

OF HUMANS AND HOVERFLIES

Holiday beaches in summertime display an awkward minuet, danced by advancing waves of spume and retreating waves of deck-chairs. As the tide rises, the deck-chairs are repeatedly moved up the shore. Only when they are safe above the high-water line do the sunbathers really relax, knowing that their territory can be encroached upon no further.

The history of science shows a similar pattern, the advance of scientific theory being matched by the retreat of anthropocentrism. Copernicus, Darwin, and Freud successively challenged comfortable beliefs: that Earth is the centre of the universe, that *homo sapiens* was created in the image of God, and that people are fundamentally rational creatures. Since the Renaissance, the deck-chairs of our self-glorification have been moved several times.

Human creativity, in this scenario, lies even further up the beach than rationality does. Inspirationists and romantics lounge there at their ease, confident of being safe from science. But is their confidence misplaced? Sometimes, after all, the high tide covers the beach, and the deck-chairs must be abandoned. Is creativity inviolable?

Well, the three intellectual revolutions cited above each showed some cherished belief to be false. Geocentrism, special creation, rational self-control: one by one, these bit the dust. If modern science were to claim that creativity is an illusion, we could sadly add a fourth example to the list.

But science claims no such thing. The previous chapters have acknowledged creativity over and over again. In brief, a scientific psychology does not deny creativity: it explains it.

To say this, however, is not enough. Many people fear that explanation *in and of itself* must devalue creativity. Forget computers, for the moment: the conviction is that *any* scientific account of creativity would lessen it irredeemably. Even an explanation in terms of brain-processes (never mind silicon-chips) would undermine our respect for creative thought.

A prime source of this common attitude is the widespread feeling that science, in general, drives out wonder. Wonder is intimately connected with creativity. All creative ideas, by definition, are valued in some way. Many make us gasp with awe and delight. We are enchanted by the water-snakes, and fascinated by the benzene ring. To stop us marvelling at the creativity of Bach, Newton, or Shakespeare would be almost as bad as denying it altogether. Many people, then, regard the scientific understanding of creativity more as a threat than a promise.

ANTI-SCIENTIFIC PREJUDICE OF this sort is not new. William Blake had a word for it – or rather, many. ‘May God us keep’, he wrote, ‘From Single vision & Newton’s sleep!’ And again:

I turn my eyes to the Schools & Universities of Europe
And there behold the Loom of Locke, whose Woof rages dire,
Wash’d by the Water-Wheels of Newton: black the cloth
In heavy wreathes folds over every Nation: cruel Works
Of many Wheels I view, wheel without wheel, with cogs tyrannic
Moving by compulsion each other, not as those in Eden, which
Wheel within Wheel, in freedom revolve in harmony & peace.

To some extent, this passage is a protest against the machine-shops of the Industrial Revolution. But ‘the Water-Wheels of Newton’ are the wheels of science, as well as technology.

Blake was not simply objecting to machines, and the way they were changing our culture. Nor was he declaring a belief in the crystalline spheres (the wheels within wheels) of mediaeval cosmology. He was reacting against the scientific enthusiasm that had led Alexander Pope to declare: ‘God said “Let Newton be!”’, and all was light.’ For Blake, Newton’s light made only *single* vision possible. Matters not dealt with by natural science, such as freedom and harmony, were insidiously downgraded and ignored – even tacitly denied.

Science withstood Blake’s attacks, and grew apace. It spawned many new theories and many, many, new facts. But reservations about the scientific world-view remained, and remain to this day.

Some doubters wielded the weapon of humour: Dickens, in the 1830s, mocked the infant British Association for the Advancement of Science (widely referred to as ‘the British Ass’) in his *Mudfog Papers*. The anti-scientific writers of the 1960s ‘counter-culture’ – Theodore Roszak,

for one – were more passionate, if less witty. Like Blake, they criticized science for what they saw as its mechanistic denial of freedom (a criticism discussed later). And they called, specifically, for a return to religious reverence, or wonder, if not to theological dogma.

A nineteenth-century contributor to the *Athenaeum* put it in a nutshell. He had shared a stage-coach with some scientific worthies travelling to the 1834 meeting of the British Association in Edinburgh. This is how he described the experience:

We entered Scotland over the Cheviot Hills. Their appearance attracted the notice of all, and it was soon evident that our fellow-travellers were members of the Association, full of their respective subjects, eager to impart and receive information. . . . *Science destroyed romance* – the field of Chevy Chase scarce elicited a remark – the cross marking the spot where Percy fell was observed by one of the geologists to belong to the secondary formation; the mathematician observed that it had swerved from the perpendicular, and the statisticians began a debate on the comparative carnage of ancient and modern warfare. [Italics added.]

We have all encountered scientifically-minded individuals with manners bad as these, full of their own knowledge but ignorant of history and blind to the beauties of landscape. (We have all met personally obnoxious artists, too.) But does science *necessarily* destroy romance?

SOMETIMES, TO BE SURE, it does. Science is fundamentally opposed to the romance of superstition (which includes the inspirational and romantic ‘theories’ of creativity). When our wonder is based on ignorance, error, or illusion, it must fade in the light of understanding.

But science can lead in turn to a new form of wonder, which is not so easily destroyed.

A friend who is an engineer once told me how, as a very young child, he was for a while utterly fascinated by circles. He would collect circular things – coins, bottle-tops, tins – which he kept in his toy-cupboard, and which he used to draw circles of many different sizes. One day his parents told him that there is an instrument that can draw any circle whatever (up to a certain size). He wondered greatly at this idea, and could hardly wait to receive this marvel as a gift. He thought of it as some sort of magically changing item that could transform itself into

equivalents of all the different objects he had collected in his cupboard.

Then, he was given a compass. He was horribly disappointed, for there was nothing magical at all about the compass. It was boringly simple, and its 'power' was transparent even to an infant. He still remembers this day of disillusion as a traumatic event in his childhood.

Today, however, he has the maturity to see that the compass was indeed wonderful – and the mathematical principle it embodied, even more so. Its simplicity (which can generate many superficially varying cases) is 'boring' only to those who feel that baroque confusion is a necessary mark of the wonderful. Even Blake did not believe this: hence his reference to harmony.

Mysteries were contrasted, in Chapter 1, with puzzles. Physics, chemistry, and molecular biology have already transmuted many mysteries into puzzles, and solved them to boot. Now, psychology is helping us to understand how the previously mysterious behaviour of humans and other animals is possible – and can be scientifically understood.

In some cases, this added understanding makes us react much as the infant engineer did to the compass. That is, the newly-discovered simplicity drives out our wonder, stifles our sense of awesome mystery, leaving us only with the brute facts of science.

Consider the hoverfly, for example. A hoverfly is able to meet another hoverfly in mid-air – which is just as well, since they need to be at the same place if they are to mate. How does this mid-air meeting come about?

One might assume that a hoverfly does something like what a person does, on recognizing a friend across a city-square: altering direction immediately, and adjusting their path as necessary if the friend suddenly swerves. Sentimentalists would expound on the wonders of nature, as illustrated by the marvelous powers of the humble hoverfly. More sober souls (given the assumption in question) might feel some sympathy for such a view. It turns out, however, that this assumption about how the hoverfly manages its social life is false.

On closer examination, there is nothing like the flexible selection and variation of pathways that are involved in truly intelligent friend-seeking action. For the fly's flight-path is determined by a very simple and inflexible rule. This rule, which is hardwired into the insect's brain, transforms a specific visual signal into a specific muscular response. The fly's change of direction depends on the particular approach-angle subtended by the target-fly at the time. The creature, in effect, always assumes that the size and velocity of the seen target (which may or may not be a fly) are those corresponding to hoverflies. When initiating its

new flight-path, the fly's angle of turn is selected on this rigid, and fallible, basis. Moreover, the fly's path cannot be adjusted in mid-flight, there being no way in which it can be influenced by feedback from the movement of the target animal.

This evidence must dampen the enthusiasm of anyone who had marvelled at the similarity between the hoverfly's behaviour and the ability of human beings to intercept their friends. The hoverfly's intelligence has been demystified with a vengeance, and it no longer seems worthy of much respect.

To be sure, one may see beauty (like the beauty of the compass) in the evolutionary principles that enabled this simple computational mechanism to develop, or in the biochemistry that makes it function. But the fly itself cannot properly be described in anthropomorphic terms. Even if we wonder at evolution, and at insect neurophysiology, we can no longer wonder at the subtle mind of the hoverfly.

Many people fear that this disillusioned denial of intelligence in the hoverfly is a foretaste of what science will say about our minds, too. But this is a mistake. The mind of the hoverfly is much *less* marvelous than we had imagined, so our previous respect for the insect's intellectual prowess is shown up as mere ignorant sentimentality. But computational studies of thinking can increase our respect for human minds, by showing them to be much *more* complex and subtle than we had previously recognized.

Think of the many different ways (sketched in Chapters 4 and 5) in which Kekulé could have seen snakes as suggesting ring-molecules. Consider the rich analogy-mapping in Coleridge's mind, which drew on naval memoirs, travellers' tales, and scientific reports to generate the shining water-snakes that swam through Chapter 6. Or remember the mental complexities (outlined in Chapter 7) underlying a plausible, and grammatically elegant, story. Even relatively simple computational principles, such as the jazz-program's rules for melodic contours, can – like the compass – have surprisingly rich results. A scientific psychology, by identifying the mental processes involved in examples like these, can help us to appreciate just how wonderful the human mind is.

Admittedly, poets and novelists have long had an intuitive sense of some of the psychological subtleties concerned. Consider Proust's insightful depiction of memory, or Coleridge's (and Livingston Lowes') comments on mental association. Theoretical psychologists such as Freud relied on similar insights, when discussing symbolism in the dreams and dramas of everyday life. But such notions have remained literary and intuitive, rather than scientifically rigorous. Moreover, even Freud underestimated the degree of complexity of the mental processes

he described. So did Koestler, in his attempts to define 'the bisociation of matrices'.

To cease to wonder at creativity because it had been explained by science would be to commit 'the fallacy of the compass'. The temporary disaffection of the infant engineer was irrational and unnecessary. Now, he still values circles for the beauty of their superficial form. But he also appreciates the underlying mathematical principle, which enables them to be generated by anyone with a compass to hand. He has gained, not lost.

A scientific psychology, then, allows us plenty of room to wonder at Mozart, and even at Grandpa's jokes. Much as geology leaves the Cheviot Hills as impressive as ever, and Chevy Chase as poignant, so psychology leaves poetry in place. Indeed, it adds a new dimension to our awe on encountering the water-snakes, or the theory of the benzene ring.

Darwin made a similar point about biology. Rather than denying our wonder at God's creation, he said, evolutionary theory can increase it:

[I think it] an idea from cramped imagination, that God created the Rhinoceros of Java and Sumatra, that since the time of the Silurian he has made a long succession of vile molluscous animals. How beneath the dignity of him, who is supposed to have said 'let there be light,' and there was light.

'The more magnificent view,' he said, is that all these creatures, along with their more aesthetically appealing cousins, have been produced by 'the body's laws of harmony'.

'**A**_H, ' SOMEONE MIGHT say, on reading Darwin's words above, 'But there's the rub! The *body's* laws of harmony are one thing. Brains may well have creative powers. But computers are quite another matter!'

This objection can be interpreted in at least three ways. These recall the first, second, and last Lovelace-questions, respectively.

It may mean that computers are utterly irrelevant to human creativity, and cannot possibly help us to understand it. It may mean that computer performance could never match ours, that there will never be a computerized Chopin, or Donne. Or it may mean that computers, unlike people, cannot *really* be creative.

Let us consider each of these, in turn.

AS REGARDS THE first interpretation (that computers are utterly irrelevant to human creativity), it is *computational concepts and theories*, not computers as such, which are crucial for psychology. Computational psychology tries to specify the conceptual structures and processes in people's minds. The material embodiment of a particular computation need not be silicon, or gallium arsenide, or anything else dreamt up by computer-engineers. It may be good old-fashioned neuroprotein, working inside human heads.

Computers are very useful nevertheless, because their programs are effective procedures. If a jazz-program produces acceptable music, we *know* that essentially similar computations grounded in the brain could be the source of real, live, jazz.

Anyone who doubts whether the brain does things in quite that way may be right. But they should provide specific evidence, not vague intuitive prejudice, to support their case. (Johnson-Laird does this when he argues that limits on short-term memory make it impossible for jazz-melodies to be improvised by hierarchical grammars, and that limits on long-term memory prevent them being strings of motifs.)

Preferably, the sceptics should offer an alternative psychological theory, with equal clarity and more appropriate generative power. If they manage to do so, then the original hypothesis will have been scientifically fruitful, even though rejected (a common fate of creative ideas in science). If they do not, then the contested theory stands, as the most promising explanation so far.

Promising explanations abound, as we have seen throughout this book. Even motivation and emotion have been analysed in computational terms, though these theories are relatively sketchy as yet. Despite such theoretical gaps, there are many ways in which a computational approach can help to explain human creativity. It enables us to see what sorts of process may underlie our ability to learn new concepts (patterns), and to combine them in novel ways. Moreover, conceptual spaces – and the ways in which they can be mapped, explored, and transformed – are made more precisely intelligible by thinking of them in computational terms.

The concept of a generative system, for instance, enables us to understand how ideas can appear which, in an important sense, *could not* have appeared before. It helps us to focus on styles of thinking in science and the arts, and to analyse how they can be changed in more or less radical fashions. These questions are the concern of musicologists, literary critics, and historians of art and science, whose insights we need if we are to understand creativity. Many of these insights are highly subtle, and not easily expressed in precise terms. But our discussions of

harmony, jazz, and line-drawing – not to mention scientific creativity of various sorts – showed that computational psychology can already say something specific about such matters.

Computer modelling enables us to test our psychological theories by expressing them as effective procedures, as opposed to verbal theories, vaguely understood. It helps us to see how complex, yet how constrained, is the appreciation of a piece of music – or its improvisation. It replicates some scientific discoveries from centuries past, and sharpens our sense of what certain methods of reasoning *can* and *cannot* achieve. And it confirms the insight that changing the representation of a problem may make it much less difficult, for identifiable reasons.

The notion of heuristic search, controlled in specifiable ways, helps us to understand how a wide range of H-creative ideas could have arisen. Scripts, and related concepts, indicate some ways in which knowledge may be organized within the mind, and help us to appreciate the many varied constraints involved in writing a story. The ‘un-naturalness’ of step-by-step programming is largely irrelevant: even though the brain often uses heuristics, scripts, and frames in parallel, it may be exploring conceptual spaces whose dimensions are those embodied in sequential models.

The richness and subtlety of ‘ordinary’ psychological abilities are highlighted by this approach. We can see, much more clearly than Coleridge or Koestler could, how everyday functions of memory and comparison might underlie creative thinking. And we can see how even the mundane task of describing a game of noughts-and-crosses requires the integration of many different constraints, not only the rules of grammar.

Moreover, computational psychology is to some extent inspired by ideas about the brain. Connectionist models give us some grip on the fleeting psychological processes involved in poetic imagery, scientific analogy, and serendipity. They define mechanisms capable of global recognition, where no single feature is necessary but many features are sufficient. And they show how a mind (and even a computer) can accept imperfect pattern-matches, and retrieve an entire associative complex on being given a fragment of it. Snakes and water-snakes, given their creative contexts of molecules and mariners, are made less mysterious accordingly.

Connectionism must be combined with other kinds of computational theory, in order to model the deliberate thinking found in the evaluation phase (and often in the preparatory phase, too). One of the most active research-areas at present is the design of ‘hybrid’ systems, combining the flexible pattern-matching of connectionism with sequential processing

and hierarchical structure. A psychological theory of creativity must explain both types of thinking, and how they can co-exist in a single mind.

Many other examples have been given in support of my claim that the first Lovelace-question merits the answer 'Yes'. If you are not convinced by now, nothing I could say at this point would help. With respect to the first interpretation of 'Computers are another matter!', then, I rest my case.

UNDER THE SECOND interpretation (that computer performance could never match ours), the objection is less easily countered.

If future programs are to match all our creative powers, then future psychology must achieve a *complete* understanding of them. Our knowledge of the human mind, and of its potential, has been greatly increased by the computational approach. Neuroscience will doubtless improve our understanding of these matters, too. But there are many still-unanswered questions about human thinking, most of which have not even been asked. Why should anyone believe that we shall find answers to every single one of them?

The conceptual space of theoretical psychology is too large for us to explore every nook and cranny. Perhaps those never-to-be-explored regions hold some secrets of Chopin's music, and Donne's poetry. Likewise, some physical diseases may never be understood: it does not follow that physiology and molecular biology are a waste of time. Science cannot answer, or even ask, all possible scientific questions.

Even if it could, scientists would not want to waste their time, and money, integrating all these explanatory principles within a single computer model. (Combining BACON with its four P-creative cousins is child's play by comparison.) Some special-purpose computer systems will surely be built, to do for other sciences what DENDRAL did for chemistry – indeed, much more. But there are easier, and more enjoyable, ways of generating new wits and new poets. As Koestler said,

The difficulty of analysing the aesthetic experience is not due to its irreducible quality, but to the wealth, the unconscious and non-verbal character of the matrices which interlace in it, along ascending gradients in various dimensions.¹

The preceding chapters have suggested how many, and how very various, these dimensions are. To put all of them into one AI-model would in practice be impossible.

Moreover, we have seen that human creativity often involves highly idiosyncratic experiences (Proust's madeleine, for one). But theoretical psychology is concerned with general principles, not personal biographies or gossip. Even when a psychologist does discuss detailed personal evidence (as in Darwin's notebooks, within which one can see the concept of evolution being gradually developed), these are used as grist for the theoretical mill.

A computer-model embodying general psychological principles has to be given some content, something idiosyncratic to chew on. It may be tested on Socrates' philosopher-midwife analogy, as we have seen. But it would not be worth anyone's while, even assuming it were feasible, to build in all of Socrates' knowledge and experience – including his marital tiffs with Xanthippe and his filial bonds with the real midwife, Phainarete. Unless that were done, however, the rich texture of Socrates' thought would not have been modelled in detail, and his creativity could not be fully captured. This has nothing to do with Socrates' greatness: your creativity, and your next-door neighbour's, cannot be exhaustively modelled either.

Future computers will perform 'creatively', up to a point. (Arguably, some already do.) So the answer to the second Lovelace-question is a guarded 'Yes'. But to expect a computer-model to match the performance of Chopin or Donne is unrealistic. Even to mimic the wit and wisdom of a schoolgirl's letter to her best friend is probably too great a challenge. To await a computer Shakespeare is to wait for Godot.

THE THIRD INTERPRETATION of 'Computers are quite another matter' raises very different – and very difficult – issues. It holds computers to be intrinsically incapable of genuine creativity, *no matter how impressive their performance may be*.

For the purpose of argument, let us assume that computers could one day appear to be as creative as we are. They might have their blind spots, as we do too: sneezing and chilblains (two examples drawn from Chapter 1) might be understood by them only in a theoretical, book-learned, way. But they would produce countless ideas – cantatas, theorems, paintings, theories, sonnets – no less exciting than ours. And they would do so by means of computational processes like those which, according to theoretical psychologists, go on in human heads.

Even so, this objection insists, they would not really be creative. Indeed, they would not really be intelligent. Artificial intelligence (on this

view) is analogous not to artificial light, but to artificial five-pound notes. Far from being an example of the same sort of thing, it is something utterly different – and to pretend otherwise is fraudulence.

‘Well, the question’s clear enough!’ you may say, ‘What’s your answer to it?’ – Not so fast! We are dealing, here, with the fourth Lovelace-question: could a computer *really* be creative? And this question is not clear at all. Indeed, the answer ‘No’ may be defended in (at least) four different ways. Let us call these the brain-stuff argument, the empty-program argument, the consciousness argument, and the non-human argument.

The brain-stuff argument relies on a factual hypothesis: that whereas neuroprotein is a kind of stuff which can support intelligence, metal and silicon are not. The empty-program argument makes a philosophical claim: that all the symbols dealt with by a computer program are utterly meaningless to the computer itself. The consciousness argument claims, similarly, that no computer could conceivably be conscious. And the non-human argument insists that to regard computers as *truly* intelligent is not a mere factual mistake (like saying that a hoverfly’s blood is exactly the same as ours), but a moral absurdity.

We must consider each of these four arguments, for someone who insists that computers cannot *really* be creative may have any one (or even all) of them in mind.

HOVERFLIES, AS WE have seen, are not very bright. According to the brain-stuff argument, however, even hoverflies have an infinitely greater claim to intelligence than computers do. For the measure of computer-intelligence is precisely zero.

More accurately, the claim is that computers made of inorganic materials must be for ever non-intelligent. Only ‘biological’ computers, built from synthetic or naturally-occurring organic compounds, could ever achieve real thinking. Hoverflies, who share the same genetic code as we do, and whose bodies contain chemicals broadly comparable to ours, can perhaps claim a fleeting antenna-hold on intelligence. But computers cannot. In a word: no biochemistry, no creativity.

The main factual assumption driving this argument may, conceivably, be true. Possibly, computers are made of a sort of material stuff which is incapable of supporting intelligence. Indeed, neuroprotein may be the *only* substance in the universe which has this power.

Then again, it may not. Even carbon-strings and benzene-rings may not be necessary: there may be creative intelligences on Mars or Alpha

Centaurs, with alien chemicals filling their heads. Science does not tell us this is impossible.

‘Never mind Martians!’ someone may say, ‘We’re talking about computers. It’s obvious that metal and silicon can’t support intelligence, whereas neuroprotein can.’

But this is not obvious at all. Certainly, neuroprotein does support intelligence, meaning, and creativity. But we understand almost nothing of how it does so, *qua* neuroprotein – as opposed to some other chemical stuff. Indeed, insofar as we do understand this, we focus on the neurochemistry of certain basic *computational functions* embodied in neurones: message-passing, facilitation, inhibition, and the like.

Neurophysiologists have discovered the ‘sodium-pump’, for instance. This is the electrochemical process, occurring at the cell membrane, which enables an electrical signal to pass (without losing strength) from one end of a neurone to the other. And they have studied the biochemistry of neurotransmitters, substances (such as acetylcholine) that can make it easier – or harder – for one nerve-cell to cause another one to fire. In a few cases, they have even been able to say something about how a cell’s chemical properties (and connectivities) enable it to code one sort of information rather than another: picking up colours or light-intensity gradients, for instance, or sounds of varying pitch.

If we could not recognize sounds of varying pitch, we could not appreciate music. Heuristics of harmony, like those described earlier, can be applied only if one can hear (for instance) that one note is a semitone lower than another. If the neurophysiologist can tell us not only which auditory cells enable us to do this, but what chemical processes are involved, all well and good. But the neurochemistry is interesting only to the extent that it shows how it is possible (in human heads) to compute tonal relationships. *Any other chemistry would do*, provided that it also enabled harmonic intervals to be computed.

Likewise, we need to see lines if we are to draw (or appreciate drawings of) acrobats. But any chemistry would do in the visual system, so long as light-intensity gradients could be identified by means of it. Again, any chemical processes would do at the cell-membrane, and at the synapse, provided that they allowed a nerve-cell to propagate a message from one end to the other and pass it on to neighbouring neurones.

Recognizing sounds and lines are computational abilities which some computers already possess. Our own mental life, however, contains much more than tonal harmony and line-drawing. Conceivably, there may be other sorts of computation, going on inside human heads, which simply cannot be embodied in anything made of metal and silicon. But we have no specific reason, at present, to think so. Conceivably, too, only

neuroprotein can implement the enormous number of stable yet adaptable structures involved in human thought, and/or do so in manageable space and time. Again, however, we have no particular reason to think so (and what is to count as ‘manageable’?).

The fact that we cannot see how metal and silicon could possibly support ‘real’ intelligence is irrelevant. For, intuitively speaking, we cannot see how neuroprotein – that gray mushy stuff inside our skulls – can do so either. *No* mind–matter dependencies are intuitively plausible. Nobody who was puzzled about intelligence (as opposed to electrical activity in neurones) ever exclaimed ‘*Sodium* – of course!’ Sodium-pumps are no less ‘obviously’ absurd than silicon chips, electrical polarities no less ‘obviously’ irrelevant than clanking metal. Even though the mind–matter roles of sodium-pumps and electrical polarities are scientifically compelling, they are not intuitively intelligible. On the contrary, they are highly counter-intuitive.

Our intuitions will doubtless change, as science advances. Future generations may come to see neuroprotein – and perhaps silicon, too – as ‘obviously’ capable of embodying mind, much as we now see biochemical substances in general as obviously capable of producing other such substances (a fact regarded as intuitively absurd, even by most chemists, before the synthesis of urea in the nineteenth century). As yet, however, our intuitions have nothing useful to say about the material basis of intelligence.

In sum, the brain-stuff argument is inconclusive. It reminds us that computers made of non-biological materials may be incapable of real creativity. But it gives us no reason whatever to believe that this is actually so.

YOU MAY HAVE come across the empty-program argument on your TV-screen or radio, or in your newspaper. For a recent version of it, based on John Searle’s intriguing fable of the Chinese Room, has figured in the international media.²

Searle imagines himself locked in a room, in which there are various slips of paper with doodles on them. There is a window through which people can pass further doodle-papers to him, and through which he can pass papers out. And there is a book of rules (in English) telling him how to pair the doodles, which are always identified by their shape. One rule, for example, instructs him that when *squiggle-squiggle* is passed in to him, he should give out *squoggle-squoggle*. The rule-book also provides for more complex sequences of doodle-pairing, where only the first and last steps

mention the transfer of paper into or out of the room. Searle spends his time, while inside the room, manipulating the doodles according to the rules.

So far as Searle-in-the-room is concerned, the *squiggles* and *squoggles* are meaningless. In fact, however, they are characters in Chinese writing. The people outside the room, being Chinese, interpret them as such. Moreover, the patterns passed in and out at the window are understood by them as questions and answers respectively: the rules ensure that most of the questions are paired, either directly or indirectly, with what they recognize as a sensible answer. Some of the questions, for example, may concern the egg foo-yong served in a local restaurant. But Searle himself (inside the room) knows nothing of that. He understands not one word of Chinese. Moreover, he could never learn it like this. No matter how long he stays inside the room, shuffling doodles according to the rules, he will not understand Chinese when he is let out.

The point, says Searle, is that Searle-in-the-room is acting as if he were a computer program. He is all syntax and no semantics, for he is performing purely formal manipulations of uninterpreted patterns. It is the shape of *squiggle-squiggle*, not its meaning, which makes him pass a particular piece of paper out of the window. In this sense, his paper-passing is like the performance of a 'question-answering' program, such as the restaurant-program or BORIS (mentioned in Chapters 5 and 7 respectively). But Searle-in-the-room is not *really* answering: how could he, since he cannot understand the questions?

It follows, according to Searle, that a computational psychology cannot explain how it is possible for human beings to understand meanings. At best, it can explain what we do with meanings, once we understand them. (So Searle's answer to the *first* Lovelace-question – whether computational concepts can help us understand human creativity – would be 'Perhaps, but not at a fundamental level'.) *A fortiori*, no program could ever give a computer the ability to understand. Complexity does not help, since all it does is add more internal doodle-matchings: no mega-BORIS of the future will really understand stories. The answer to the last Lovelace-question, on this view, must be 'No'.

If you are not immediately convinced by Searle's argument, your first response may be to object that he has cheated. Of course Searle-in-the-room will never learn Chinese, for he is not causally plugged-in to the world. No one can really understand what egg foo-yong is without being able to see it, smell it, and poke at it with chopsticks. A computer capable of constructing real meanings would need to be more than a VDU-screen

attached to a teletype – which is the sort of computer that Searle-in-the-room is mimicking.

Thus far, Searle would agree with you. But he responds by imagining that the room, with him (and a new rule-book) in it, is placed inside the skull of a giant robot. Now the *squiggle-squiggles* are caused by processes in the robot's camera-eyes, and the *squoggle-squoggles* cause levers in the robot's limbs to move. Accordingly, the egg foo-yong can be picked up, and transferred into the robot's mouth. But, Searle argues, Searle-in-the-robot has no idea what it is to move a chopstick, and does not know an egg foo-yong when the camera records it.

Notice that Searle, in these imaginary situations, must understand the language (English, we are told) of the rule-books. Notice, too, that the rule-books must be at least as detailed as AI-programs are. They must include rules for using language grammatically (no simple matter, as we have seen), and rules covering vision and motor control. In short, in order to write the rule-books one would need a powerful computational psychology, involving theoretical concepts like those discussed in previous chapters. Whether English is sufficiently precise for this task is – to put it mildly – doubtful. What would be needed is something like an AI-programming language. And this language would have to be taught to Searle before his incarceration, for the story depends on his being able to understand it. We shall return to this point later.

Searle's crucial assumption is that computer programs are semantically empty. That is, they consist of abstract rules for comparing and transforming symbols not in virtue of their meaning, but merely by reference to their shape, or form. The so-called symbols are not really symbols, so far as the computer is concerned. To the computer, they are utterly meaningless – as Chinese writing is to Searle-in-the-room. Human beings may interpret them as concepts of various kinds, but that is another matter.

That programs are empty in this sense is taken for granted by Searle. But it is true only if we think of them in a particular way.

Remember the necklace-game: although I described it in terms of red, white, and blue beads, I said that it was based on Hofstadter's 'pq-system'. The pq-system is a logical calculus, whose rules generate strings of letters. As such, it has syntax (only letter-strings with certain structures are allowed), but no semantics. Considered purely as a logical calculus, wherein the letters are not interpreted in any way, it is meaningless. But considered as a set of rules for necklace-building (or for doing addition), it is not.

Remember, too, the distinction (made in Chapter 3) between the two sorts of computational *could*. One is a timeless, abstract, sense: *could* this

set of numbers (successive squares, for instance) be generated, in principle, by such-and-such a mathematical rule? The other is a procedural, and potentially practical, sense: *could* this rule actually be used to produce the relevant set of numbers? The former question asks whether a certain structure lies within a particular conceptual space. The latter asks whether the space can be explored, in a specific manner, so as to find it. Both types of question can be applied to computer programs.

One can, for some purposes, think of a computer program as an uninterpreted logical calculus, or abstract mathematical system. This is useful if one wants to know whether it is capable, in principle, of producing certain abstractly-specified results. (For example: could it get stuck in an infinite loop, or could it distinguish grammatical sentences from nonsensical word-strings?) But one must not forget that a computer program is *a program for a computer*. When a program is run on suitable hardware, the machine *does* something as a result. It builds computational necklaces, as one might say.

There is no magic about this, just as there is no magic about the hoverfly's finding its mate. Input-peripherals (teletypes, cameras, sound-analysers) feed into the internal computations, which lead eventually to changes in the output-peripherals (VDU-screens, line-plotters, music-synthesizers). In between, the program causes a host of things to happen, many symbols to be manipulated, inside the system itself. At the level of the machine code, the effect of the program on the computer is direct, because the machine is engineered so that a given instruction elicits a unique operation. (Instructions in high-level languages must be converted into machine-code instructions before they can be obeyed.)

A programmed instruction, then, is not merely a formal rule. Its essential function (given the relevant hardware) is to make something happen. Computer programs are not 'all syntax and no semantics'. On the contrary, their inherent causal powers give them a toehold in semantics.

In much the same way, the causal powers of the hoverfly endow its 'mind' with primitive meanings. But because the fly's internal computations are not complex enough to enable it to plan, or even to react to changes in another hoverfly's flight-path, its meanings are neither diverse nor highly structured. It is, as we noted before, not very bright. Indeed, its computational powers are so limited that we may refer to its mind only in scare-quotes, as I just did.

Because Searle assumes the utter emptiness of programs, he draws the wrong analogy. At base, a functioning program is comparable with Searle-in-the-room's understanding of English, not Chinese.

A word in a language one understands is a mini-program, which causes certain processes to be run in one's mind. Searle sees Chinese

characters as meaningless doodles, which cause nothing (beyond appreciation of their shape) to happen in his mind. Likewise, he hears Chinese words as mere noise. But he reacts in a very different way to words from his native language. English words trigger a host of computational procedures in his head: procedures for parsing grammatical structures, for accessing related ideas in memory, for mapping analogies, for using schemas to fill conceptual gaps . . . and so on. And some English words set up computations that cause bodily actions (for example, 'Pass the slip of paper out of the window').

To learn English is to set up the relevant causal connections, not only between words and the world ('cat' and the thing on the mat) but between words and the many non-introspectible processes involved in interpreting them. The same applies, of course, to Chinese (which is why Searle-in-the-room's rule-book must contain rules for parsing, interpreting, and constructing sentences in Chinese). Moreover, the same applies to any new (AI-like) language which the people outside the room may teach Searle, in order that he should understand the rule-book. He would have to learn to react to it (automatically) as the electronic machine is engineered to do.

Where does our discussion of the empty-program argument leave us, with respect to the fourth Lovelace-question? Could a computer *really* understand anything at all?

Certainly, no current 'question-answering' program can really understand any natural-language word. Too many of the relevant causal connections are missing. BORIS, for example, does not really understand why (in the story mentioned in Chapter 7) Paul phones his friend Robert for legal advice on discovering his wife's infidelity. It does not really know what a telephone is, still less what lawyers, friendship, and jealousy are.

But BORIS does have the beginnings of an understanding of what it is to compare two symbols, and of what it is to plan and to parse. You may want to say that it cannot really do these things, either – that it cannot really interpret plans, nor really parse sentences. But the complaint that it cannot parse is surely dubious. As for plans, BORIS has access to some of their abstract structural features, such as means–end relationships and the possibility of cooperation or sabotage at certain points. In short, the understanding that BORIS possesses is of a very minimal kind: how to compare two formal structures, for example, or how to build a new one by using certain hierarchical rules.

Perhaps you feel that the 'understanding' involved in such a case is so minimal that this word should not be used at all? So be it. For the purpose of explaining *human* creativity, we need not ask 'When does a computer (or, for that matter, a hoverfly) understand something?'

Moreover, this question is ill advised even if we are interested in computers as such, for it misleadingly implies that there is some clear cut-off point at which understanding ceases. In fact, there is not.

The important question is not 'Which machines can understand, and which cannot?', but 'What things does a machine – whether biological or not – need to be able to do in order to be able to understand?'³ These things are very many, and very various. They include not only responding to and acting on the environment, but also constructing internal structures of many different kinds.

Hoverflies can do relatively few of these things, which is why their intellectual powers are unimpressive. Current computers can do a few (others) too, but they lack the ways of situating themselves in a meaningful external world which even hoverflies possess. Robots with efficient senses and motor organs would be more like hoverflies. But they would need many different world-related goals, and ways of constructing new ones, to be likened to mammals.

In general, the more types of concept and conceptual space that can be built, and the more flexibly and fruitfully they can be combined, explored and transformed, the greater the understanding – and the greater the creativity. Computational psychology has provided a host of theoretical ideas with which to consider novel combinations of concepts, and with which to map the conceptual spaces constructed within human minds. Moreover, *pace* Searle, it even helps us to see how creative understanding is possible at all.

THE THIRD WAY in which people commonly deny the possibility of 'real' creativity in computers is to appeal to the consciousness argument. 'Creativity requires consciousness,' they say, 'and no computer could ever be conscious.'

We have seen, time and time again, that much – even most – of the mental processing going on when people generate novel ideas is not conscious, but unconscious. The reports given by artists, scientists, and mathematicians show this clearly enough. To that extent, then, this argument is misdirected.

But it does have some purchase, for the concept of creativity includes the notion of positive evaluation, which typically involves the deliberate examination and modification of ideas. (Remember Kekulé's adjustments to the benzene ring, in the light of valency?) In the four-phase account of Poincaré and Hadamard, this is the phase of 'verification'. Our discussion (in Chapter 4) of the development of self-conscious

reflection in children indicated how crucial to creativity this is. To be creative, one must be able to map, explore, and transform *one's own mind*.

It follows that computers can be creative only if, in this sense, they can be conscious. 'Well then,' you may say, 'our imaginary objector must be right. After all, no mere tin-can computer could be *conscious*!'

Wait a minute. 'Conscious' (like 'consciousness') is a word with several different meanings. The sort of consciousness that is essential for creativity, because it is involved in the very definition of the term, is self-reflective evaluation. A creative system must be able to ask, and answer, questions about its own ideas. Are elliptical orbits – or benzene rings, or modulations to a distant key – acceptable, given the relevant constraints? Are they more illuminating than the previous ways of thinking, more satisfactory than any alternatives currently in mind? Is it interesting – or even intelligible – to refer to thrushes' eggs as 'little low heavens', or to compare sleep with a knitter, or a bath?

Such questions can certainly be asked by computer programs. Lenat's program AM, for example, asks whether its newly-generated categories (multiplication, primes, maximally divisible numbers . . .) are mathematically 'interesting'. If it decides that they are, it explores them further (as well as reporting them in its print-out). DALTON and related models of scientific discovery examine their newly-produced conceptual structures in various ways, representing verification of different kinds. The combined ARCS-ACME system can reflectively assess the strength of the literary or scientific analogies it has generated. And COPYCAT can evaluate and compare the interest of the many different analogies it produces for one and the same thing. Presumably, the computer systems of the future will be able to represent, and so to examine, their novel ideas in even more subtle ways. In *this* sense of the term, then, there seems no reason in principle why a computer could not be conscious.

There are several other senses of 'conscious'. Some of them can (like self-reflective thinking) be accepted within a computational psychology. And some puzzling facts about consciousness can even be explained in computational terms.

For example, consider the strange non-reciprocal co-consciousness sometimes found in clinical cases of 'split personality'. In one famous case, the sexually aware personality known as Eve Black had conscious access to the thoughts of the demure Eve White, and even commented spitefully on them to her psychiatrist. It was as though Eve Black could introspect Eve White's mind – except that Eve Black vehemently denied that it was *her own* mind she was looking into, so 'introspection' hardly seems the right word. But Eve White, like the rest of us, had conscious access to nobody's thoughts but her own. Consequently, Eve Black could

– and did – play highly embarrassing tricks on Eve White, gleefully recounting them later to her doctor.⁴

How are such things possible? Taking the notion of conscious access for granted (as theoretical psychologists usually do, if they use the notion at all), this sort of psychological phenomenon can be understood in computational terms. Broadly, we can think of the two ‘personalities’ as different modules of one overall computational system, alternately controlling the same motor facilities and sensory apparatus (the patient’s body), and having different degrees of access to and control over each other’s current processing or memory-store. This does not explain why the psychological dissociation happened in the first place, what events in the person’s life-history caused it. But the problem here was to understand *how it is possible at all*, not what triggered it.

In considering our original question, however, the notion of conscious access cannot be taken for granted. It involves, among other things, the problematic idea of felt experience, or sensation: that introspective *something* that is so remarkably difficult to describe, but which we all know in our own case. (Just try pinching your arm!) What of that?

Well, what of it? It’s not obvious that consciousness in *that* sense is essential to creativity. It’s not obvious, either, that the various *somethings* can actually be distinguished independently of their causal relationships with other internal processes and with things in the world outside. If they cannot, then perhaps some conceivable computer could (as we say) experience them. (A machine capable of feeling would doubtless be very unlike the computers we know today, but that is another matter.)

Our previous discussion of the brain-stuff argument remarked that the relation of mental phenomena to neuroprotein, as such, is utterly mysterious. So why shouldn’t some future tin-can have feelings and sensations, too? Why shouldn’t there be *something which it is like to be* that computer, just as there is *something which it is like to be* you or me – or a bat?⁵ Admittedly, it seems intuitively unlikely, perhaps even absurd. But intuitions are not always reliable.

The crucial point here is that we understand so little about this particular sense of consciousness that we hardly know how to speak about it, still less how to explain it. When we say with such confidence that we have consciousness, we do not know what it is that we are saying. In these circumstances, we are in no position to prove that no computer could conceivably be conscious.

In sum, *if* creativity necessarily involves conscious experience over and above the self-reflective evaluation of ideas, and *if* no computer could have conscious experience, then no computer could ‘really’ be creative. But these are very iffy ifs. The question must remain open – not just

because we do not know the answer, but because we do not clearly understand how to ask the question.

WHAT OF THE fourth way of denying ‘real’ creativity to computers: the non-human argument? Unlike the brain-stuff argument, this is not a scientific hypothesis. And unlike the empty-program argument and the consciousness argument, it is not a disinterested philosophical debate. Rather, it is the adoption of a certain attitude towards computers, an uncompromising refusal to allow them any social roles like those enjoyed by people.

I suggested in Chapter 7 that if a future version of AARON were to draw acrobats with ‘triangular’ calves and thighs, art-connoisseurs might refuse to accept these as aesthetically valuable. They could hardly deny the analogy between limb-parts and wedges. But they would dismiss it as uninteresting, even ugly. And I explained: ‘In their view, it is one thing to allow a human artist to challenge our perceptions, and upset our comfortable aesthetic conventions, but quite another to tolerate such impertinence from a computer program.’

What has impertinence to do with it? Well, to be impertinent is to say something (perhaps something both true and relevant) which one has no right to say. A person’s right to be heard depends largely on their social status and topic-specific authority. These distinctions are necessary, because we cannot attend equally to everything that anyone says about anything.

At a certain level of generality, however, everyone has some authority. We all have aims, fears, and beliefs, each of which – unless the contrary can be specifically shown – deserves to be respected. Everyone has a right to be heard, a right to try to persuade others, and a right to further their interests. Each of us has these rights merely by virtue of being a member of the human community.

And that is the point. Computers are not automatic members of the human community, in the way that members of the biological species *homo sapiens* are. They are not even members of the animal community, like dogs – or hoverflies. If they are not members of our community, then they have none of our rights. So ‘impertinence’ makes sense.

However, automatic membership is only one way of entering a community: someone can be invited to join. It is up to the community to decide who – or what? – is acceptable.

Prima facie, the science-fictional computers we are discussing would have a strong claim to honorary membership in human conversational

groups. For they could do many humanlike things – sometimes, better than us. Many social functions – story-teller, jazz musician, financial adviser, marriage-counsellor, psychotherapist – could be carried out by these non-biological ‘intelligences’. So we might decide to do away with the scare-quotes entirely, when using psychological words to describe them.

But this decision, to acknowledge computers as *really* intelligent, would have far-ranging social implications. It would mean that, up to a point, we should consider their interests – much as we consider the interests of animals. For interests, *real* interests, they would be assumed to have.

Many current programs set up goals and sub-goals, and try to achieve them in various ways. (It is because we discovered that the hoverfly patently does not do this, that we refuse to credit it with intelligence.) The highly-advanced systems we are imagining would be able to do this too – and to ask for our cooperation. Suppose a computer-poet, tussling with a new composition, requested a good analogy for winter, or asked you to check whether thrushes’ eggs are blue. If you had accepted it as a genuinely intelligent creature, you would be bound, within reason, to interrupt what you were doing in order to oblige it. (Even insects sometimes benefit from our fellow-feeling. You might squash an irritating hoverfly without much compunction, but would you deliberately pull a hoverfly to pieces, just for fun? And have you never got up out of your comfortable chair to put a ladybird into the garden?) I leave it to you to imagine scenarios in which more problematic conflicts of interest might occur.

Similarly, to regard computer-systems as *really* intelligent would mean that they could be deceived – and that, all things being equal, we should not deceive them. It would mean, too, that they could *really* know the things they were apparently saying, so we could *really* trust them.

These two examples have already come up in the English law-courts. Ironically, the court’s refusal in both cases to acknowledge real deception or knowledge seems unsatisfactory. The first example (which involved a mechanical device, not a computer) concerned a man who had lifted the ‘arm’ of a car-park machine without putting any money in. The magistrate acquitted him, on the grounds that to commit fraud one must deceive someone, and ‘a machine cannot be deceived’. In the other example, the defendant was accused of stealing banknotes. The prosecution submitted a list of banknote-numbers, some of which matched notes found in his possession. In law, documents accepted as evidence must be produced by someone ‘having knowledge of their contents. But the crucial list had been produced by the bank’s computer. Because a computer (so the judge said) cannot have any knowledge of anything, the accused was acquitted.

My aim in giving those examples was not to start you on a life of crime, nor even to show that the law is an ass. The law will presumably be changed in some way, so as to prevent such absurdities. But just how should it be changed? Would you advise today's judges to accept that computers *can* have knowledge of the documents they produce? What about the judges living in the futuristic society we have been imagining? Would you be happy for one of that society's 'creative' computers to adjudicate such tricky legal decisions?

Whatever your answers, the crucial point is that the decision to remove all scare-quotes, when describing programs in psychological terms, carries significant moral overtones. So, like moral decisions in general, it cannot be forced upon us by the facts alone.

To answer 'No!' to the fourth Lovelace-question (on these grounds) is to insist that, no matter how impressive future computers may be, we must retain all the moral and epistemological authority, and all the responsibility too. There can be no question of negotiating with computer programs, no question of accepting – or even rejecting – their advice: the role of *adviser* is barred to them. Quasi-intelligent, quasi-creative, programs would be widely used, much as pocket-calculators are today. But it would be up to us to take the entire responsibility for relying on their 'knowledge', and for trusting their 'advice'.

Whether people actually would answer 'No!', in the imaginary situation we are discussing, is not certain. Their answer might even hang on mere superficialities. For our moral attitudes and general sympathies are much influenced by biologically-based factors, including what the other person – or quasi-person – looks like, sounds like, and feels like.

Fur or slime, cuddliness or spikiness, naturally elicit very different responses. Walt Disney profited from the universal human tendency to caress and protect small animals with extra-large heads and extra-large eyes. Even robots can profit from this tendency. You may remember the two robots in the film *Star Wars*: the life-sized golden tin-man, C-3PO, and the little, large-headed, R2D2. At one point in the story, R2D2 toddles after C-3PO as fast as he can on his two little legs, calling to him in his squeaky voice 'Wait for me!' When I saw the film, a chorus of indulgent 'Ooh's and 'Ah's rose spontaneously from the cinema audience around me. So if our futuristic computers were encased in fur, given attractive voices, and made to look like teddy-bears, we might be more morally accepting of them than we otherwise would be.⁶ If they were made of organic materials (perhaps involving connectionist networks constructed out of real neurones), our moral responses might be even more tolerant.

You may have little patience with this science-fictional discussion. You may feel that we cannot know, now, what we would do in such a

hypothetical situation. You may argue that we cannot even imagine it clearly, never mind decide what a morally appropriate response would be.

If so, I sympathize with you. The fourth Lovelace-question is, in large part, a disguised call for a complex moral-political decision concerning a barely conceivable situation. I have offered firm answers to the first three Lovelace-questions. As for the fourth, it can be left undecided.

SCIENCE IS WIDELY believed to destroy not only romance (converting the wonderful into the prosaic), but freedom as well. The deck-chairs of creativity lie alongside those of freedom: if the tide covers one, it covers all.

Blake's reference to freedom, in the passage cited above, is only one of many expressions of this view. Over a century before, on realizing that mankind is but a slender reed in the terrifying spaces of the newly-conceived scientific universe, Blaise Pascal had consoled himself by noting that it is a *thinking* reed – and part of what he had in mind was our ability to make free choices. Our subjectivity, our ability to think of things and choose actions that have never been, and to construct the world of the mind over and above the external environment, was for him our saving glory. This, at least, lay beyond the cold impersonal touch of science.

But freedom, too, can be touched by science without being destroyed by it. A computational psychology can allow that much human action is self-generated and self-determined, involving deliberate choices grounded in personal loyalties and/or moral principles. Indeed, it shows us how free action is possible at all.

To a significant extent, arguments defending freedom against science parallel arguments pitting creativity against a scientific psychology. Worries about predictability and determinism, for example, crop up constantly in such discussions. The arguments detailed in Chapter 9 with respect to creativity apply, *pari passu*, to freedom too.

In either case, pure indeterminism gives mere chaos. To choose one's actions freely is not to hand responsibility over to the unpredictable. Luke Reinhardt's novel *The Dice Man* shows how un-free, and un-human, a life would be in which most choices were determined by chance. And in either case, predictability may be a virtue. One would not go for advice on a moral problem to someone whose judgments are not usually reliable, nor admire someone whose good deeds were always grounded in passing whims.

Sensible, responsible, thinking requires a highly structured conceptual

space, and a discriminating exploration of the possibilities involved. Thoughts triggered by chance (R-random) events, in the environment or within the brain, may be fruitfully integrated into the psychological structures of the mind concerned. If so, all well and good. If not, they are either mere idle irrelevancies, or (sometimes) tyrants forcing us to venture beyond the space of our self-determination.

An obsessional idea is a tyrant inside the mind. It monopolizes the person's attention, bypassing many of the self-reflective computations involved in genuinely free action. Acting responsibly involves careful deliberation, not enthusiastic spontaneity.

You may recall William Golding's remark, that some incidents in his novels came to him rather than from him: 'I heard it. . . . [At such moments] the author becomes a spectator, appalled or delighted, but a spectator'. One might 'hear' a recommendation for action, too: many people have. But if matters of any moral importance are concerned, one should think hard before following it. People who claim to be obeying inner voices are surrendering their responsibility to someone supposedly more worthy (this is one way in which a tyrannical idea can bypass the normal deliberative channels). For Joan of Arc, the authority-figures were St. Margaret and St. Catherine; for the mass-murderer known as the Yorkshire Ripper, the voices seemed to come directly from God. (Hypnosis puts temporary tyrants into the mind; but hypnotists can rarely, if ever, force people to act in ways that conflict with their most fundamental evaluative principles.)

There are tyrants outside the mind, too. Someone who threatens you with imprisonment, or who holds a gun to your head, gives you such an overwhelming (even urgent) reason for doing what he commands that your normal reasoning can get no purchase. You are not so constrained as the theatregoers who, because of some computational short cut in their minds, run without thinking when they hear someone behind them shout 'Fire!' You have more freedom than they, because you do have the computational capacity (especially in a non-urgent situation) to choose to disobey the tyrant. But the sanctions involved are so extreme, and your fear probably so great, that you are most unlikely to choose the action you know to be right. What in normal circumstances is unremarkable, in these situations would be heroic.

Free choice is structured choice, not mere mental coin-tossing. Even the *acte gratuit* of the existentialist (as depicted by Albert Camus in *L'Etranger*, for instance) is purposefully generated, to make a philosophical point. But the philosophy concerned is mistaken. Human freedom is not wholly unconditioned, any more than creative 'intuition' is.

Our ability to perform apparently random, motiveless, acts is

undeniable – but only if they are done with some purpose or other in mind do they count as *acts* in the first place. Someone who kicks the cat while having an epileptic fit is not only not acting freely: their kicks are not actions at all. Nor are the hoverfly's journeyings to meet its mates. Actions must be generated by certain sorts of computational structure (goal-hierarchies, for example) in the mind. They have a complex psychological grammar, whereas epileptic kicks do not. Camus' hero, who commits suicide as a defiant gesture in face of what he sees as the absurdity of the world, is indeed acting freely. That is, he is making a fundamentally self-determined choice, as opposed to following unexamined habits and unquestioned principles.

The person who acts out of 'bad faith' does not take the trouble to explore – still less, to transform – the conceptual space within which their habitual actions are situated. Sartre's waiter plays the waiter's role (follows the waiter-script), without ever considering any other possibilities. Someone who does not even ask 'What else could I do?' is rather like the nineteenth-century chemist who takes it for granted that all molecules must be strings. The question of their not being strings does not even arise, for how *could* they be anything else? (It does not follow that we should constantly raise questions about every aspect of our habitual behaviour; as noted in Chapter 5, to dispense with role-scripts entirely would be to drown in computational overload.)

Choices that arouse our world-weary cynicism are like hack novels, written in an undemanding literary style and displaying a shoddy sense of priorities. Choices that take us, with some subtlety, down previously unexplored pathways already marked on familiar moral maps are like the pleasing improvisations of a jazz-musician who can play in only one style. Choices that lead to fundamentally unexpected actions, arousing a shock of admiration or contempt, are comparable to Kekulé's creative transformation of an accepted chemical constraint. And someone who changes the basic topography of our moral landscape (for instance, by considering the negative and saying 'Love your *enemy*') is like someone who composes music in a radically new style.

All these are examples of free choice. But, as George Orwell might have put it, some are more free than others. We have seen throughout this book that it is usually unhelpful to ask 'Is that idea creative: *yes* or *no*?', because creativity exists in many forms, and on many levels. Much of the interest, and the illumination, lies in the details. So, too, it is usually unhelpful to ask 'Was that action free: *yes* or *no*?' Action in general, of which the generation (and evaluation) of new ideas is a special case, has a highly complex psychological structure. What may count as 'free' (or 'creative') in light of one of the many relevant structural

aspects may count as relatively 'unfree' (or 'unimaginative') in light of others.

A PERSON'S PSYCHOLOGICAL structure includes their myriad beliefs, goals, anxieties, preferences, loyalties, and moral-political principles – in a word, their subjectivity. Each of us is different from every other human being, and these differences influence what we do, and how we think, in systematic ways. We all construct our lives from our own idiosyncratic viewpoints.

In Newton's vision, however, we are all alike. Indeed, the magnificence of his theory lay in its ability to see apples, tides, and planets – and human bodies, too – as instances of one unifying principle. Even the clinical anatomist or biochemist, who makes distinctions between one person and another, is not professionally interested in *personal* distinctions. The natural sciences (including neurophysiology) are not concerned with personal, subjective, phenomena. These cannot even be described in natural-scientific vocabulary – and what cannot be described is very likely to be ignored, or even denied. It is hardly surprising, then, that science *as such* is widely seen as inevitably dehumanizing.

It does not follow that *no* science can admit our individual differences, or explain them in terms of general principles of mental function. But the concepts used in such a science must be able to describe ideas, and subjective thought. They must be able to depict the mental processes that generate our idiosyncratic representations of the world, including our ideas of other people's minds and meanings.

This, as we have seen, is what a computational psychology can do. It cannot hope to rival the novelist's eye for human detail, or the poet's insight into human experience. But that is not its job. The task of a scientific psychology is to explain, in general terms, *how such things are possible*.

Over a century ago, people were asking how the diversity of biological species, and the layered order within the fossil record, are possible. Darwin criticized the cramped imagination of those who could wonder only at piecemeal special creations. Understanding the body's laws of harmony, he said, provided a more magnificent view.

Similarly, to attribute creativity to divine inspiration, or to some unanalysable power of intuition, is to suffer from a paucity of ideas. Even to describe it as the bisociation of matrices, or as the combination of familiar concepts in unfamiliar ways, does not get us very far. But, surprising though it may seem, our imagination can be liberated by a

computational psychology. The explanatory potential of this approach is much – very much – more than most people imagine.

The Loom of Locke and the Water-Wheels of Newton had no room for notions like creativity, freedom, and subjectivity. As a result, the matters of the mind have been insidiously downgraded in scientific circles for several centuries. It is hardly surprising, then, if the myths sung by inspirationists and romantics have been music to our ears. While science kept silent about imagination, anti-scientific songs naturally held the stage.

Now, at last, computational psychology is helping us to understand such things in scientific terms. It does this without lessening our wonder, or our self-respect, in any way. On the contrary, it increases them, by showing how extraordinary is the ordinary person's mind. We are, after all, humans – not hoverflies.