7. The self

‘Each self is a divine creation.’

Sir John Eccles

‘My one regret in life is that I'm not someone else.’

Woody Allen

What are we? Each of us has buried deep within our consciousness a strong sense of personal identity. As we grow up and develop, our opinions and tastes change, our perspective of the world shifts, new emotions surface. Yet through it all we never doubt that we are the same person. *We* have those changing experiences. The experiences happen to *us*. But what is the ‘we’ that has the experiences? That is the long-standing mystery of the self.

When dealing with other people we usually identify them with their bodies, and to a lesser extent their personalities, but we view ourselves quite differently. When someone refers to ‘my body’ it is in the sense of a possession, as in ‘my house’. But when it comes to mind, that is not so much a possession as a *possessor*. My mind is not a chattel: it is *me*.

The mind is, then, regarded as the *owner* of experiences and feelings, the centre or focus of thoughts. My thoughts and my experiences belong to me; yours belong to you. In the words of the Scottish philosopher Thomas Reid:

> Whatever this self may be it is something which thinks, and deliberates, and resolves, and acts, and suffers. I am not thought, I am not action, I am not feeling; I am something that thinks, and acts and suffers.[1]

What more natural than for theologians to identify the self with the elusive mental substance or soul? Furthermore, as the soul is not located in space, it cannot be ‘pulled apart’ or disseminated, so the integrity of the self is assured. For it is one of the most fundamental properties of the perceived ‘self’ that it is indivisible and discrete. *I* am *one* individual, and *I* am quite distinct from *you*.

The concept of the mind (or soul), as we saw in the previous chapter, is, nevertheless, a notoriously difficult one and can involve paradox. The question ‘What *am* I?’ is not an easy one to answer. As Ryle points out: ‘Gratuitous mystification begins from the moment that we start to peer around for the beings named by our pronouns.’[2] Still, the question has to be answered if one is to make any sense at all of the idea of immortality. If I am to survive death just what is it that I can expect to survive?

According to David Hume, the self is nothing but a collection of experiences:

> When I enter most intimately into what I call *myself* I always stumble on some particular perception or other, of heat or cold, light or shade, love or hatred, pain or pleasure. I can never catch *myself* at any time without a perception, and never can observe anything but the perception.[3]

So adopting this philosophy, the answer to the question 'What am I?' is simply 'I am my thoughts and experiences'. Yet there is a feeling of unease about this. Can thoughts exist without a *thinker*? And what is there to distinguish *your* thoughts from *my* thoughts? What, indeed, does one mean by 'my' thoughts? In fact, Hume was later to write of his first assessment: 'Upon a more strict review of the section concerning *personal identity*: I find myself involved in a labyrinth.'

It has to be conceded, however, that the concept of self is nebulous, and that experiences go a long way to shaping the quality of the self, even if they do not explain it completely away. Some aspects of the self seem to lie on the borderline of personal identity. Where are we to locate (figuratively) emotions for example? Do you *have* emotions (as you have a body) or are your emotions an integral part of *you*? It is well known that emotions are strongly influenced by physical effects, such as the chemical composition of the blood. Hormone imbalances can produce various emotional disorders. Drugs can produce or depress a variety of mental states and emotional dispositions (as any consumer of alcohol knows). More drastically, brain surgery can produce major alterations of personality. All this makes us reluctant to clothe the soul with too many of the trappings of personality. On the other hand, ifall emotions are removed, what is left? Christians, for example, might accept shedding negative emotions, but would wish the soul to retain feelings of love and reverence. Morally neutral feelings like boredom, vigour and a sense of humour are presumably debatable.

Of greater concern is the question of memory and the whole issue of our perception of time. Our conception of ourselves is strongly rooted in our memory of past experiences. It is not at all clear that, in the absence of memory, the self would retain any meaning whatever. It might be objected that a person suffering from amnesia may still wonder 'Who am I?' but does not for a moment doubt that there exists an 'I' to whom the 'who' pertains. Still, even an amnesiac is not completely deprived of memory, He has no difficulty, for example, in knowing the use of everyday objects, such as cups and saucers, buses and beds. Furthermore, his short-term memory remains unaffected: if he decides to walk in the garden he does not a few moments later wonder what he is doing there.

If a person did lose the ability to remember his experiences of even a few seconds previously then his sense of identity would completely disintegrate. He would be unable to act or behave coherently at all. His bodily movements would not be coordinated in any conscious pattern of action. He would be totally incapable of making any sense of his perceptions, and could not even begin to order his experiences of the world about him. The whole notion of *himself* as distinct from his perceived world would be chaotic. No pattern or regularity in events would be apparent, and no concept of continuity especially personal continuity — could be maintained.

It is thus largely through memory that we achieve a sense of personal identity, and recognize ourselves as the *same* individual from day to day. Throughout life we inhabit one body, but the body can undergo considerable changes. Its atoms are systematically replaced as a result of metabolic activity; it grows, matures, ages, and eventually dies. Our personalities also undergo major changes. Yet through this continuous metamorphosis, we believe that we are one and the same person. If we had no memory of earlier phases of our life, how could the concept 'same person' have any meaning, save in the sense of bodily continuity?

Suppose a man claimed to be a reincarnation of Napoleon. If he did not look like Napoleon the only criterion by which you could judge his claim would be that of memory. What was Napoleon's favourite colour? How did he feel before the battle of Waterloo? You would expect him to relate some specific (and preferably verifiable) information about Napoleon before taking the claim seriously. Suppose, however, that the man declared that he had lost all memory of his previous life, save only that he *was* Napoleon, what should you make of it? What would it mean for him to say 'I was Napoleon'?

'What I mean,' he would perhaps counter, 'is that, although my body and my memory, and indeed my entire personality, are now those of John Smith, the soul of John Smith is none other than that of the late Napoleon Bonaparte. I *was* Napoleon, *now* I am Smith, but it is the same me. Only my characteristics have changed.' But is this not jibberish? For what is to identify one person's mind from another other than their personality or their memory? To claim that there is some sort of transferable label — the soul — which is otherwise quite devoid of properties save to display some mystical registration mark, is a totally meaningless conjecture. What would we say to someone who denied its existence? Could we not invent souls for everything in this way — for plants and clouds, rocks and airplanes? 'This looks like an ordinary diesel locomotive,' one might declare, 'but in fact it contains the essence, the soul, of Stevenson's original Rocket! The design is different, the materials are different, the performance bears no resemblance to the Rocket, but it is actually the *same* locomotive with a totally new structure, appearance and design.' What is the use of such an empty assertion?

To take a more plausible example than reincarnation, suppose that a close friend were to undergo major surgery which is so comprehensive that he is physically totally unrecognizable afterwards. How would you know he was the same person? If he related to you facts about his earlier life, reminded you of small incidents and personal conversations, and generally displayed a good acquaintance with his former circumstances, you would be inclined to conclude it was indeed the same man. 'That's him all right. Nobody else could know that.' But if the surgeon also removed much of your friend's memory, or perhaps damaged it, your judgment of his identity would be far less confident. If he had no memory at all, you would have no grounds (except perhaps some residual bodily evidence) for saying that the man before you was your friend. In fact, it is not clear that someone with *no* memory can really be thought of as a person at all; he would possess none of the coherent features, such as a personality, that we would normally associate with an ‘individual’. His responses would either be totally random, or pure reflex, and as such his behaviour would be little different from a rather badly programmed automaton.

The difficulty that one sees here for the dualist who believes in the survival of the soul is obvious. If the soul depends on the brain for memory storage, how can the soul remember anything after the death of the body? And if it can't remember anything, what right have we to attribute a personal identity to it? Or are we to suppose that the soul has a sort of non-material back-up memory system that functions in parallel with the brain but can equally well cope on its own?

Sometimes an attempt is made to break out of this deadlock by asserting that the soul transcends time. Just as the soul cannot be located in space, so it has no location in time. But this manoeuvre brings with it a fresh crop of difficulties, as we saw in the previous chapter.

We seem to approach closer to an understanding of the self by noting a point made by many philosophers: that human consciousness does not consist merely of awareness, but of self-awareness — we *know* that we know. In 1690 John Locke emphasized that it is ‘impossible for anything to perceive without *perceiving* that he does perceive.’[4] The Oxford philosopher, J.R. Lucas, expresses this point as follows:

> In saying that a conscious being knows something, we are saying not only that he knows it, but that he knows he knows it, and that he knows he knows he knows it, and so on… The paradoxes of consciousness arise because a conscious being can be aware of itself, as well as of other things, and yet cannot really be construed as being divisible into parts.[5]

In similar vein A.J. Ayer has written: ‘There is a temptation to think of one's self as a set of Chinese boxes, each surveying the one it immediately encloses.’[6]

There is no doubt that the quality of self-reference is a key one in unravelling the mystery of the mind. We have already encountered the importance of feedback and self-coupling in Prigogine's dissipative structures which have the capacity for self-organization, and there seems to be a natural progression from the inanimate through the animate to the conscious — a hierarchy of complexity and self-organization. But there is another hierarchy buried in this progression — a hierarchy of conceptual levels discussed in the previous chapter. Life is a holistic concept, the reductionist perspective revealing only inanimate atoms within us. Similarly mind is a holistic concept, at the next level of description. We can no more understand mind by reference to brain cells than we can understand cells by reference to their atomic constituents. It would be futile to search for intelligence or consciousness among individual brain cells — the concept is simply meaningless at that level. Clearly, then, the property of self-awareness is holistic, and cannot be traced to specific electrochemical mechanisms in the brain.

The study of self-reference has always encountered a touch of paradox, not only in the philosophical question of self-awareness but in the arts and even at the logical and mathematical level. The Greek scholar Epeminides drew attention to the problems of self-referring statements. Normally we assume that every meaningful statement must either be true or false. But consider Epimenides' proposition (which we call A) that can be paraphrased thus:

A: This statement is false.

Is A true, or false? If true, then the statement itself declares it is false; if false, the statement must be true. But A cannot be both true and false, so the question ‘Is A true, or false?’ has no answer.

We ran into a similar problem in the form of Russell's paradox in Chapter 3. In both cases, absurdity seems to follow from perfectly innocuous statements or concepts when they are looped around and directed at themselves. An equivalent form of A is:

A: The following statement is true. Al
The preceding statement is false. A2

In this form, each individual statement, A1 and A2, is perfectly straightforward and free of paradox, but joined together into a self-referring loop they appear to make nonsense of logic.

In his remarkable book, Hofstadter points out how ‘locally’ sensible concepts that loop into paradox ‘globally’ have received dramatic artistic representation in the work of the Dutch artist M.C. Escher. Consider his *Waterfall* for example. If we follow the path of the water around the loop, at each stage its behaviour seems perfectly normal and natural until suddenly, with a shock, we find ourselves back where we started. The entire loop, taken as a whole, is manifestly an impossibility, yet at no point on the path around the loop does anything go ‘wrong’. It is purely the global, or holistic, aspect that is paradoxical. Hofstadter also finds a musical equivalent of these ‘strange loops’ in Bach's fugues.

Penetrating investigations of self-reference have been carried out by mathematicians and philosophers concerned with the logical foundations of mathematics. Perhaps the most startling accomplishment of this program is a result proved by the German mathematician Kurt Gödel in 1931, known as the Incompleteness Theorem, which forms the linking theme of Hofstadter's book. Gödel's theorem sprang from the attempt by mathematicians to systematize the process of reasoning in order to clarify the logical basis on which the edifice of mathematics is built. Russell's paradox, for example, arose from efforts to organize concepts in as general and non-committal a way as possible by allocating them to ‘sets’ — with disastrous results.

Gödel hit upon the idea of using mathematical objects to codify statements. In itself that is nothing new or sensational. Anyone who has read an enumerated contract is familiar with the practice. The novel feature which Gödel explored was the use of mathematics to codify statements about mathematics — the self-referring aspect again. Perhaps inevitably, something similar to the Epeminides' paradox emerged, but as a statement about mathematics; in fact, about good old-fashioned numbers 1,2,3… Gödel demonstrated in his theorem that there always exist statements about numbers that can *never*, even in principle, be proved either true or false (like A above), on the basis of a fixed set of axioms. Axioms are the things you assume are true without proof (e.g. 1 = 1). Thus, even a mathematical system as relatively simple as the theory of numbers possesses properties that cannot be proved (or disproved) on the basis of a fixed set of assumptions, however complex and numerous those assumptions may be!

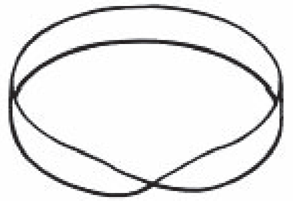
The importance of Gödel's Incompleteness Theorem is that, by mixing subject and object, it demonstrates how, even at the fundamental level of logical analysis, self-reference can produce either paradox or indecision. It has been taken to imply that one can never, even in principle, understand one's own mind completely. Hofstadter conjectures: 'Gödel's Incompleteness Theorem (has) the flavour of some ancient fairy tale which warns you that "To seek self-knowledge is to embark on a journey which… will always be incomplete".'[7]

Gödel's theorem has also been used to argue for the non-mechanical nature of the mind. In an essay entitled 'Minds, Machines and Gödel', Lucas asserts that human intelligence can never be attained by computers: 'Gödel's theorem seems to me to prove that Mechanism is false, that is, that minds cannot be explained as machines.' The essence of his argument is that we, as human beings, can discover mathematical truths about numbers which a computer, programmed to work within a fixed set of axioms, and therefore subject to Gödel's theorem, cannot prove for itself:

> However complicated a machine we construct it… will be liable to the Gödel procedure for finding a formula unprovable-in-that-system. This formula the machine will be unable to produce as being true, although a mind can see it is true. And so the machine will still not be an adequate model of the mind.[8]

No doubt many people would feel uncomfortable about basing the mind's superiority on esoteric mathematics, when it is usually qualities such as love, appreciation of beauty, humour, and so on that are cited as evidence for a non-mechanical mind, or 'soul'. In any case, Lucas's argument has been attacked on a number of grounds. For example, Hofstadter points out that, in practice, the human mind's capacity for discovering complicated mathematical truths is limited, so one could still program a computer that could successfully prove everything that a given person can ever discover about numbers. Moreover, it is easy to convince oneself that *we* are just as vulnerable as computers to Gödel incompleteness because of Epimenides-type statements; it is possible to construct logical truths about the world involving Smith that can never be proved by Smith!

As emphasized in the foregoing, consciousness, the impression of free will and the sense of personal identity all involve an element of self-reference and can have paradoxical aspects. When a person perceives something — a physical object, for instance — the observer is by definition external to the observed object, though coupled to it through some sensory mechanism. But during introspection — an observer observing himself — both subject and object coincide in a most perplexing way. It is as though the observer is both inside and outside himself.



10 .The famous Möbius band is made by putting a single twist in a strip and joining the ends to form a loop. Careful examination reveals that the band now has only one side and one edge.

Some intriguing representations can be given for this curious mental topology. Consider the famous Möbius band (see Fig. 10), for example. The band is constructed by making a single twist in a strip of material and then joining it into a closed loop. At any particular point on the band, there would seem to be both a front side and reverse side of the band. But if you follow a route around the loop, you will see that there is actually only one side. Locally there appears to be a division into two categories (analogous to subject and object) but a glance at the global structure shows there is only one.

Another suggestive representation of self-reference is given by Hofstadter in the language of his Strange Loops:

> My belief is that the explanations of 'emergent' phenomena in our brains – for instance, ideas, hopes, images, analogies, and finally consciousness and free will – are based on a kind of Strange Loop, an interaction between levels in which the top level reaches back down towards the bottom level and influences it, while at the same time being determined by the bottom level… The self comes into being the moment it has the power to reflect itself.[9]

The essential feature in all these attempts to grope towards a better understanding of the self is the convolution of hierarchical levels. The hardware of brain cells and electrochemical machinery supports the software level of thoughts, ideas, decisions, which in turn couple back to the neural level and so modify and sustain their own existence. The attempted separation of brain and mind, body and soul, is a confusion born of trying to sever these two convoluted levels (or 'Tangled Hierarchy' in Hofstadter's parlance). But it is a meaningless enterprise, for it is the very entanglement of the levels that makes you *you*.

Remarkably, modern Christian doctrine has moved a long way towards this picture of the integrated brain and mind, with its emphasis on the resurrection of the *whole man* through Christ, rather than the traditional idea of a distinct immortal soul being cast adrift from its material counterpart to carry on a disembodied existence somewhere.

However, nothing that has been said about the mind is specifically restricted to human beings. There seems to be no scientific evidence for any special divine quality in man, and no fundamental reason is apparent why an advanced electronic machine should not, in principle, enjoy similar feelings of consciousness as ourselves. This is not, of course, to say that computers have souls, but rather that the complex tangle of convoluted levels which produce what we understand as mind can arise in a variety of systems.

Yet there still remains one aspect of the self that seems to be contradicted by the lowlevel, deterministic description, and that is the *will*. All human beings believe that they are capable of choosing, in a limited way, between various courses of action available to them. Can such an apparent freedom to initiate actions ever be programmed into a computer?

Hofstadter argues that in principle we can. He describes the feeling we have of free will as a delicate balance between self-knowledge and self-ignorance. By incorporating an appropriate degree of self-reference into a computer programme, Hofstadter claims that it too would start to behave as though it had a will of its own. He tries to tie in the will with the Gödel-like incompleteness which inevitably arises in any system capable of monitoring its own internal activity. (The subject of free will and determinism will be explored in greater depth in Chapter 10.)

Suppose that one were persuaded by these arguments that human brains are marvellously complex electrochemical machines, and that other types of artificial mechanisms, such as computers, can be programmed for free will and human-like emotions. Does this devalue the human mind? Recall the trap of 'nothing-buttery'. To assert that the brain is a machine does not deny the reality of mind and emotion, which refers to a higher level of description (the ant colony, the plot of the novel, the picture on the jig-saw, the Beethoven symphony). To say a brain is a machine need not imply that the mind is *nothing but* the product of mechanistic processes. To claim that the deterministic nature of brain activity renders free will an illusion is as misconceived as the claim that life is an illusion because of the underlying inanimate nature of atomic processes.

A number of science fiction writers have developed the idea of machines with minds, most notably Isaac Asimov in his robotics stories, and Arthur C. Clarke in the novel *2001: A Space Odyssey*. More penetrating analyses have also been given by some writers who visualize 'mind transplants' in an attempt to clarify the definition of the self.

Consider, for example, what would happen if your brain could be removed and placed in a 'brain support system', remaining coupled to your body via some sort of radio communication network. (Of course, such a procedure is utterly beyond foreseeable technology but there is no logical reason why it could not be achieved.) Your eyes, ears and other senses remain functioning as usual. Your body can operate without impediment. In fact, nothing would seem any different (perhaps a feeling of light-headedness!), except that you could look down upon your own brain. The question is, where would *you* be? If your body takes a train journey, your experiences are those of someone on the journey, exactly as they would be if your brain were still in your skull. You would certainly *feel* as though you were on the train.[10]

Perplexity mounts if we now envisage your brain being transplanted into another body instead. Would it be correct to say that *you* had a new body, or that *it* had a new brain? Could you regard yourself as the same person, with a different body? Perhaps you could. But suppose the body were of the opposite sex, or that of an animal? Much of what makes *you*, your personality, capabilities and so forth, is tied to chemical and physical conditions of the body. And what if your memory were wiped out during the transfer? Does it then make any sense at all to regard the new individual as *you*?

Fresh problems arise when one speculates about duplication of the self. Suppose the entire information content of your brain were put on a giant computer somewhere, and your original body and brain died. Would *you* still survive — in the computer?

The idea of putting minds into computers raises the prospect of multiple duplicates of you being copied on other computers. Of course, much has already been written about ‘multiple personality' mental disorders, and the cases where patients have had the connections between the left and right hemispheres of their brains severed, leading to mental states where, crudely speaking, the left hand literally does not know what the right hand is doing.

Though some of these ideas may seem fearsome, they do hold out the hope that we can make scientific sense of immortality, for they emphasize that the essential ingredient of mind is *information*. It is the pattern inside the brain, not the brain itself, that makes us what we are. Just as Beethoven's Fifth Symphony does not cease to exist when the orchestra has finished playing, so the mind may endure by transfer of the information elsewhere. We considered above how, in principle, the mind can be put on a computer, but if the mind is basically ‘organized information’ then the medium of expression of that information could be anything at all; it need not be a particular brain or indeed any brain. Rather than ‘ghosts in machines’, we are more like ‘messages in circuitry' and the message itself transcends the means of its expression.

MacKay expresses the viewpoint in computer language:

> If a computer operating a given program were to catch fire and be destroyed, we would certainly say that that was the end of that particular embodiment of the program. But if we wanted that same program to run in a fresh embodiment, it would be quite unnecessary to salvage the original computer parts or even to replicate the original mechanism. Any active medium (even operations with pencil-and-paper) which gave expression to the same structure and sequence of relationships could in principle embody the very same program.[11]

This conclusion leaves open the question of whether the ‘program’ is re-run in another body at a later date (reincarnation), or in a system which we do not perceive as part of the physical universe (in Heaven?), or whether it is merely ‘stored’ in some sense (limbo?). As far as the perception of time is concerned, we shall see that it is only during the running of the program, as in the actual playing of a symphony, that any meaning can be attached to the flow of time. The existence of a program, like that of a symphony, once created, is essentially timeless.

In this chapter it has been argued that research in the cognitive sciences has tended to emphasize the similarities between mind in man and machine, with mixed implications for religion. While on the one hand these studies leave little room for the traditional idea of the soul, on the other hand they leave open the possibility of survival of the personality.

Minds, being complex, are not usually studied in the framework of physics that, as we have seen, operates best at the reductionist level on simple elementary things. However, there is one important area of the new physics into which mind has intruded at a fundamental level, much to the mystification of physicists. It is called the quantum theory, and it leads us into an Alice-in-Wonderland world that cuts right across the traditional framework of religion.