts or indeed from the brains of lower animals. Nerve cells, whatever else they may be, are exquisitely sensitive and elaborate processors of information. They are normally linked in cooperative groupings which respond to specific features of incoming traffic, listening out for particular patterns of coincidence or covariation of signals in different nerve fibers. In this sense the brain could perhaps be described not as a computer, but as a vast community of linked microcomputers. The principles on which it is organized are quite different from those employed in the digital hardware of our conventional computing machines. A typical brain cell has a continuously variable probability of "firing" (or being activated) which depends not only on its metabolic state, but also on the precise relative
timing and layout of hundreds or even thousands of incoming signals that converge on its branching dendrites or the cell body. This means that the momentary state of an interacting group of nerve cells must be specified not just by saying which cells are firing, but by a vast matrix of ever-changing "conditional probabilities," for example, the probability that if A happens but not B, then C will happen, and so on (MacKay 1954; 1956; 1961; 1983a). Many extraneous factors, such as chemical messengers in the blood supply, may play their part in the modulation of these conditional probabilities and in the continual biophysical tug-of-war that determines (statistically) whether and when each cell will deliver an impulse to the ongoing chatter (Eccles 1980). There is nothing here of the quiet and disciplined sobriety of an array of digital computing elements, each awaiting its simple yes-or-no instruction to change its state from one to zero or vice versa!

Despite these many disanalogies, however, there is nothing in the known biophysics of the nervous system to give comfort to an anti-mechanist or to suggest any clue to its mysterious link with our conscious experience. There are plenty of promising correlates. For example, changes in the efficacy of connections between nerve cells (synapses), which can occur when both cells are simultaneously active, offer one plausible basis for the storage of information in learning and memory (Eccles 1980, chap. 7). Again, large groups of cells can show what is called cooperativity—massive changes of state which can propagate rather like a flame in an igniting gas mixture. These mechanisms further enrich our stock of possible correlates of psychological processes. None of them, however, answers the question why the brain, rather than the heart, for example, should have this baffling link with our immediate experience.

**Informational Organization**

When we look at the global organization of the brain as a whole, at the cybernetic level of systems analysis, some special distinguishing features do emerge. By studying the effects of brain injury or malfunction, neurologists have accumulated much evidence as to which brain areas are essential to which psychological functions (Penfield 1975). By analyzing the minimal information-processing requirements for intelligent purposeful agency (whether in automata or living organisms) the systems analyst can develop a skeleton information flow model which can be used (as a tool of research) to suggest possible mechanistic interpretations of the clinical data (MacKay 1954; 1956; 1961; 1983a). In its simplest form it can be thought of as a hierarchic structure: there is an organizing level concerned with the selection and deployment of various alternative courses of action under feedback from sensory
apparatus: this is programmed by a higher supervisory level that selects the goals to be pursued by the organizing system and also the relative priorities to be assigned to different possible goals or criteria of evaluation in case of goal conflict. A system organized on these lines can show behavior of the kind we would describe as intelligent and accept as evidence of consciousness in living animals. Furthermore, if the supervisory level were equipped to form and evaluate internal representations not only of its environment and its goal-priorities, but also of itself and its own evaluative process, its behavior could show many of the features we take as evidence of self-consciousness and self-evaluative autonomy (MacKay 1951; 1953a; 1966; 1980a; 1981).

My suggestion is not that an automaton on these lines must be considered self-conscious (see below), but rather (conversely) that the special feature of our brain which must be functioning if we are to enjoy conscious experience may well be its self-supervisory and self-evaluative information system. Clinical evidence shows that we can lose consciousness when relatively small portions of deep central brain structures are damaged, even while the rest of the brain remains capable of organizing quite complex sensorimotor coordination (Penfield 1966). Conversely, large areas of the cerebral cortex can be damaged or cut out altogether without rendering the patient unconscious (although leaving him deprived of corresponding sensory or motor functions). This has an obvious bearing on the twin practical problems of determining at what stage in fetal development personal life has begun and at what stage in bodily disintegration personal life has come to an end (MacKay 1984). It also bears on the vexed question as to whether surgical section of the corpus callosum (the millions of connecting fibers between left and right cortical areas) turns so-called split-brain patients into two separate conscious agents. Experimental evidence to date suggests that the two half-systems in these cases can certainly sustain independent and potentially conflicting goal-directed agency, and even a form of low-level dialogue between them; but it gives no indication that the topmost self-supervisory level of organization has been split (MacKay & MacKay 1982).

**What is Special About Meeting?**

From the standpoint of the information engineer, something quite special can happen when two information systems of the kind we are discussing (call them A and B) become cross-coupled in the process we call dialogue. To the extent that each seeks to form an internal representation of the goals and evaluative criteria of the other and opens itself reciprocally to the address of the other, the joint system (A plus B) so constituted can be logically-indeterminate for both of them (MacKay
This means, not that it violates physically-deterministic laws (if any), but that, even if all its constituent processes were physically determinate, no completely detailed specification of its present or immediately future state exists which could claim to be the inevitable truth for either party. Because of the reciprocal causal chain-mesh set up in true dialogue (as distinct from mere alternate monologue), any completely detailed specification of B’s information system must include details that reflect B’s image of A’s cognitive processes, and conversely. In general these details cannot (logically cannot) be independent of what A or B is thinking and feeling. Hence no such complete (future) specification can claim to be true irrespective of what A or B thinks of it. In that sense, their joint future has an ingredient that is irreducibly indeterminate-for-them until they jointly determine by their dialogue what form it shall take (MacKay 1978a).

**Role-Playing and Rule-Following**

Recognition of another person is not a matter of strict deduction from logically sufficient signs. It is a matter of commitment, demanded of us *prima facie* by the human appearance. Occasionally, as when we consult the policeman at the waxworks, we may be let down; but normally our commitment is justified by events.

There is, however, an instructive anomaly in the case of someone playing or acting a part. On the stage the real person Bill Bloggs may act the part of Macbeth. Watching him, we see and hear Macbeth, quite a different person. Are there then two persons on the stage? Clearly not, if by a person we mean a center of personal awareness. Only Bloggs is out there. We may indeed agree that the personality currently expressed is not Bloggs’s, but Macbeth’s, but this does not add to the number of personal centers of awareness in our contemporary world.

Consider next the case where Bloggs is a secret agent acting a role by way of disguise. The real Bloggs hates classical music and capitalists, say, and has a thriving family. His alter ego, Joe, is a bachelor who professes a love of the classics and of conservatism. Here we can ourselves meet Joe and form our impression of him as a “person.” Bloggs, his impersonator, does not have only memorized lines to recite, but can respond *ad lib* in character, as we say, so that we may easily be fooled. Joe’s personality is beyond dispute; but again it would be absurd to conclude on these grounds that Joe is a separate and additional member of the cognitive community of persons which already included Bloggs. Joe is a contrived personality, carefully contrived by the emission of personal behavior according to strict rules, but not a person to be counted alongside Bloggs among the living human population.
MACHINES AS ROLE-PLAYERS

Our brief excursion into the mechanics of personal behavior and its imitation by human actors brings out several points to keep in mind when assessing claims made for computing machinery and artificial intelligence. In the first place, to speak of a computer as "thinking" would be as semantically inept as to speak of a brain's doing so (MacKay 1962). It is not brains, but people, that think. Thinking is a concept belonging to the I-story. When I think, my brain doubtless goes through corresponding physical motions; but these motions are the correlates, or the embodiment, of my thinking.

Second, if it would be irrational to conclude that an additional conscious person has come into existence merely because a human being acts the part of Joe or Macbeth, it is not obviously more rational to do so when the part in question is acted by a machine. Indeed, if the acting is dictated wholly by rule, as in a typical artificial intelligence program, the parallel with human role-playing is so close as to make derisory any claims to personal status, conscious awareness, or the like on behalf of the artificial contrivance. Admittedly, like our human actor, the artificial intelligence artifact manifests a personality; but insofar as this results from mere rule-following, there are no rational grounds in either case for concluding that there is anyone there present as a conscious cognitive agent, whose personality it is. With due respect to A. M. Turing (1950), the experience of meeting the contrived personality, in the sense only of enjoying a satisfactory exchange of messages, no more proves the reality of the artificial person than the meeting enjoyed by children at the Christmas toy shop proves the reality of Santa Claus.

ARTIFICIAL CONSCIOUSNESS

By raising these valid objections to the nonsense sometimes talked about contemporary computing machinery, however, I do not suggest that we dispose of the possibility in principle of artificially begetting conscious persons. True, it would be inept to claim that any machine thinks; but it is far from nonsensical to ask whether a suitably constructed system might not embody the mental activity of a conscious agent in the same sense in which the information system of your brain or mine embodies our mental activity. Moreover, although role-playing by rule-following, whether by people or machines, does not generate new centers of conscious experience, we know of at least one genetically programmable biological process that does; so it is far from obvious that the use of rules to determine the construction (as distinct from the functioning) of an artifact need disqualify it ipso facto from embodying conscious agency.
The notion seems to be current that any machine must be equivalent in some sense to a set of rules for its behavior; but this is a mistake. At a branch point in the time course of a mechanical system, the outcome can be determined in one of two ways. The first, embodied in digital computing machinery, is to consult a rule. The second, embodied in old-fashioned analogue computing machinery such as the slide rule and also in the nervous system, is to make a physical experiment. In this case, although the selection of the experiment to make may be determined by rule, the outcome is not; it is the result of an appeal to the real world to which no finite set of rules is completely equivalent.

I have elsewhere argued that mechanisms suitably combining the principles of rule-following and physical experimentation can transcend the demonstrable limitations of purely digital symbol-manipulators and can in principle meet any specifiable test for mindlike behavior (MacKay 1951; 1965b; 1970; 1971b). If this is right, any attempt to find an explicit specification that no machine could meet is foredoomed to failure. On the other hand, any attempt to produce a complete and explicit specification of human behavior is equally foredoomed in principle, if one of the major determinants of human behavior is a process equivalent to physical experimentation. It is arguable that much of what Michael Polanyi called "tacit knowledge" may be embodied implicitly in the brain structures involved in such internal experimentation, in the sense in which the multiplication table is embodied implicitly in the structure of a slide rule (MacKay 1974).

In summary, while dismissing as nonsensical the claims to consciousness made on behalf of contemporary robots, I see no technical reason to deny (or to assert) that an artificially constructed information system could be the brain of a conscious person. In our present state of ignorance it seems to be a genuinely open question.

**Theological Implications**

It is not uncommon to find protagonists of artificial intelligence who see their art as a growing threat to religious, especially Christian, ideas of our human nature and destiny; and some Christians appear to share this perception. To those of us who cheerfully enjoy the excitement of the scientific chase with no such thoughts in our heads, this raises a serious question. Are we missing something insidious (from a religious standpoint) in this whole mechanistic approach, or does the trouble arise from some conceptual confusions shared by both sides in the debate?

There is, of course, something potentially insidious in any invitation to step out of our normal human situation and imagine ourselves as isolated spectators of the human phenomenon (MacKay 1955). By
cutting ourselves off, even in imagination, from dialogue with our fellows we can in principle depersonalize them and reduce them to mere objects of scrutiny, losing sight for the moment of their own status as centers of awareness and potential terminals of dialogue. But this is a mere perceptual illusion to be remedied, like all such illusions, by shifting our standpoint so as to make possible a more veridical experience of the situation temporarily misperceived (MacKay 1983b). There is no more need for a practitioner of artificial intelligence to remain under such illusions than for the person in charge of the weighing machine at a boxing match to see human beings only as ponderous masses.

More serious at first sight might be the objection that unless we rule out a priori the possibility of synthesizing the brain of a conscious person (a possibility that I have deliberately left open in principle, albeit with tongue in cheek), we run into conflict with Christian teaching that the creation of mankind is a divine prerogative. Here I think we can identify a straight confusion between creation—the divine prerogative—and procreation—for which God has licensed human beings. True, the natural process by which two people normally bring about the embodiment of a new person is one that they do not understand in any detail; but it would still be only a case of procreation even if they did, and even if they could call upon the resources of genetic engineering to determine in detail the characteristics of their offspring. Should it ever prove possible to beget conscious human beings by more artificial procedures, this would simply be a case of artificial procreation, a totally different concept from what theologians mean by divine creation (MacKay 1953b).

Another objection sometimes raised, by both believers and unbelievers, is that artificial intelligence threatens to dispel the mystery and wonder surrounding our human nature. Here again there would seem to be a confusion between two quite different notions. One, discussed in detail by Margaret Boden (1985), is the sort of wonder evoked by a clever performance, which is dissipated once we learn how the trick is done. Such wonder, which goes back basically to curiosity and puzzle- ment, is progressively eroded as science advances, though to be fair, we should note that each advance in science tends in its turn to throw up a crop of fresh puzzles.

The other notion could be termed awe, the recognition, deeper than words, of a mystery that must be accepted and responded to with, as Job puts it, our hand upon our mouth. Such a mystery, it seems to me, is presented by our experience of what it is like to be an embodied conscious agent and what it is like to be engaged in dialogue with another such agent. As I have argued above, this mystery stands wholly
untouched by any conceivable advance in either brain science or artificial intelligence, since the acceptance of it is the precondition of all our thinking. To see artificial intelligence as a threat to the sense of awe appropriate to the mystery of personal being would be to misconceive the nature of both. If artificial means were ever achieved to bring conscious personal beings into existence, then their existence (as distinct from that of their embodiment) would still present us with the same mystery, demanding the same awe in response.

Implicit in our analysis has been a particular answer to the question: What makes people special and why do we value other persons? Ultimately, I suggest, not for what they provide by way of resources, whether physical or informational, but because only other persons can engage with us in what Martin Buber (1937) called the I-Thou relationship of dialogue. It is characteristically in dialogue that we have our personal being, as participants in a situation which is for us undetermined until our participation determines its unfolding. It is in dialogue that we have the special experience, wholly without parallel in non-personal encounters with our world, of the union of two or more centers of awareness—the correlate of the mutual interpenetration of their information systems with the resulting formation of closed loops of logically indeterminate activity, at which we looked earlier.

This is not to suggest that we are persons only in dialogue with others. What seems to be distinctive about the human embodiment is our capacity for internal dialogue: the turning inward of skills elicited and developed in the course of dialogue with those who first nurtured our capacity for it. Talking to ourselves, whether internalized or not, is a characteristically personal activity in which our own situation is undetermined—that is, it has no completely determinate specification as to its present or future with an unconditional claim to our assent—until we, by our thinking, valuing, and deciding, determine its unfolding.

Equally compatible with our mechanistic analysis is the claim of Christian theism, that we enter most fully into our personal being when in dialogue with God, its Giver. “Thou hast made us for Thyself; and our hearts are restless till they find rest in Thee,” said Augustine. He was not, I think, referring to an alternative to the modes of being of our daily life, but rather to a complementary perception of the significance of that daily life as an ongoing dialogue with its divine Giver. If I am right, then no extrasensory channel of communication need be added to our mechanistic model in order to make room even for our knowledge of God; for it is in and through the exercise of our ordinary capacities in response to the challenges of daily events and in loving bondage to Christ that that knowledge is meant to grow and develop as the pathway to life eternal.
Here, however, we must face one final objection. Granted what has been said in relation to life here and now, do not such mechanistic theories of our embodiment dispose completely of the Christian hope of eternal life? I have explained elsewhere why I think this objection is misconceived (MacKay 1978b; 1980b, 100-2). Briefly, if the relation between our conscious experience and our brain activity is one of embodiment rather than one of quasi-physical interaction between two different worlds, then the destruction of our present embodiment would certainly imply the termination of our conscious experience in this space-time; but it would not at all rule out the possibility, if our Creator so willed it, that we should find ourselves reembodied, perhaps (as Paul hints in I Cor. 15) in some unimaginably different embodiment, "in the resurrection." Even an artificial intelligence, after all, could in principle be "restored to life" in a fresh embodiment if the computer in which it was originally programmed were destroyed. For the reasons outlined in earlier sections I am far from suggesting that mere replication of behavior patterns would suffice to guarantee ontological continuity between the original individual and his resurrected counterpart; but the daily illustration of ontological continuity between Joe Snooks who goes to bed one night and Joe Snooks who awakes next morning ought to reassure the Christian believer that the provision of such continuity is not beyond the powers of his Creator.

In summary, we have found a number of ways in which developments, both in brain science and in artificial intelligence, are forcing us to reexamine traditional assumptions about the capacities of mechanistic systems, including those of our own brains; but I have argued that the upshot only sympathetically illuminates, and in no way discredits, the biblical emphasis on our awesome nature and destiny.

REFERENCES


