This Chancy, Chancy, Chancy World

Mir Publishers
Moscow

L. Rastrigin

Л. Растригин

ЭТОТ СЛУЧАЙНЫЙ, СЛУЧАЙНЫЙ, СЛУЧАЙНЫЙ МИР

Издательство
«Молодая гвардия»
Москва

This Chancy, Chancy, Chancy World

by L. Rastrigin

Translated from the Russian
by R.H.M. Woodhouse

Mir Publishers
Moscow

First published 1973
Second printing 1984

На английском языке

Have you ever sat down and thought about how often chance affects your life? If you have, then you probably realize that chance literally hits us from every side.

We live in a world more vulnerable to the vicissitudes of chance than the wildest imagination could devise.

Chance abounds in an endless variety of forms. Some darken our existence, confound our plans and prevent us from realizing our most cherished ambitions. Others do not affect us, while others still illuminate our lives with all the colours of the rainbow and bring happiness and success (eureka!).

But is it really worth talking about chance? What is there to say about it? Chance is chancy, and that's that.

In fact there is a great deal we can say about chance and there is even more we can ask about it.

For example: how does chaos arise? What is control? How should we act in circumstances involving chance? How can we come to terms with the difficulties that arise from chance obstacles in our lives? What is the Monte Carlo method? Why is learning necessary? What role does chance play in evolution and progress? How is it that our chancy, chancy, chancy world gets along quite well? Is it possible to make it better still?

Answers to these and many other questions will be found in this book.

Contents

Part II. Welcome Chance

WHAT IS CHANCE ?

(IN LIEU OF A PREFACE)

Before setting off on our journey into the world of chance I thought it might be well to find out how chance is defined by various authorities. Turning first to my usually excellent philosophical dictionary, I was immediately disconcerted by the following brief, but bold, entry: 'CHANCE. See: Necessity... .' (For those who doubt my veracity—and are able to read Russian—the offending edition is: Philosophical Dictionary, Moscow, 1968, page 323.) Thrusting my doubts aside for the moment, I pressed on. But imagine my dismay when the article referred to proceeded to identify chance with inessentiality: 'Chance does not arise from the essence of a phenomenon... .'

After such an unambiguous statement in a philosophical dictionary, it seemed out of the question to write a book about chance. After all, a popular science book should deal only with essential things (and not haphazard ones). Quite frankly, with thoughts like that I was ready to throw my hand in. Was it really worth writing a book about chance?

However, gritting my teeth, I tried another of my 'authorities', this time Clement Timiryazev. His absolutely definite, and alas! widely held, view of chance likewise brought little joy. His text reads: '... what is chance? An empty word used to conceal ignorance, the dodge of a lazy mind. Does anyone really suppose that chance exists in nature? Is it even possible? Is it possible for any action to take place without a cause?' (*A Concise Exposition of Darwinian Theory*). True, Timiryazev can be excused on the ground that he lived at a time when some modern sciences were still in their infancy and most of them had not even been thought of.

After this cavalier treatment of chance it was impossible *not* to write a book on the subject, if only to rehabilitate it. Otherwise, as things stood, the theory of probability was to be numbered among the pseudo-sciences, and mathematical statistics to be regarded as 'the dodge of a lazy mind'.

In both the extracts quoted above, chance seems to be regarded as something improper and shameful, something not talked about in polite society. Behind them both there is obviously some such educative thought as this: 'Leave nasty Chance alone, children, ugh! don't touch it: you'll only get your hands dirty! Here, come and play with this—good old Certainty. See how nice and clean and straightforward it is. There now, that's the way.' And with this sort of upbringing the child soon comes to believe that there is something 'not quite right' about chance; whereas certainty—ah, yes! that's what we are after. If you got to where you were going, that was strictly in accordance with certainty; but if you slipped on the way and broke your nose, then chance was entirely responsible.

This view of chance reminds one of a weighted coin: only one side of the coin is ever seen, and it happens to be the unpleasant, irksome one. Unfortunately, the thorny path of human progress is strewn with the coin of chance mostly showing just this side (later we shall see why). We have inherited a pessimistic view of chance because of the multitude of broken noses with which the whole history of suffering humanity is forever scarred.

What part does chance play in our lives? If you have thought about this at all, you are sure to have noticed how much our lives depend on chance. Chance occurrences hit us from every side.

In science and technology chance has generally been regarded as an enemy—an irritating obstacle to accurate investigation. Chance obscures our view of the immediate future and makes prognostications for the more distant recesses of time quite impossible. (Let us but remember the sad reputation of our weather prophets.) Chance interference not only impedes, but often completely severs, connections between widely separated points. Even in everyday life chance gives rise to a great deal of unpleasantness.

The age is long since past when man first took up arms against chance. Man's war with chance has always been fought on two fronts. The first is characterized by attempts to discover the causes of chance occurrences with a view to eliminating chance itself entirely. For example, until recently, it was thought that the sex of a baby was a matter of pure chance. But geneticists have now unravelled the principles of sex-determination. What the geneticists have done, in effect, is to extract from nature one of her secrets, thereby annihilating an element of chance, which then stands revealed as having been nothing more than a cloak for our ignorance.

Similar situations crop up in life and in science with great frequency. It was just this state of affairs that impelled Timiryazev to make his angry pronouncements. True, he identified chance with causelessness—and these are by no means the same thing.

In point of fact, every event has a completely definite cause; that is, every event can be regarded as the effect of its cause. Every chance event has such a cause. That cause in its turn is the effect of some other cause, and so on. There is no particular difficulty involved when the chain of causes and effects is simple and obvious and can be examined with ease. In such a case the end result cannot be regarded as a chance event. For example, if we are asked whether a tossed coin will land on the floor or the ceiling, we can give a perfectly definite answer since it is perfectly obvious what will happen and chance has nothing to do with it. But if the cause-effect chain is complex and parts of it are hidden from view, the event becomes unpredictable and is said to be a chance event.

Suppose, for example, we want to know whether our tossed coin will turn up heads or tails. Here it is still possible to write an accurate description of the chain of causes and effects. However, we should then have to investigate such factors as: the pulse-rate of the person tossing the coin, his emotional state, and so on. Practically speaking, it is impossible to carry out such an investigation because we do not know, for example, how to measure the emotional condition of a person tossing a coin. So although the cause is nonetheless there, we are still unable to predict the result. Here the complexity of the cause-effect chain makes the event unpredictable—in other words, a chance event.

But what exactly is an *unpredictable* event? Can we really say nothing about it? Are we obliged to throw our hand in every time we meet up with chance?

No, of course not. A long time ago people began to notice that chance is endowed with certain properties, and that a great deal could, in fact, be said about any 'unpredictable' occurrence. For instance, in the light of our experience of tossing coins we are now able to state that roughly half the results will be heads and half tails. It follows that chance can and should be investigated. And indeed, as far back as the seventeenth century the beginnings of the theory of probability—the mathematical study of chance events—are already discernible.

This last constitutes the second front of man's struggle with chance. Here, the aim is to study the laws governing chance events. The investigation of these laws does not make the individual chance event any the less chancy: what it does do is to provide us with a clear picture of the inner structure of chance. Knowledge of this structure enables us to conduct an effective battle against the unpredictability of chance events.

Investigations like this are directed toward reducing the role of chance in science and technology and in the ordinary life of the community. A great number of methods have been devised that permit of either the total exclusion of chance or at least a lessening of its destructive consequences. One of the most interesting and most important problems of this kind concerns the separation of a useful signal from a mixture containing chance interference ('noise') along with the necessary signal. (In everyday life we do a fair job of solving a mass of similar problems at almost every step, even though we do not pause to consider how we do it.) In this book we shall have a look at the most interesting and useful methods available for diminishing the role played by chance.

So far, we have spoken only of the tiresome side of chance, the side that clutters our lives with uncertainty, despondency and alarm. But it has long been observed that there is more than this to chance and that chance has a happy, useful and desirable side to its character.

Whereas once upon a time people were content to record beneficial chance occurrences and to marvel at their good fortune, nowadays more and more attempts are being made to put chance to work and to make it serve mankind. Apparently, the first people to understand and to employ the usefulness of chance were those engaged in the selective breeding of new plants and new varieties of livestock, poultry, and fish.

More recently, engineers have begun taking an interest in chance and have succeeded in producing a series of wonderful machines possessing extraordinary capabilities because they have an element of chance incorporated in their design.

Economists and military men have also learned to understand and appreciate the importance and usefulness of chance in solving problems that require the selection of the best course of action in situations involving conflict. They have discovered that the best course of action is generally one that relies on chance.

In this book we shall examine the most important of chance's useful applications.

Chance is not a passive thing: it plays an active part in our lives. On the one hand it confounds our plans, while on the other it presents us with new opportunities. It is hard to overestimate the influence of chance on nature and on our lives. We need only remember that life itself originated in a series of chance events. In nature chance follows its own laws and is inescapable. It can be blind: it can also be remarkably perspicacious. Chance destroys just as inevitably as it creates; it arouses regret just as often as delight; it hinders and at the same time helps.

The two-edged sword of chance is thus an unusually misleading, not to say dangerous, partner in man's struggle against the blind and menacing forces of nature.

This book is dedicated to chance in both its aspects: chance the obstructor—and chance the helper; chance the destroyer—and chance the creator; chance the enemy—and chance the friend.

CNAHCE IS...

In my preface I asked 'What is chance?' without giving a direct answer. There were two reasons for this.

The first has to do with the style of popular science books—a style this writer has no intention of abandoning. The idea of this style is that the writer begins by asking some apparently elementary question and allows a certain amount of fog to gather around it. Then he proceeds to show that things are not quite so simple as they seemed and that they are, in fact, just the opposite: the problem turns out to be extremely complicated and in no sense elementary. After this the writer is supposed to proceed with measured step and a large number of fascinating, and at the same time comprehensible, examples to shock the surprised reader into an awareness of his or her ignorance of the subject. And then he is ready to set forth the current state of knowledge in the given field.

The second reason is much more important. The plain fact is that scientists and scholars have not been able to reach any agreement as to how chance should be defined. Consequently, any writer who takes it upon himself to answer this question runs the risk of inviting a host of unflattering remarks from his colleagues. In this case the writer finds himself in a most uncomfortable situation because he has to state his opinion. So, summoning up all our reserves of courage, let us try to answer the question 'What is chance?'

Well then, chance is first and foremost ... *the unpredictability that is due to our ignorance*: to our being badly informed, to the absence of necessary data, and to our lack of essential knowledge.

Thus defined, chance is essentially a measure of ignorance: the less the information we possess about an object the more chancy is its behaviour. Conversely, the more we know about an object the less is its behaviour a matter of chance, and the more definite we can be in predicting its future behaviour.

From this point of view Timiryazev was perfectly correct. Reference to the chance nature of this or that fact or process—given this definition of chance—is an affirmation of ignorance and of the investigator's incompetence in whatever the particular question happens to be.

Let us construct a model for this conception of chance. We shall represent the cause-effect relationship of an event graphically by means of a circle and two arrows (Fig. 1). The cause of the event is represented by the white arrow entering the circle from the left,

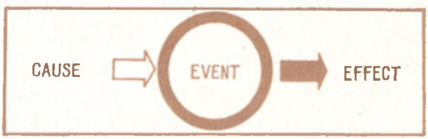


Figure 1

and its effect by the black arrow leaving the circle on the right.

We shall have occasion to use transformations of this type at every stage of our discussion. When we ring somebody's front doorbell, we press the button beside the door and thus initiate a cause that produces an effect: the ringing of the bell inside the apartment (Fig. 2).

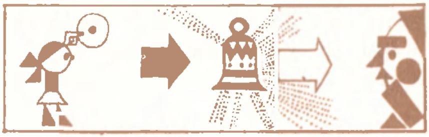


Figure 2

If we want to light a burner on our gas stove, we have to initiate two causes. We have to turn on the tap controlling the flow of gas; and we have to hold a lighted match near the burner. These two causes will produce the effect, namely: the lighting of the burner (Fig. 3).

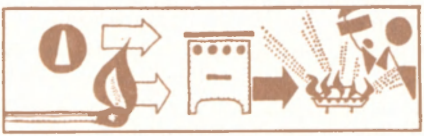


Figure 3

This method of depicting cause-effect relationships is extremely convenient and is widely used in cybernetics (Fig. 4). Here a signal A causes a signal B to

Figure 4

appear. The relation between the signals A and B takes the form of a transformation that produces signal B at its output as soon as signal A is fed into its input. In symbols:

$$A \rightarrow B$$

Let us return to the example of the doorbell. Can we be quite sure that whenever we press the button, the bell will always ring? Of course not. Before we can say that our summons will be heard we have to know whether the power in the circuit containing the bell is switched on. In other words, two conditions are necessary if the bell is to ring: the circuit must be supplied with electricity, and the button must be pressed. Only if these conditions are fulfilled will the ringing of the bell be a complete certainty.

But as we approach the front door we do not know whether the power in the bell circuit is on. This makes the ringing of the bell a chance event, so far as we are concerned, because we lack certain information. If we telephoned our host beforehand to ask whether the doorbell was working—in other words, if we obtained the necessary information—then to us the events *press the button* and *the bell rings* would be connected in a rigorously determinate way, and chance would not come into the picture at all.

And so we see that in this example an element of chance is usually present for the simple reason that of the two conditions necessary for the bell to ring, A_1 and A_2 (pressing the button and supplying power to the circuit), only one is usually observable, namely: pressing the button. The bell's chancy behaviour is due to the uncertainty regarding the second condition (Fig. 5). (Here we have tacitly assumed that the bell

Figure 5

itself is in working order. If we cannot assume this, we have yet another source of chanciness.)

The same sort of situation arises when we go to light the gas stove (Fig. 6). In addition to the two cau-

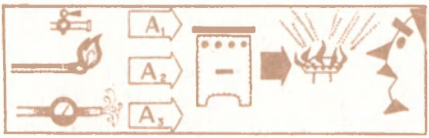


Figure 6

ses we have already treated, A_1 and A_2 (turning on the tap and lighting the match), it is essential that we know the state of a third cause A_3—the presence of gas pressure in the mains or in our gas cylinder as the case may be. If we took the precaution of telephoning the gas authority and making certain the supply was in order—or if we measured the pressure in our cylinder—then the lighting of the burner would not be a chance event. However, if we neglect to ascertain this information, we must reconcile ourselves to the fact that the burner will not light up on every occasion: the event becomes a chance event because it is no longer one hundred percent predictable.

Thus chance is essentially a function of our level of ignorance. The more ignorant a person is, the more his particular world is subject to chance. The opposite is also true: the world as seen through the eyes of a scientist does not seem so depressingly vulnerable to the whims of chance.

We see, then, that chance is a subjective matter that depends on the amount of information possessed by the subject. If one is prepared to concede the existence of a truly omniscient god, then obviously such a being would not find the slightest trace of chance in our world. Unfortunately for the angels, however, the biblical legend endowed the god alone with omniscience: even the most senior among them—for all their unearthly holiness—were not favoured with such great powers for acquiring the necessary information. As for sinful man—how could he possibly be expected to compete with the ultra-telepathic abilities of the Almighty? Through the five slender channels linking him to the outside world (sight, hearing, smell, touch, and taste) man receives but scanty information about his surroundings. His title of Lord of Nature derives only from the resourcefulness of his brain; and it is this resourcefulness that allows him to explain the mechanism of chance roughly as follows.

Every event (B) is the consequence of a set of causes that may be either small or large in number. In Figure 7 the dots below the cause arrows, A_1, A_2, A_3 ..., indicate that the number of causes may be extended without limit as far as we please: A_4, A_5, A_6 ... and so on. To predict an event man must know exactly all the causes responsible for the occurrence of this event.

Where the causes are few in number and can all be readily observed, the event is not considered to be a chance event (it is often called a regular or deter-

Figure 7

minate event). For example, if we throw a stone into the air we can safely predict with complete accuracy that it will fall on the earth and not on the moon. Here, knowledge of the law of gravity provides all the information we need about the event to enable us to determine where the stone will come to rest.

However, if the number of causes is so large that they cannot all be grasped simultaneously (for instance, if event B requires, say, a million causes), the event can no longer be accurately predicted; and, being unpredictable, it is a chance event. Here, chance arises from insufficient knowledge, from incompetence, and from paucity of information.

Does this mean that one fine day, when we have all become terribly clever, chance will suddenly disappear from the face of the earth?

No, it certainly does not. At least three factors will prevent such a thing from ever happening—three staunch and steadfast defenders of chance.

First, there is *the infinite complexity of the world.* We shall never succeed in exhausting the endless variety of our world: we shall never find out all that there is to know about it. Any pretension to divine omniscience of this kind is more stupendous than mere belief in a god: it constitutes an aspiration to being a god oneself, omnipresent and omnipotent.

To put it more simply: there is a sort of natural ban on completely exhausting the world of its mysteries. However much we investigate it, there will always be something left behind 'in the bottom of the barrel'; for the world is truly inexhaustible. Nowhere is this limitation better expressed than in the aphorism of Kozma Prutkov: 'It is impossible to fathom the unfathomable.'

Obviously, we shall never be able to predict which way (heads or tails) a tossed coin will fall, because the fate of the coin is determined by four factors at least. They are: the person tossing the coin, the medium in which the coin falls, the surface on which it lands, and the properties of the coin itself. Each of these factors is vital to the outcome, and each in its turn is the result of a huge number of causes. In fact the number of all these causes is practically infinite, so it is unlikely that they could all be taken into account simultaneously even for a single toss of the coin.

Another sure defender of chance and of the unpredictability of our world is to be found in our *limited accuracy of measurement.*

It is well known that the accuracy with which an event can be predicted often depends on the accuracy with which its causes can be measured. But the accuracy of any measurement is limited. As science and technology develop, so our accuracy improves; but it always remains—and will always remain—finite. In other words, there is no such thing as absolute accuracy and there never will be—not even if we base our measurements on the atomic structure of matter.

This state of affairs limits the possibilities of prediction and, as a consequence, ensures the survival of chance.

For example, if we wanted to determine the point of impact of a ballistic missile we would need to know with great accuracy all of the factors influencing the missile's trajectory. Here, we would be primarily concerned with the condition of various layers of the atmosphere through which the missile would pass. However, it would be tremendously difficult—and practically impossible—to measure accurately movements of air masses in the atmosphere all the way to the target area. We would therefore have to restrict ourselves to using approximated estimates of the required factors. And this in turn would guarantee that it would be a matter of chance whether the missile hit the target, because we could not predict accurately that it would.

And so, the impossibility of being certain about hitting the target would be due to our lack of accurate information, resulting principally from the approximateness of our measurements.

Finally, chance arise not only as the result of our ignorance, of the endless complexity of our world, and of our limited measuring accuracy: it is also inherent in the famous *principle of indeterminacy* or *uncertainty** first formulated by the German physicist Werner Heisenberg.

Essentially, the uncertainty principle means that every event the outcome of which is determined by the interaction of individual atoms is of its nature a chance event. The more detailed working of this principle may be presented as follows.

---

* The Heisenberg uncertainty principle was established in 1927 and earned Heisenberg a Nobel prize in 1932.—*Tr.*

It is common knowledge that to determine any future state of a given particle in space, we need accurate values for its initial position and velocity, and no more. The uncertainty principle specifies a kind of limitation to accuracy whenever the object under investigation is a subatomic particle. Briefly, the accuracy with which we can determine one of the parameters of a subatomic particle—its position, for example—is related to the accuracy with which we determine the other parameter—velocity, say, or rather momentum. The more accurately we measure one parameter the *less* accurately are we able to measure the other. It is impossible to measure both parameters with the necessary high degree of accuracy. This is a crucial feature of the micro-world; and no advance in the technology of measurement will ever enable us to improve our accuracy beyond this limit, just as no advance in science will ever enable us to interfere actively with the past. (We should note that while it is possible to interpret the past—and this we often do—we can never change it.)

The following simple experiment is a good illustration of the uncertainty principle. Suppose we have an ordinary television picture tube. Inside the tube there is a source of electrons called the electron gun. This is simply an ordinary incandescent filament, such as we have in an electric light bulb, burning at red heat. The heated filament is a source of electrons. The electrons are accelerated by means of an electric field and then passed through two orifices placed one behind the other in the barrel of the gun. These two orifices concentrate the electrons into a narrow beam that issues from the electron gun much as a stream of bullets from a machine-gun. This narrow beam is directed towards a screen containing a layer of a special material that is sensitive to the impact of electrons.

When a single electron strikes the screen, it is scattered (loses energy), the lost energy reappearing as a tiny flash of light visible to the eye. Consequently the continuous stream of electrons in the electron beam produces a glowing spot of light on the screen. By controlling the motion of the beam by means of electric or magnetic fields, we can move this bright spot all over the screen, which is the basis of television.

However, that is not the main point at present. Suppose, now, we want to make the spot on the screen as small as possible. To do this we have to reduce the diameter of the beam of electrons issuing from the electron gun. How can we achieve this? It would seem that nothing could be simpler: all we have to do is make the muzzle of the electron gun smaller. Let us imagine that we have succeeded in making an electron gun with a variable orifice giving us a range of apertures from fairly large down to a minimum equal to the diameter of an electron (we would not need to go any smaller than this: if we did, the electrons would simply jam in the barrel). You can picture the mechanism for this as being like the diaphragm of a camera. By fitting some sort of diaphragm like this to our electron gun we can vary the diameter of the electron beam.

Now we are ready to begin our experiment. As we reduce the aperture we find at first that the spot on the screen actually does get smaller. But after a while the spot stops decreasing in size and faint rings of light form around it. As we close down our diaphragm still further, these rings spread out across the screen. And with the muzzle of the gun at its smallest (equal to one electron diameter) the spot disappears altogether and we see a series of tiny flashes appearing one after another, now here, now there, evenly distributed over the entire surface of the screen.

How can we explain the electrons' strange behaviour? One would think at first that when the aperture was at its narrowest the beam would reduce to a single stream of electrons all hitting the screen at exactly the same point; and that the diameter of the bright spot would therefore be equal to that of one electron. But in our experiment we observed nothing of the kind. Where did we go wrong?

The fact of the matter is that the hoped-for result contradicts the uncertainty principle that we have been talking about. What happens is this: as we reduce the aperture of the electron gun the error in determining the position of the moving electrons becomes less and less. This error is equal to the difference between the diameter of the orifice and the diameter of an electron. As the orifice becomes smaller, the error tends to zero; so that when the orifice has the same diameter as an electron the position of the electrons is precisely determined. At the moment of passing through the orifice the electron is located within the orifice and its coordinates coincide exactly with those of the orifice. In accordance with the uncertainty principle, such high accuracy in fixing the position of the electron severely prejudices the possibility of determining anything about the electron's subsequent behaviour, that is, anything about its subsequent motion (velocity). This is what we observed in our experiment when we found that we could detect an electron with equal probability at any point on the screen.

It follows that when we fix the position of an electron we are, and always will be, utterly unable to determine the direction of its subsequent motion—that is, its velocity—with any accuracy greater than that permitted by Heisenberg's uncertainty principle. Here, chance has a fundamental quality that is not in any way altered by improvements in the accuracy of measurement.

Such is our world, such are its objective laws. The hope that at some future time we shall succeed in ridding ourselves of chance is as futile as the dream of journeying back into the past. (It is true that science-fiction writers have long worked this doubtful vein: however, their efforts have been more in line with their enterprising ability to turn molehills into mountains and collect on the mining rights, than with attempts at scholarly prevision.)

The micro-basis of our world behaves in random fashion: the uncertainty principle lies at its heart.

From this we can draw an exceedingly important and instructive conclusion concerning the uniqueness of every concrete experiment and the strictly unreproducible nature of all experimental results—a conclusion that runs counter to the whole fabric of classical science.

The old, orderly science of the last century maintained that it was absolutely necessary for one and the same set of conditions to produce identical results every time. Yet this is exactly what does *not* happen. Even if we could reproduce a second time all the conditions of a given experiment with complete accuracy, we should still obtain a different result. What does this portend? The ruin of science?—No, of course not: it means that knowledge has taken a great new step forward. The fact that our world is ruled by probability imposes a ban on accurate prediction as a matter of principle. Every extrapolation into the future will always be essentially a statement of a probability rather than of a certainty. 'Nevertheless', as the well-known American physicist Richard Feynman wittily puts it, 'in spite of this, science lives on'.

How can we go on living in a world in which nothing can be accurately predicted? It turns out that this is not such a terrifying prospect after all.

In the first place, the limit of accuracy set by the uncertainty principle is very small—of the order of the dimensions of an atomic nucleus; and this principle makes its presence felt only in measurements at the atomic level. Secondly, unpredictability does not make our world any the less comfortable.

It is true that unpredictability can be regarded as an impediment to precise measurement. However, modern science has developed powerful methods for dealing with errors in measurement (of these we shall speak later) and enables us to eliminate—more precisely, to render painless—the difficulties associated with unpredictability.

But let us go back to the macro-world.

We have already seen that in the micro-world we can never predict accurately the future position of a micro-particle. Now every macro-interaction—that is, every interaction involving sizeable bodies—is made up of a large number of micro-interactions the results of which, in accordance with the uncertainty principle, cannot be accurately predicted. Consequently, we cannot predict the future behaviour of large bodies accurately either, but only approximately and with only a finite degree of reliability.

Let us illustrate this with a simple example. Suppose we have an ordinary roulette wheel consisting of a shallow circular dish with hundreds of holes in the central part of its surface. A small, light ball is released into this dish with a definite velocity, and as it bounces around the rim of the dish it gradually loses speed until it falls into one of the holes. Now suppose that we have made the ball and the roulette wheel perfectly smooth right down to the atomic level (imagine for the moment that this is possible) and that the ball is released by some hypothetical perfectly accurate mechanism so that it always starts from exactly the same position, travelling in exactly the same direction with exactly the same speed. Does this mean that it will always finish up in the same hole?

No, it most certainly does not. In accordance with the uncertainty principle, the direction taken by the ball after each collision with the edge of the wheel will be predictable only in terms of probability. The exact trajectory followed in each case will be impossible to predict because it will be determined at the atomic level, that is, by the interaction of the atoms of the ball and the wheel at the point of contact. And since the conditions of our experiment specify that the initial velocity of the ball is known to a high degree of accuracy, its final position will be somewhat uncertain.

Obviously, with each bounce of the ball the uncertainty of its position increases in cumulative fashion and reaches a maximum when the ball comes to rest. It is this that makes the roulette wheel an essentially random machine. The results it produces are only approximately predictable—and no new methods permitting of even the most accurate measurements will ever enable us to predict the final resting place of the ball with any more accuracy than the principle of uncertainty will allow.

Not so long ago the sensational news spread round the world that a group of young mathematicians had apparently succeeded in computing a system for roulette by means of a modern high-speed computer, and beaten the bank. The 'inexpertise' of such news is obvious. It comes from a desire (no more) to peer into the future, not approximately as in scientific forecast, but absolutely, which contradicts the uncertainty principle. It is difficult to say whether the sensation was an invention or an advertisement for a computer-firm.

Let us note once again that we have been talking so far about an *ideal* roulette wheel, which, despite its perfect construction, turns out to be a random machine. A real roulette wheel behaves with a significantly greater degree of unpredictability owing to the natural roughness and unevenness of both the wheel's surface and the real ball's. This means that the uncertainty of a real roulette wheel is made up of the uncertainty demanded by the Heisenberg principle plus the uncertainty arising from the roughness of the contacting surfaces: and the latter uncertainty outweighs the former to a significant degree. In other words, a real roulette wheel is a random machine in which the basic source of chance is found, not at the atomic level, but in the imperfections of the surfaces that come in contact—in the 'poor workmanship' of these surfaces if you like. Nonetheless, the behaviour of an ideal roulette wheel would also be unpredictable.

It is worth noting that the proprietor of a roulette wheel is always interested in achieving maximum unpredictability. Naturally, if the wheel began to exhibit a preference for certain holes so that the ball fell into them more often than into others, the players would notice what was happening and start placing their bets only on the preferred holes—and the roulette proprietor would be quickly ruined. So to avoid such a misfortune the proprietor tries to keep his roulette wheel operating at maximum randomness with maximum unpredictability of the result.

For the reasons we have outlined above, our world is a world of chance—a world of probabilities. Its random nature is due as much to the properties of the world itself as to the limited abilities of human beings, for whom the exact location of the element of chance is often not very important—whether it be in the essence of a phenomenon or the result of an interaction of man with the world around him.

To sum up, we can borrow an idea from the mythology of our forefathers and say that chance in this world rests securely on these three 'whales':

(1) the principle of uncertainty or indeterminacy;

(2) the inexhaustibility of the universe;

(3) the limitedness of human ability (at the particular moment in time, of course).

The interaction of these three factors compounds the uncertainty of our thrice chancy world.

How should we act in such circumstances?

To begin with, we must rid ourselves of the delusion that we can ever escape from chance completely—this would only be possible if we could invent another world quite different from the one we live in.

Such was the world conceived by the French scientist Laplace, who said that all phenomena were precisely determined by the world's immediately preceding state. Here is how he formulated this notion:

'We must regard the present condition of the universe as the result of its former condition.

'A mind that knew for a given instant of time all the forces existing in nature and the relative positions of all her component parts and, in addition, was so vast as to be able to analyse all these data would embrace in a single formula the motions of the greatest bodies in the universe and the motions of its lightest atoms. There would be nothing such a mind would not know: the future would arise before its gaze just as clearly as the past.'

Obviously, the world of Laplace would be nothing but an endless motion picture unfolding endlessly before our eyes. We ourselves would be part of this film and act strictly in accordance with a script written by God knows whom.

The unreality of such a world is plain for all to see. It is just not our world. Apart from being simply insulting (indeed it is most degrading to be a puppet in someone else's hands), Laplace's universe arouses more serious objections. Laplace's world is predestined, and therefore fantastic. Everything would be just as it is in the script—and this would have fixed everything beforehand. Try as you might, you would not be able to change it a jot. And your struggles to change things would also have been written into the script long before.

Such is the world according to Laplace.

And what about our thrice chancy universe? How should we proceed in one or another chance situation when the situation itself will be impossible to predict? In fact, can we act rationally at all in circumstances involving chance? How can we utilize chance to our own ends?

We shall try to answer all these questions in the succeeding chapters of this book, dealing first with the bad and then with the good consequences of chance. The first part examines the means at our disposal for doing battle with chance; the second discusses ways of using chance for the benefit of mankind.

PART I

Chance
The Obstacle

> Chance plays such a large part in the affairs of the world that I usually try to allow it as little room for manoeuvre as possible: because I am quite certain that it will be able to look after itself without any assistance from me.
>
> *Alexandre Dumas*

1. CHANCE AT THE CRADLE OF CYBERNETICS

In 1940 fascist Germany, having unleashed the Second World War, had air superiority. The German planes were capable of high speeds and could easily escape the fire of the British anti-aircraft batteries because military aircraft were then already flying at speeds comparable to that of an anti-aircraft shell. So it became necessary to aim, not directly at the target, but at a point some distance ahead where plane and shell could be calculated to meet. So long as the speed of aircraft remained low the gunner could determine this point intuitively. Hunters are well aware of the principle involved: for moving game you have to aim ahead of the animal, anticipating its motion by up to a whole body-length depending on how fast it is moving and how far it is away from you. Anti-aircraft gunners of the time adopted the same procedure.

With the appearance of high-speed fighters and dive-bombers it became necessary to anticipate the target by some twenty to thirty lengths—and such a task was beyond the intuitive powers of the gunner. On top of this, when a plane came within a zone of anti-aircraft fire it would begin carrying out evasive tactics that reduced the effectiveness of the gunner's anticipation to zero. These tactics were essentially as follows: when the pilot came into a zone of fire he would deliberately change to a curved flight-path that enabled him to avoid an unpleasant encounter with any shell that had already been fired.

As a result the Germans were able to bomb the towns and cities of Great Britain, inflicting heavy damage and getting off practically scot-free. The British High Command was forced to appeal to leading scientists of the Allied Powers to solve the problem of predicting the position of a plane engaged in evasive anti-flak tactics. The complexity of the problem was due chiefly to the fact that the plane was under the control of a human being whose actions had to be guessed in advance. Naturally the pilot flew his plane in such a way that the gunner could not guess its future position; in other words, he tried to achieve maximum unpredictability of his plane's behaviour. The gunner, on the other hand, tried to work out the pilot's intentions: from the gunner's point of view the plane's evolutions were haphazard because he did not know which way the pilot would move the control wheel next. Did this mean that the plane was to remain forever invulnearable and that anti-aircraft artillerv would have to de dispensed with?

No, it did not.

The point is that a pilot's intentions do not coincide with their fulfilment. Having decided to turn his plane in a certain direction, he moves the controls accordingly and the plane changes course: but not immediately—only after a finite interval of time. Consequently, the pilot does not have unlimited freedom of manoeuvre. In addition, the behaviour of the aircraft always lags behind the pilot's wishes, that is, its motion at any instant corresponds to the positions of the controls a few moments earlier. This enables the anti-aircraft gunner following the plane's flight-path to estimate its behaviour in the immediate future. But how exactly does he do this?

The problem boils down to the prediction of random behaviour. We can soon convince ourselves that this is possible by means of the following simple experiment. The curve depicted in Figure 8 represents the

Figure 8

motion of some object—either a mechanical device or a living creature. The idea is to cover the right-hand side of the diagram with a sheet of paper so that it cannot be seen and to ask your friends to continue the curve on past the dotted line simply by looking at the left-hand part. The right-hand branch of the curve will be unknown to them and will therefore correspond to a chance event. However, in spite of this, most will guess the path of the curve fairly accurately.

Why is this so?

The truth is that the observable left-hand branch of the curve contains information about the right-hand branch; so your observer has no difficulty in predicting its behaviour. But if you asked him to explain why he continued the curve the way he did and not some other way, you should not expect a reasoned answer. At best he will say something like: 'Well, that just seemed to me the best way to do it.'

So a human being can solve this problem. How—we do not know: the fact is that he can. Well, what if we tried to build a machine that could do all this just as well as a man, but faster? And if we could get this machine to control the fire of our anti-aircraft gun we would have a splendid system for effectively knocking out any plane no matter how adroit its evasive tactics were.

But before we can make such a machine we have to know how to solve the problem mathematically. This most difficult problem is called the problem of extrapolating (prolonging, continuing) random trajectories.

It was just this problem that attracted the attention of the American mathematician Norbert Wiener, the founder of cybernetics. His brilliant solution resulted in the timely provision of all Allied anti-aircraft batteries with new devices that automatically selected the point the guns had to aim at the instant of firing.

Thus did the infant science of cybernetics take its first step. Here, cybernetics entered the field against chance and vanquished it—and demonstrated that not all chance events are so impenetrably unforeseeable, and that many of them can be successfully predicted and so deprived of their mysterious veil of unpredictability; and that to do this we only have to look carefully into the particular process and attempt to extrapolate it.

It is a most interesting fact in this connection that the physical realities of the object do not have the slightest effect on our calculations. The method allows us to predict—approximately, of course—not only the trajectory of a controlled flight, but also the behaviour of animals, the future demand for a particular type of product, the magnitude of flash floods, and many other chance happenings of the most varied nature.

We are able to do this because the world about us is not quite so chancy as it appears to be at first glance. If we look carefully into the haze of chance, we can often make out the distinct contours of a natural law: and this enables us to overcome chance and to make fairly accurate prognoses.

Here chance plays a negative role. It retards knowledge; it creates difficulties; it interferes with the lives of men; and in general it hinders progress. We can assert without fear of contradiction that the struggle for progress is by and large a struggle with chance.

Chance seldom assists. More often than not it assumes a destructive role; but even then its machinations are hampered by a powerful factor inherent in progress, namely control—to which we shall now turn our attention.

2. CONTROL

THE BIBLICAL LEGEND AS A LESSON IN CYBERNETICS

According to the well-known biblical legend, the god Sabaoth and his angels created the universe in six days, but they enjoyed its brilliance, its novelty and its harmony for only one day and one night. By the second day the world already possessed a one-day history in the course of which odd parts of it had started to crack up. Here and there the original lustre was becoming tarnished, and someone somewhere had managed to quarrel with his neighbours and the original harmony was upset. The one-day-old world was no longer a model of order and virtue. And, the legend continues, things got worse with each passing day. It is said that a certain Satan was not entirely innocent of responsibility for this sad state of affairs. But Satan did not trouble himself with trifles: he preferred operations conducted on a global scale. Among his specialities, for example, could be numbered such dainty morsels as the wholesale conversion of humanity to vice and the invention of a world-wide system of fire-breathing volcanoes.

However, it was not chiefly thanks to Satan's efforts that the world went downhill—although he certainly did his best. The world lost its lustre because Sabaoth took his hand off the wheel and lost control. As a result, abomination and havoc spread abroad upon the face of the earth. And by the time Sabaoth had snapped out of his trance it was too late. The rot had set in too deep. To have attempted any sort of correction would have been pointless because the whole thing was in need of a thorough overhaul. So the divine will decreed a world-wide flood, with the idea of destroying all earthly ugliness. But somebody thought of a plan to save life on earth from utter extinction and whispered in Sabaoth's ear the idea of building an ark, into which Sabaoth's candidate and captain Noah gathered the finest specimens of terrestrial fauna and flora—including, of course, Noah himself and his sons. And their job was to found a new, properly organized world.

Sad to say, these great expectations were not to be fulfilled. Noah took to drink, and his sons quarrelled and went their separate ways. The world obviously needed constant minding. But Sabaoth was a good-for-nothing lazy-bones. Only now and then, when the sheer tedium of his idleness drove him into a frenzy, did he turn his attention to trying to restore some sort of order. But being an impulsive soul, he was unable to adopt a systematic approach to the business of improving (or *correcting*) living conditions on earth. And the abomination and ugliness continued to flourish.

In the end, Sabaoth—at bottom a sensible sort of a fellow—realized that the world had to be systematically controlled; and that to exert proper and effective control it was necessary to collect information about the controlled object in a systematic fashion. He also saw that effective control from the lordly heights of his heavenly throne was impossible. (This is now understood by every junior devil, but in those days it was a significant break-through.) So Sabaoth sent his son Christ into the world with the job of fixing up a dependable system for gathering information about the state of affairs on earth. Christ, however, failed to complete his high mission: he walked on the waters; he fed a multitude on seven loaves; and he healed the occasional invalid by psychotherapy. It would not have been so bad if he had confined himself to odd tricks of this sort and had got on with the job at the same time; but when he gathered about him a gang of ne'er-do-well apostles and set about creating a personality cult centred on himself, Sabaoth's patience snapped, and Christ was crucified.

From that day on, Sabaoth washed his hands of earthly affairs. Secretly he hoped that Satan would step into the breach and wreak vengeance on these unreasonable human beings that refused to live in accordance with divine laws.

At first Satan had considerable success. After all, who else but Satan, acting on Sabaoth's authority and with Sabaoth's express permission, lit the fires of the Inquisition in the dark days of the Middle Ages. Satan's plan was as cunning as it was simple: to commit to the flames everything that was new and progressive and that could possibly change the existing 'divine' order. If Sabaoth's experiment with the flood destroyed all that was evil, leaving only the best things behind, Satan, predictably enough, did exactly the opposite and destroyed all that was good, so that abomination blossomed exuberantly in conditions that guaranteed it perennial luxuriance.

By this time, however, man had reached the stage where he was able to take upon himself the function of control and topple Satan from his perch. Satan is now confined to working with trifles.

This simple story, like any other fairy-tale, reflects the naive imaginings of people of a distant era concerning natural phenomena that they were otherwise at a loss to explain.

You do not have to be particularly observant to notice that there are two powerful tendencies operating in the world around us. One of them is bent on destruction, the other on creation. Thanks to the first tendency, our world is shaken by a variety of catastrophes that result in a mass of unpleasant and awkward situations: bridges and houses crumble away; plants and animals age and die; and so on. This 'evil' tendency was obviously responsible for the appearance, in a primitive age, of the superstitious concept of a devil (Satan) personifying the destructive principle in our world. Modern science, however, describes this aspect of natural phenomena in terms of the second law of thermodynámics, which may rightly be called the law of chaos.

The second law of thermodynamics was first formulated in 1829 by the French engineer Sadi Carnot. The essence of this law can be stated as follows: every closed system—that is, every system that is completely isolated, and is not connected in any way with any other system—tends towards its *most probable state.* This most probable state is complete chaos. So, in accordance with the second law, all closed systems gradually become disorganized, decay, and die. In engineering practice this process is often called depreciation; in biology—aging; in chemistry—decomposition; in sociology—decay; in history—decline.

In order to measure the degree of disorder, or chaos, of a system, we usually make use of the concept of *entropy.* This is a property of a system such that the greater the disorder of the system the greater its entropy. We can now re-formulate the second law of thermodynamics thus: *the entropy of a closed system does not decrease.* In other words, a closed system cannot of itself increase its state of organization.

We should note that the second law of thermodynamics is an experimentally established law. However, no instance has ever been observed where it has not been valid.

At this point someone is bound to ask why the world around us is not in a state of complete chaos, and is plainly not even tending toward such a state, contrary to what the second law of thermodynamics would seem to require. Biological systems, for example, that is, living organisms, are highly organized systems with an extremely low level of entropy. How are we to reconcile the existence of such improbable systems with the second law? Or again: modern progress is aimed at making life more orderly in defiance of the second law—and the successes achieved in this direction are plain for all to see.

There is actually no contradiction here at all; nor has anybody seriously challenged as yet the absoluteness of the second law. The point is that the concept of a closed system, for which this law is enunciated, is a rather remote abstraction. In the real world there is simply no such thing as a closed system: all real systems are interrelated and interdependent. The connections between them may be strong, or they may be weak; but they are always there. Moreover, it is impossible to make a system closed by artificial means because every system is always under the influence of thermal and gravitational effects due to other systems, however slight they may be.

So we cannot consider our earth a closed system, because the earth receives energy from the sun.

Similarly, our solar system is not a closed system, because it is affected by galactic radiation and gravitation. Admittedly, the rate at which the solar system receives radiation is low; but in the course of thousands of millions of years this accumulated radiation has had a significant effect on it.

These facts give the second law a somewhat academic flavour and dissipate the emotional tension associated with the 'heat death' of the universe.

It is worth going into the spectre of the heat death in a little more detail. During the last century many scientists made the mistake of attempting to apply the second law of thermodynamics to the whole of the universe taken as a single closed system. (Even today there are a few scientists who still adhere to this view.) This led to the hypothesis of the heat death of the universe: the 'dead' universe was visualized as an expanse of evenly heated material in which there were no rises or falls in temperature. Certainly it is true that the increase in entropy of a closed system tends to even out the temperatures at all points of the system. Life would be an impossibility in such a universe because no engine (in the widest sense of the word) can function in the absence of a temperature difference. Any machine that produces work does so at the expense of cooling a heated part and heating a cold one.

Living things are no exception to this rule. A living organism is a highly complex engine that likewise cannot work without a temperature difference between itself and its surroundings. If there is no temperature difference, life ceases to exist. Hence the term 'heat death'.

Despite its superficially convincing and incontrovertible logic, the heat-death theory suffers from one grave defect: it is based on a false premise. All the terrors of the heat death are possible only in a closed system; and such a thing, fortunately, does not exist. We are saved from the heat death, therefore, by the *law of the overall interrelationship and interdependence of phenomena and objects throughout the world.* Neither the universe in its entirety nor any part of it can in any way be considered a closed system; the applicability of the second law of thermodynamics does not extend over either the one or the other. In other words, we are not threatened by any 'heat death'.

But let us return to the second law of thermodynamics. It does not exclude the possibility of a localized decrease in entropy even within a closed system. It allows some local organization—but only at the expense of more intense destruction of everything else. Localized ordering of a certain part of a closed system is possible only on condition that the remainder of the system become more disorganized. However, in accordance with the second law, the degree of order taken over the whole system does not increase.

The problem of raising the degree of order of a system was first taken up by Maxwell in 1871. Maxwell formulated the problem in terms of a paradox now known by the rather unusual name of '*Maxwell's demon*'. (A demon is not to be confused with a devil. There is nothing nasty about a demon: unlike a devil, a demon fulfils useful functions and may be regarded as man's ally in the struggle with the devil of chaos.)

MAXWELL'S DEMON

In 1871, that is, before the appearance of cybernetics, Maxwell's clever paradox was incompatible with the second law of thermodynamics. We shall now see why.

Let us suppose we have an empty, isolated box divided internally by a partition into two sections (Fig. 9). Suppose we fill both sections of the box with

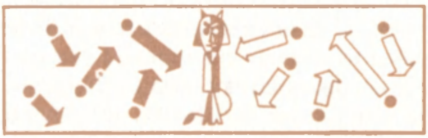


Figure 9

gas at the same initial temperature. This system—a box containing gas at a uniform temperature—has maximum entropy. If the temperature in one compartment differed from that in the other, the system would be more organized and its entropy would be lower correspondingly. In accordance with the second law of thermodynamics, the temperatures in the two compartments tend to converge; and this may be observed experimentally, as everybody knows.

Now we make a hole in the partition and fit it with a shutter that we can open and close at will. Suppose the shutter is under the control of our hypothetical demon (the duties he is to perform are beyond the capabilities of anything but a mythical creature having unlimited powers). The demon acts in accordance with the following instruction or *algorithm*: he opens the shutter so as to allow only fast molecules to pass in one direction from one compartment to the other, and only slow molecules in the opposite direction.

The motions of the gas molecules in the box may be likened to the motions of a number of billiard balls all moving with a variety of velocities, colliding, rebounding, colliding again and so on, and all the time exchanging energy.

The speeds of the molecules vary considerably; and the actual speed of any given molecule at any given moment is entirely a matter of chance. However, the average speed of the molecules is related to the temperature of the gas: the greater the average speed the higher the temperature, and vice versa. Consequently, both compartments of the box always contain faist molecules moving towards the shuttered hole, as well as slow ones. The demon's task is to let or not let molecules pass from one compartment to the other, depending on their speed.

It is easy to see that after the demon has been at this for a while, one compartment will contain a higher proportion of fast molecules, the other a higher proportion of slow molecules. The temperature of the gas in the first compartment will be correspondingly higher, and in the second it will be lower. The entropy of the system will now be lower than it was initially, because it contains a temperature difference.

Here we have an obvious paradox. Our closed system—the box containing the gas and the demon—is able to increase its degree of order, apparently in defiance of the second law of thermodynamics. The explanation of this paradox was only forthcoming after the creation of cybernetics.

The point is that when the demon manipulates the shutter he feeds *information* into the system, and this information organizes the system. By sorting and classifying the molecules he makes the system more organized: he *controls* the system, that is, he acts upon it in such a way that it becomes more orderly.

This is all very well; but it cannot be achieved at no cost. If the demon is to control effectively, he must receive information about the speed of the molecules. And since the system is supposed to be a closed system, and we cannot even let light into the box from outside without invalidating its closed status, the demon himself has to provide the energy required to obtain the necessary information. For example, he may have to illuminate the molecules with a torch, thereby partially discharging—that is, destroying—the torch battery.

Consequently, the organization that is achieved is of a purely local character because it affects only the gas. The local decrease in the entropy of the gas is produced at the expense of increasing the entropy of the battery. And if we summed these two effects algebraically, we would find that the total disorder of our closed system (gas and battery) had increased.

The relationship between the gas, the demon and the battery are shown schematically in Figure 10. Here the demon receives information about the motion of the molecules via channel A and on the basis of this information he controls the shutter via channel B, all the time drawing energy from the battery. As a result, *organization* is pumped, so to speak, from the battery to the gas. In the process a certain amount of organization is inevitably lost (these losses are represented in the diagram by the dashed arrow).

The diagram contains all the essential features of any control system. The controlled object in this case

Figure 10

is the gas. The demon fulfils the role of a control device acting in accordance with a set algorithm (or instruction). And the source of organization is the energy of the battery.

There is nothing idle about our diagram: it captures the idea of control in all its profundity. Figure 11

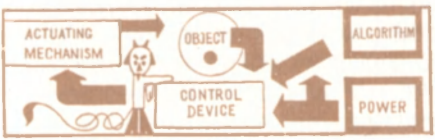


Figure 11

shows a control diagram for any object in general. As you can see, it is a virtual copy of the demon diagram of Figure 10.

Here information about the condition of the object is obtained in exchange for a definite expenditure of energy and fed continuously to a control device fulfilling the functions of our demon. The control device compares the information with the requirements sta ted in the control algorithm (shown in the diagram a a heavy rectangle) and controls the object on the basi of this comparison.

It is evident that two factors are absolutely essen tial if the system is to work:

1. A source of organization or entropy reduction
2. A control algorithm, that is, a rule enablin control to be effected on the basis of the informatio received.

The question of the energy source causes no grea trouble because it can be solved by modern powe engineering using fairly simple devices. But the tas of formulating (or synthesizing) the control algorith is by no means always so easy.

Control of any kind is concerned with organizin an object. It is a purposeful activity directed toward transferring an object from a more probable state t a less probable one.

The problem of synthesizing control and analysin its operation is an information problem and, as such it constitutes the basis of modern cybernetics as science. Control is a means of acting upon our sur roundings, of subjugating nature to mankind and of ra tionally altering our world.

In this sense control is in opposition to the secon law of thermodynamics. Control lowers the entrop of an object: the second law of thermodynamics postu lates increasing disorganization for the object. Bu control produces local organization only, whereas th second law applies to a closed system as a whole. The refore we cannot say that control actually contradict the second law, because the two things operate o different planes.

Control is always local. The second law of thermody namics is always integral. It is a universal law.

Let us examine the processes of aging and healing as illustrations of disorganization and control. Aging is a typical process of entropy increase. Healing is a typical control process that lowers the entropy of a living organism. Aging proceeds in all organs simultaneously and in parallel fashion. It is a general, integral process encompassing all cells of an organism. Healing, on the other hand, is local: it is directed towards improving the functioning of one specific organ, but not of the organism as a whole. Not for nothing is modern medicine split up into a mass of 'ologies', each devoted to the healing (or control) of a single organ, for example: cardiology for the heart, neurophysiology for the brain, stomatology for the oral cavity, and so on. Obviously, aging extends over the whole organism and healing aims to control individual parts of it only.

To carry this idea further: the second law of thermodynamics acts all the time and everywhere throughout the universe; but control operates only where there are information processes, that is, where there are programmes indicating what has to be done to achieve control. Programmes of this type are products of deliberate activity, and they result from the functioning of living things. And this gives us the basis for linking control with life.

To take this a step further still: we can assert that every process of control is the result of the activities of living things—and of living things only. This means that until the appearance of life on earth control or organization of any sort was completely out of the question.

'But what about crystals!' the observant reader will exclaim. Who, to be sure, has not had occasion to admire the fantastic, but strictly regular, shapes and facets of mineral crystals and snowflakes? Surely this

4*

is organized matter in its highest form? And yet crystals are formed without the help of any sort of deliberate activity, much less that of man. What has happened here? How can we reconcile this contradiction? First we should note that the crystallization process involves loss of energy. While a system is undergoing crystallization it ceases to be a closed system: consequently, the second law of thermodynamics no longer applies to it as a whole. But there is yet another essential feature.

To explain this we can do the following simple experiment. We put some ordinary beach sand into a glass of water and stir it vigorously as if we were trying to dissolve it. So long as we are expending effort on stirring, the sand and water form a fairly homogeneous mixture. But as soon as we cut off the supply of energy the sand falls to the bottom of the glass, so that the sand and the water separate.

Now, which of these two states of the contents of the glass possesses the greater degree of organization? At first it would appear that the stirred-up sand bears a strong resemblance to chaos, so that the first state seems to be completely devoid of organization; and that the distinct boundary between the sand and the water in the second state points to a high level of organization.

Actually the reverse is true. The stirred sand has the lower entropy, and is maintained at this low level by the continuous supply of energy. The distinct boundary between the sand and the water in the calm state, on the other hand, is obtained only at the cost of giving up energy, namely: the potential energy of the sand. It is no secret that all processes proceed in the direction of decreasing potential energy: this circumstance, in fact, is the basis of one formulation of the second law of thermodynamics.

And so it is with crystals. The formation of crystals is a process involving an increase in entropy and, consequently, a loss of organization—even though it seems superficially to constitute the formation of the highest type of organization. Crystallization, in fact, is a process of transition from a less stable to a more stable state with an accompanying loss of energy.

We see now that the concept of order as used in cybernetics often differs markedly from our everyday understanding of the term. In cybernetics, order means that condition that satisfies a particular aim. Sometimes the stated aim coincides with the result postulated by the second law of thermodynamics. In such cases the aim is very easy to achieve. For example, in order to destroy a building, it is sufficient to blow it up with dynamite. The second law takes care of the rest: it turns the building into a heap of rubble that testifies to the greatest possible entropy and to the triumph of chaos. Exactly the same thing happened to the sand in the glass. The sand falls to the bottom of the glass under the influence of gravity. Here we have a perfect example of the second law for a closed system—the system being the glass full of sand and water together with the earth we live on. (The earth is included as the source of gravity without which the sand would not fall to the bottom of the glass.) So before the sand settles, the entropy of the system is less than it is after the sand settles. In fact, we can easily picture a device that would make use of the energy of the falling sand—a paddle-wheel, for example, that would turn under the action of the descending grains.

However, if we wish to restore the building—and this would be in direct opposition to the second law—we shall have a good deal of work to do to lower its entropy and get it back into an orderly condition again. Here, when we talk about 'work' we are thinking in terms not of energy expenditure (though this is vital to the project), but of information expenditure—expenditure on control.

Here we encounter another approach to control in which the concept of aim plays a decisive role.

CONTROL AS A MEANS OF ACHIEVING SPECIFIC AIMS

What do you think the following have in common: Maxwell's demon, a thermostat, a yard-keeper, a machine-tool operator, an administrator, a design engineer, and a research worker?

Maxwell's demon is a hypothetical creature that was invented by Maxwell for the purpose of constructing a paradox that could not be resolved without reference to the concept of control.

A thermostat is a device for controlling temperature. It works like this: if the temperature in the room falls below a prearranged value, the thermostat switches on a heater; when the temperature rises above the preset value, it switches off the heater.

The yard-keeper, the machine-tool operator, the administrator, the design engineer, and the research worker are all people fulfilling various functions in human society.

At first sight there seems to be nothing common to all of these. We cannot even say that they share material being, because the demon is an imaginary creature and has no real existence.

And yet there is something common to them all. What they share is the purposefulness of their activity. They are all control devices whose action is directed towards achieving specific aims. Essentially, they organize an object and bring it closer to perfection; in other words, they lower its entropy.

The distinctive characteristic of any control device is its purposeful behaviour aimed at fulfilling a specific objective; and this activity is applied directly to the object under control, its only purpose being to make the object attain a clearly defined ideal or aim.

For Maxwell's demon the aim was to raise the concentration of fast molecules in one part of the box and of slow molecules in the other.

For the thermostat the aim is to keep the temperature of the room at a particular level.

The reader will have no difficulty in describing for himself the aims of such control devices as a yard-keeper, a machine-tool operator, an administrator, and so on.

However, for effective control it is not enough simply to know the objective: we also need to know how it can be attained—we have to be able to influence the object under control in such a way that our plan is fulfilled. And this is often a good deal more difficult than the problem of defining the aim.

In some cases, of course, this problem can be solved fairly easily. For example, with the thermostat it is perfectly clear that when the temperature in the room is too low the correct solution is to switch on a heater, rather than to go basking in the African sunshine; and when the temperature is too high, to switch off the heater or switch on the air-conditioner.

This sort of simplicity, however, is the exception rather than the rule. It is usually extraordinarily difficult to determine just how the stated aim is to be achieved. This brings us to a consideration of one of the most fundamental concepts of modern cybernetics, namely the *control algorithm.*

A control algorithm is a method of achieving the stated aim—a sort of rule for action. For Maxwell's demon, this rule was the instruction for sorting molecules according to their speed. The control algorithm for the thermostat is a rule specifying when the heater or the refrigerator is to be switched on and off. The yard-keeper achieves his aim—a clean street—by the algorithm of removing litter, which he realizes with the help of dustpan and broom. The machine-tool operator achieves his aim—the manufacture of a component in accordance with a drawing—by applying the algorithm of removing surplus metal with the help of a metal-cutting machine-tool. The administrator works towards his objective—the fulfilment of a plan or quota—using control algorithms in the form of rewards and punishments for his subordinates according to the diligence with which they work towards their objectives. The list of examples could be extended indefinitely. It is more useful, however, to consider a general control system independently of the specific peculiarities of controlling a particular object.

Figure 12 shows the block diagram for a general control system. Here the interaction of the object with

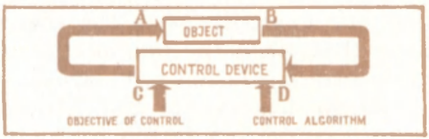


Figure 12

the control device is indicated by two arrows, *A* and *B*, representing communication channels between the object and the control device. The control device receives information about the object via channel *B* (because you cannot hope to control anything if you do not know what it is up to) and then acts upon the object via channel *A*, and thus controls it (because unless you can do something to the object, control is, once again, impossible).

However, this is not the full picture, as we have already indicated. We still need to know *what* to do with the information received, *how* to use it to control the object, and *exactly what* we should be aiming at in the process. In order to satisfy these requirements we have two further inputs to the control device: the objective or aim of the control process (arrow *C* in the diagram); and the algorithm or method of control (arrow *D*). These data have to be fed into the control device beforehand. So, if the control system is to impose the required order on the object, it must contain two essential elements: (1) the objective of control; (2) the control algorithm showing how the objective is to be attained.

This control system is valid for any controlled object. Once again we emphasize that it will only work in conjunction with a control programme, or algorithm (arrow *D*, Fig. 12). Such a programme of local organization—a programme of purposeful changes to the object—must be incorporated in the control device: this is what enables the control device to organize the object and to bring it into the desired, less probable condition. Only when this has been done can we hope to improve the object.

Every act of control is the result of purposeful behaviour. But we know that there is no such thing as purposeful behaviour in inanimate nature. There the second law of thermodynamics reigns supreme, acknowledging only chaos and having maximum chaos as its 'aim'. So it is natural to suppose that the purposefulness and the teleology of our world are due to the existence of some degree of purposefulness and teleology in the past.

If we want to switch on a desk lamp, for example, we press the button switch. This in fact is a control process that brings the system (in this case the darkened room) into the required (brightly lit) condition. Here, the control algorithm is the rule 'press the switch'. But if we did not know this rule we would have absolutely no way of getting our darkened room into the desired illuminated condition. Control of the room's illumination turns out to be possible only when we know the algorithm.

And so we see that progress and the improvement of the material world depend very largely on certain control algorithms thanks to which progress is achieved.

However, the algorithms themselves are also products of organization. Indeed, if we wish to know how to control an object we have to get the requisite instructions from someone. For example, to be able to switch on a lamp we have to be taught *by somebody* how to do it. This 'somebody' already *knows* the purpose of the switch and passes this knowledge on to us in a learning process. So an algorithm cannot be created (synthesized), or put to work in a control system, without the help of a prior algorithm.

From the example of Maxwell's demon we can see that to increase the state of organization of the gas we first have to invent the control algorithm ourselves (namely, sorting the molecules into different sections of the box according to their speeds). Then we have to construct the demon, that is, a device capable of acting in accordance with the given algorithm. But what do we mean here by 'invent' and 'construct'?

These activities are also purposeful; therefore they must be accompanied by a lowering of entropy. If we are to invent and to construct we need to know the algorithms 'how to invent' and 'how to construct', and so on.

We soon come to the conclusion that there must be a whole chain of algorithms like this. At the head of the chain we expect to find the simplest control algorithm that is capable of giving rise to all the rest. In other words, the chain must be started by something very like an 'act of creation'.

In the biblical legend of the creation of the world the creating is done by a god, the god being a highly organized, rational system that knows 'how to create' —that, after all, is why he is a god. But who created the god and taught him how to create in the first place? This the bible does not tell us.

Let us take another pretty legend—the one about Prometheus, who, according to the ancient Greeks, taught men how to obtain fire and how to use it. In the language of cybernetics, Prometheus knew the algorithm for making fire and the algorithms for roasting meat and for forging and melting metals and possessed a veritable fund of useful knowledge covering a wide variety of subjects. But who taught him all this? Who told Prometheus all these algorithms? 'Zeus!' you say? Then who told Zeus?

This kind of reasoning always finishes up in an impasse. Indeed, if all control is the result of the activities of living things, and if they themselves result from control—or rather, self-control—it is only natural to ask: where did the first instance of control on earth come from, that is, where did life on earth come from?

The answer that saves the argument—or, more precisely, the substitute for an answer—consists in an appeal to an unearthly, cosmic origin for life, as propounded by the theory of panspermy. But if we ask where this unearthly, cosmic life originated, the panspermic theory, so far from providing an answer, actually considers the question idle. Life is life, and that is all there is to it; and if there is life, well—there is also control.

To me this answer seems perfectly correct—and perfectly inadequate. Control, as a means of lowering entropy and improving organization, may well have had a most interesting, not to say breathtaking, history: and this would provide us with another way out of the dilemma.

To write such a history we have only to remember that the creation of control algorithms does not proceed by hereditary transmission alone, but also by *spontaneous organization*, or spontaneous generation. This means that control algorithms can create themselves, that is, they can synthesize themselves.

How does this happen?

During the period control has been in existence many ways of creating (or synthesizing) control algorithms have been developed. The history of control can be divided into four stages, each characterized by the appearance of new ways of synthesizing control algorithms. We can give these stages the following names:

Stage one—the *probability* stage.
Stage two—the *elemental* stage.
Stage three—the *intelligent* stage.
Stage four—the *universal* stage.

We shall now examine each stage separately.

3. THE HISTORY OF CONTROL

STAGE ONE

Control on earth began with the *probability* stage. (Here, we are considering only the history of control on earth: this in no way excludes other lines of development that may be possible under different conditions and in different planetary or stellar systems.) This stage is characterized by the chance formation of the simplest systems that we can call organized. These take the form of molecules of various elementary proteins and amino acids. These kinds of molecules were formed accidentally as the result of electrical discharges in the earth's atmosphere, which at that time consisted of water vapour (H_2O), methane (CH_4), ammonia (NH_3), and hydrogen (H_2).

These substances, interacting with each other *at random*, formed more complex structures. If these structures proved to be stable they existed for a while, reacting with other similar structures. Those that were unstable decomposed rapidly and immediately participated in new accidental combinations.

As a result of this random 'trying out' of all possible structural combinations, the most stable structures gradually proceeded to more advanced stages of development. At the same time, the most active molecules continued to take part in the 'game', while passive structures withdrew. An essential condition for such a process to continue is that the reactants be in a state of energetic motion: and the swirling, seething atmosphere of the primordial earth provided ideal conditions.

The American scientist S. Miller performed an interesting experiment that was at once fairly simple and highly edifying. He made up a mixture of gases corresponding to the supposed atmosphere of the primeval earth and passed electrical discharges through it to simulate lightning. At the end of a week he carried out a painstaking chemical analysis of the mixture. You can imagine his astonishment when he discovered that the flask contained amino acids. Amino acids are the fundamental building blocks of the proteins, the basis of life itself. In particular, he was able to establish beyond doubt the presence of the amino acids most commonly found in proteins, namely: glycine and alanine, both of which have extremely complex structures.

How were they formed?

The only reasonable answer is supplied by chance It was only by virtue of the multiplicity of fortuitou combinations and relationships that could be forme among the molecules of water, ammonia, methane an hydrogen under the conditions of high temperature produced by electrical discharges that molecules c more complex structure could arise. And for this ther was time enough: the earth's atmosphere and hydro sphere were rent by violent storms for many million of years before the nutrient broth of life—a solution c various amino acids—came into being. In this proces the role of chance was decisive.

But already this broth was coming under the in fluence of the second law of thermodynamics. Accord ing to this law large molecules could not be distribute evenly throughout the water. In much the same wa as saturated steam condenses to form a mist consist ing of minute droplets of water, the large molecule in the solution coalesced into separate clusters hel together by electrostatic forces. When these cluste reached a certain density they separated out of the so lution to form what are called *coacervate* drops tha remained floating in the solution. These drops woul have been separated from the surrounding mediu by a well-defined interface.

Although the tendency for these drops to form wa not in itself a chance affair, the actual combinatio of amino-acid molecules in each drop was. Ever amino-acid drop had a highly individual structure. A this point a peculiar process of selection goes into ope ration, a process that has been noted and describe by the Soviet academician A. Oparin.

It is fairly obvious that if the accidental structur of a drop happened to be unstable, the drop woul break up under the action of external forces. Conse quently, only drops containing stable combinations of molecules would be preserved. Unstable combinations would 'die out', so to speak, and the resulting fragments would be re-formed into further chance structures. Clearly, given a sufficiently long period of time for the operation of this process, only stable drops would remain in the end—that is, only those with the ability to withstand the destructive forces of the environment.

A stable drop, like any other body, would adsorb various molecules from the solution and thereby increase its volume. These new molecules no longer attached themselves haphazardly over the surface of the drop, but were arranged in conformity with its particular surface structure. The drop thus grew, gaining in mass. These increases in dimensions did not proceed at random, but strictly in accordance with the individual properties of each drop.

When a drop reached a certain size, it would become mechanically unstable and break up into two or three pieces under the action of external mechanical forces—rather as the droplets of an emulsion are broken up by shaking. The newly formed drops had the same structure as the original one: they inherited the peculiarities of the parent drop, and then they, too, began to grow and break into pieces. And so on.

But this was still not life: it was what is known as pre-biological structure. It possesses nearly all the attributes characteristic of life, but in forms rudimentary to the point of being grotesque. The drop is indeed similar to a living cell. The gathering of solution molecules on the surface of the drop can be regarded as a form of feeding and the mechanically induced rupture of the enlarged drop as a form of cell-division. With the latter the analogy even extends to include elements of heredity.

Just as in real life!

But life is still a very long way off. Many millions of years have yet to pass before natural selection succeeds in turning these drops into living cells. But the essential materials are already at hand. It is only matter of time.

And of time nature had ample store.

Some one to one and half thousand million years pass, and multi-cellular organisms appear. The mould-like, mucilaginous forms of earliest life gradually give way to active types like the creatures we are familiar with today.

And so the probability stage in the history of control is characterized chiefly by such an abundance of chance events as made possible the creation of life on earth. We can boldly assert, therefore, that chance was a root cause of the appearance of life on earth.

The actual creation of life in the nutrient broth was a chance process; however, given such conditions as we have described, it was inevitable that life should arise. After all, in the course of the thousand million years of random trialling of all possible structural combinations of various organic molecules it was absolutely inevitable that one combination at least should prove successful and possess the properties of a living cell.

The instant when this happened marks the beginning of the history of life and also, obviously, the end of the probability stage in the history of control. This stage was distinguished by the chance appearance of a local decrease in entropy, resulting from the operation of statistical laws.

But when life appeared, it brought with it new possibilities for control.

The *elemental* stage in the history of control is concerned with the development and improvement of living organisms. The basic control algorithm here was the algorithm of natural selection discovered by Charles Darwin.

According to this algorithm the individual that is better adapted to its environment has a greater likelihood of producing offspring; the less well adapted individual perishes without producing offspring in which its lack of adaptation could be preserved. As a result of natural selection an enormous variety of control algorithms has come into being. These include: algorithms that control the mechanical behaviour of living things—swimming, crawling, flying and walking; algorithms governing the mental functions—aggressiveness, evasion, flight, playing dead, and so forth; algorithms that take care of the functioning of the nervous system; and so on.

One way or another, during the elemental stage the synthesis of every new algorithm was controlled by and subject to the law of natural selection. Of all the algorithms that arose for the self-regulation of living organisms, only those that enabled the organism to deal more effectively with its environment survived and were preserved.

Not that nature could get by without the occasional oddity. An ostrich, for example, confronted by grave danger, buries its head in the sand. Allusion to this strange behaviour on the part of the ostrich has become a standard metaphor for stupidity characterized by an inability to face up to facts. How indeed did such an apparently senseless algorithm for behaviour in the face of danger come about in the first place? Surely it would be better to attack, or to flee? Perhaps nature blundered here?

On closer examination it becomes evident that for the toothless, hornless, hoofless ostrich this behavior algorithm is optimal in just those situations where flight is impossible.

When an ostrich buries its head in the sand it cannot see the source of danger, and stands still; and, strange as it may seem, this often helps the ostrich to escape the predator's jaws, the reason being that most carnivores feed on the meat of animals they have only just killed (yet another example of adaptational logic particularly applicable in hot countries). The motionless ostrich fails to arouse the predator's appetite. The predator would sooner chase an antelope streaking away over the horizon, than tackle a motionless mound of feathers a mere yard or so away. And this is what saves the ostrich.

But why should the bird bury its head in the sand instead of just standing still and no more? The answer is that this procedure keeps the nervous strain down to a minimum, and this renders the chosen algorithm less terrible an ordeal.

Man uses much the same device when he encounters a bear—but by design rather than by instinct, without burying his head in the sand. (So we are told, at any rate, by authors of books on the subject.)

We see then that this stage in the development of control is characterized entirely by the elemental nature of the process of selecting effective control algorithms in living organisms.

STAGE THREE

The next stage in the synthesis of control algorithms is associated with human activity. When man emerged on the orbit of history he immediately trumpeted forth his ability to create control algorithms by the application of his intellect rather than by elemental reliance on chance. This ability distinguished him from all other animals.

To be precise, the stage of *intelligent* synthesis of control algorithms begins not with the appearance of man himself, but with man's rational activity. This stage differs from the two earlier ones in that control algorithms are now created by man.

The development of crafts and sciences forms the basis of man's control activity. Man began to impose order on the world about him by inventing a multitude of algorithms for controlling and purposefully altering nature. Each of these algorithms was characterized by its uniqueness, because each was applied to a different natural object. For example, the craft of a potter differed from that of a blacksmith because the objects they worked with—their raw materials—were different: moist clay on the one hand; red-hot metal on the other. And so the control algorithms for shaping these different objects differed.

It is impossible to create algorithms for changing the world without a good understanding of what makes the world tick, that is, without the development of science as a system of ordered knowledge about nature. Such knowledge is acquired as a result of interpreting observed phenomena, of understanding their essential features and explaining their nature.

But what do we mean by *interpreting*, *understanding* and *explaining nature*? What is *knowledge*? Can we express these hazy concepts in a precise and definite form susceptible to quantitative evaluation?

Yes, we can; to do this we need only the ability to predict the progress and behaviour of the phenomenon that we are interested in. Our ability to predict is determined to a significant degree by the amount of information we possess about the object under investigation. If we know a lot about a given process, we can predict fairly accurately what it will do in various situations. Any discrepancies between the actual course taken by a real process and its predicted behaviour characterize the reliability of the prediction and at the same time determine the magnitude of our ignorance. The smaller these discrepancies are, the better is our state of knowledge concerning the given process and the greater, obviously, will be the value we can place on our understanding of the nature of the process.

Of course, we cannot assert that the ability to predict something accurately is always equivalent to having profound knowledge. But these two things are unquestionably related to each other. As a rule, our ability to predict the behaviour of any process depends on having a profound understanding of the nature of the process.

On this basis it is convenient to define knowledge as the ability to predict. A system of judgements and conclusions that allows us to predict in a particular way the behaviour of a particular phenomenon we shall call a *model* of the phenomenon.

Let us consider, for example, the phenomenon of a falling stone. By releasing the stone from various heights and measuring its time of descent we can establish a relation between height and time, and thus formulate a law of free fall. This law will then be a model that will allow us to predict the behaviour of a stone falling from various heights.

To take another example: Gregor Mendel was able to establish the laws of heredity by crossing a red-flowering variety of pea with a white-flowering variety. He showed that inheritable characters were transmitted by the parents in finite heredity units that could not be subdivided. In much the same way as energy is transferred in quanta (strictly determined, smallest, indivisible portions), so heredity is also quantified. The quanta of heredity are the *genes*, the material bearers of indivisible characters. For example, in Mendel's experiments with peas the hybrid plants always had either red or white flowers, but never flowers of an intermediate colour (such as pink). This means that the colour of the flowers is determined by one of two genes—the red-flowering gene and the white-flowering gene; other genes for determining flower colour in the pea do not exist.

Mendel reduced his observations to a law of heredity which states that parental attributes are not averaged in the offspring, but are transmitted in the form of individual characters (Daddy's nose, Mummy's eyes, and Granny's temperament). This law is a model that allows us to forecast the way in which parental characters will be inherited.

And so we see that our knowledge of the world resides in models of its phenomena. These models enable us to foresee the consequences of our interactions with the objects that make up our world. To take a simple example: if we did not know the law of falling bodies we would be unable to use the ballistic missile, because without this law we could not predict where the missile would land.

It is obvious that the creation (or synthesis) of models like these is also a process that raises the level of organization of human thought. The concrete results of this state of organization are to be seen in the purposeful actions that man performs on the basis of the models he possesses.

Consider, for example, a hunter. In the course of his training (per medium of books, stories and demonstrations) and later when he is gaining actual field experience, he learns the characteristic behavioral patterns of various animals. In other words, he forms models of their behaviour within his brain, and then makes use of them when he plans his hunt. Here, non-material (mental) model of behaviour makes it possible to control the catch effectively with a consequent gain in material benefits.

Explanations of the mechanisms of natural phenomena can thus be regarded as control because they involve the construction of models of the phenomena. The process of acquiring knowledge—that is, of constructing models—is a process in which facts are purposefully ordered and entropy is lowered. By 'purposefully' we mean that our activity is such that the constructed model differs in its effects as little as possible from the object under investigation; the less the difference the better the model. So, for example, the well known laws of Newton provide a perfectly satisfactory model for low-speed mechanical motion even though they are in fact approximations to reality. The most successful hunter will be the one who possesses the best models of animal behaviour and consequently is best able to predict what an animal will do in any given situation.

A system of models constructed by man with a view to effectively controlling his surroundings constitutes a *science*.

You will notice that at this intelligent stage in the history of control, man's activity has a dual nature: on the one hand he is changing his surroundings by actively controlling nature; and on the other he is explaining nature by creating the models necessary to achieve the afore-mentioned changes. These two functions are closely interrelated. Obviously, if we are to change the world intelligently and adapt it to our needs, we have to know what the consequences of any particular action will be. You can easily imagine the sort of mess we could get ourselves into if we pursued courses of action whose results we could not predict even approximately.

But we can only predict on the basis of models. Consequently, rational action is quite impossible without models that the action can be tried out on and tested beforehand. There is no such thing as a rational act that does not take into account its possible consequences.

If we wish to send a rocket to the moon we must construct a model of the proposed flight. We have to be able to calculate the rocket's position as a function of time and of other variables. Otherwise we shall simply be wasting time and resources on empty amusements.

It is impossible to change the world in a purposeful fashion without creating models.

Obviously, the methods for solving different problems are different. Man has created an enormous number of algorithms for explaining and changing nature. Each of these algorithms is notable for its strictly local, particular and specialized character. For example, there are many different ways of setting up a model for the behaviour of an animal in a trap or at a watering-place. Every hunter has his own pet method (or algorithm) for studying his quarry's habits. Similarly, we can apply our knowledge in different ways to achieve our aims. To go back to hunting once again, the actual location and siting of a trap will depend both on the available models of the animal's behaviour and on the personal experience of the hunter.

STAGE FOUR

With the birth of cybernetics—the science of control in the animal and the machine—the next, and apparently the last, period of this history begins: the stage of *universal* control algorithms. Algorithms of this type can be applied to any object regardless of its physical reality. Cybernetics examines control processes from a general point of view, rather than in relation to a particular concrete situation. In cybernetics we are concerned only with a model that represents not the physical, but the informational core of the events taking place within the object while it is under control.

A single model is able to describe control processes in objects that differ in their physical make-up. For example, the oscillator provides the mathematical model for such varied phenomena as the oscillation of a mechanical pendulum, the variations of current and voltage in an electric circuit, and the changes in predator populations. The control process for each of these will be identical.

Let us look at this example in more detail.

THE TALE OF THE LITTLE GIRL ON THE SWING, THE BIG BAD WOLF, AND THE ELECTRICAL CIRCUIT

Once upon a time there was a Little Girl who loved to swing on a swing, a Big Bad Wolf who loved to eat rabbits, and an Electrical Circuit. The Little Girl and the Big Bad Wolf were of the kinds usually found in fairy-tales: the Girl was pretty and clever, the Wolf fierce and greedy. But the Electrical Circuit hailed from an electronics textbook.

The Circuit was very proud of his origins; he stuck his nose in the air and swaggered. He knew that the things that went on inside an electrical circuit were electromagnetic in nature and that therefore not everybody could understand them. And because of this the Circuit puffed out his chest with pride and strutted to show how mysterious and ingenious he was.

One day they all met together. The Little Girl was swinging, as always, on her swing; the Wolf was lazily snapping his jaws to show how bad-tempered and hungry he was; and the Electrical Circuit swaggered and strutted.

'Stop your fidgeting!' growled the Wolf, who, it must be said, had been badly brought up: this was the politest sentence he knew.

'She is not fidgeting,' the Electrical Circuit said importantly; 'she is executing mechanical oscillations about her point of equilibrium.'

'You're a right pain in the neck, you are, Circuit. What d'you mean "oscillations"? Anyone can see she's fidgeting about with nothing better to do.'

'Stop it!' interrupted the Little Girl. 'You two are always quarrelling. Circuit is right, and I like my mechanical oscillations very much.'

'How can anyone like mechanical oscillations?' the Circuit exclaimed irritably. 'Electromagnetic ones are better than anything else in the whole wide world.'

And having said his piece, the Circuit withdrew into himself and was silent. It was impossible to tell just by looking at him whether he was working or not, because electromagnetic oscillations can only be detected with special apparatus.

The Wolf, however, had his own idea of what was better than anything else in the whole wide world. But he said nothing, and only snapped his jaws as he thought about a juicy rabbit.

'Why is the world such a wretched place to live in?' the Wolf queried thoughtfully. 'Last year the forest was chock full of rabbits... but this year they've become as rare as hens' teeth. It's got to the point where we have to bow and scrape to every lousy rabbit that happens along and fight with our own brother wolves.'

'What about the year before last?' asked the Little Girl.

'That was a tough one too.'

'And the year before that?'

'Chock full of 'em again. Hold on...' The Wolf started as a new thought struck him: 'One year empty, one year plenty—is that it? Yes, it must be... Damn those cunning rabbits with their uneven breeding habits!'

'That is another example of an oscillation,' the Circuit chimed in, 'the oscillation of population sizes.'

'What, what, what?' the Wolf snorted. 'Don't you try to confuse me with your oscillations and poppillations: I don't want to hear about it. All I want is a full belly.' (The Wolf was a cynic into the bargain.) 'Anyway, everything's oscillations with you: first you have *your* oscillations, then the Girl has *her* oscillations, and now we hear the rabbits have oscillations. Next thing you'll be digging up some sort of oscillations for me too, I suppose,' the Wolf concluded sarcastically.

'Naturally,' said the Electrical Circuit. 'The number of wolves is inversely proportional to the number of rabbits: the more wolves the fewer rabbits; and conversely, when the number of wolves decreases, the number of rabbits increases. And the result is oscillations.'

After a longish pause the Wolf growled: 'Hold on, hold on. Let's go back and start with the Girl again—her case might be easier to understand...'

At that point we shall leave our fairy-tale friends to their conversation and try to explain ourselves what is common to the swinging of a swing, the processes occurring inside an electrical circuit, and the number of wolves in the forest. At first sight they seem to have very little in common; but if we look more closely, we shall see that all three phenomena behave in an oscillatory manner.

We shall now convince ourselves that this is so.

The swing is in fact a simple pendulum (Fig. 13*a*).

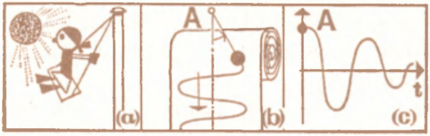


Figure 13

The displacement of the pendulum at any instant can be conveniently specified by the angle *A* between the axis of the pendulum and the vertical (Fig. 13*b*). We shall consider the angle *A* positive when the pendulum is to the right of the vertical, and negative when it is to the left. Let us displace the pendulum by the angle *A* (this will be its *initial displacement*) and release it. Under the action of gravity the pendulum moves towards its equilibrium position, namely the vertical axis, reaches it, and keeps on going by virtue of the momentum it has acquired. Then the force of gravity sends it back towards the vertical position again, and again the momentum of the bob carries it on past. And so the process goes on repeating itself, the amplitude of the motion decreasing a little each time until the pendulum finally stops.

If we fixed a small piece of slate to the pendulum bob and drew a sheet of paper along underneath it at right angles to its motion (Fig. 13*b*) the slate would leave a trace, the shape of which is reproduced in Figure 13*c* in the form of a graph. This is the typical graph of a damped oscillation (that is, one that gradually dies away).

We now note that there are two essential factors responsible for the oscillation's taking place: first, the force of gravity tending to pull the pendulum towards its equilibrium position all the time; and secondly, the pendulum's inertia (or momentum) tending to maintain its motion. The interaction of these two opposing tendencies produces the pendulum's oscillatory motion.

The first of these two tendencies we shall call the *tendency to stability*; the second, the *dynamic tendency* (the tendency towards motion).

Now let us examine the electrical circuit. It consists of two components—a condenser and a coil—connected in a series loop containing a switch (Fig. 14). The condenser has the ability to store electric charge: it becomes charged if you connect it up to a battery, the magnitude of the charge being directly proportional to the voltage of the battery.

Suppose the initial charge in the condenser is Q_0. When we close the switch, the condenser immediately starts discharging, sending an electric current through the coil. The current sets up a magnetic field in the coil. (This is why a piece of iron becomes magnetized if you put it inside the coil.)

The magnetic field thus set up induces a voltage in the coil such as to oppose any change in the current flowing through the coil. This means that the coil keeps the current flowing after the charge in the condenser has reached zero; and this current re-charges the condenser with a charge of opposite sign. So the circuit returns to its initial condition except that plus and minus have changed places. The condenser then discharges through the coil again; and so the process continues.

It is obvious that the electric charge in the condenser behaves in an oscillatory fashion just like a pendulum. Here again, it is clear that there are two tendencies that interact to produce oscillation: the tendency for the condenser to discharge and thus to reach a stable condition; and the inductance of the coil generating a magnetic field that tends to maintain the flow of current and thus to prevent the circuit from reaching the equilibrium condition ($Q = 0$). Obviously, the discharging of the condenser represents a tendency towards stability, and inductance gives rise to a dynamic tendency in the circuit.

Now for the relationship between wolves and rabbits, or the dynamics of contacting populations in general.

In biology, a population is a close-knit group of organisms belonging to the same species. So we can talk about a population of wolves and a population of rabbits. Both populations interact with each other for

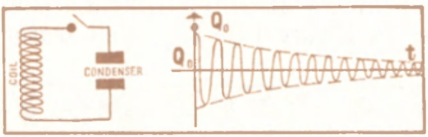


Figure 14

the simple reason that wolves eat rabbits with great relish. Let us examine the following woodland situation.

Suppose the numbers of rabbits and wolves are in equilibrium. Each time one rabbit is eaten, another rabbit is born to replace it; and each time a wolf dies, its death is compensated by the birth of a cub. This picture may not be exactly idyllic; but it is a possibility.

Suppose the number of rabbits suddenly increased as a result, say, of a breakdown in the rabbits' birth-control system. The wolves would suddenly find plenty to eat, which would favour their reproduction, and soon their numbers would also increase. As the wolves multiplied they would be eating more rabbits, so the number of rabbits would begin to drop. Eventually there would be comparatively few rabbits about, and the by now large number of wolves would find themselves facing hard times. They would begin to die of starvation and disease resulting from malnutrition until the wolf population was significantly reduced in numbers. And this in turn would result in another population explosion among the rabbits; and so on.

As we can see, an oscillation of both population sizes is set up about the equilibrium condition (Fig. 15). This oscillation is again the result of two factors. One factor is associated with the wolves' appetite and the rabbits' fecundity: these tend to keep the wolf-rabbit system in a state of equilibrium (shown by the broken lines in Figure 15). The other factor is the lag

Figure 15

between population size and living conditions. When there is a change in conditions, the population does not change immediately, but only after a time, the exact length of the delay depending on the birth-rate the population can sustain. This latter constitutes the dynamic factor.

And so the pendulum, the electric circuit, and the rabbit population are systems that exhibit oscillatory behaviour. For the purposes of cybernetics these different systems can all be treated as the same thing, thanks to the concept of the oscillator.

What is an oscillator?

By an oscillator we mean any transformation of an input X into an output Y such that a change at the input produces an oscillatory response at the output. This means that a stepwise change at the input results in the kind of output shown in the upper graph of Figure 16; the lower graph in the figure shows the output resulting from a pulse type of input.

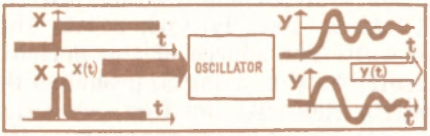


Figure 16

Clearly, this oscillator behaves in exactly the same way as the examples we have been looking at. It differs from them only in its lack of physical details. Here, only the basic idea remains—the nature of the change in the output for a given change in the input.

This is quite sufficient for the purposes of controlling such an oscillator because control is general rather than specific in nature. Let us illustrate this point using one of the above examples.

The girl on the swing can swing with constant amplitude only by applying an effort in a particular way—by pushing against a wall in front of her with her foot, for example. In terms of cybernetics we can say that pulses are impressed periodically at the input of the oscillator, the time between pulses being exactly equal to the period of the oscillation. The output of the oscillator will then be the undamped oscillation shown in Figure 17.

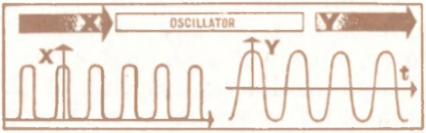


Figure 17

The girl has to push against the wall at precisely determined times, namely: whenever the swing is close to the wall. The source of the periodic pulses is the girl herself. As the swing approaches the wall she gives the wall a kick with her foot. So if the oscillator (the swing) is to produce undamped oscillations, we have to have a regulator (the girl) that uses information it receives about the output of the oscillator to determine when it should provide a pulse at the input (the kick against the wall). This system is shown in Figure 18.

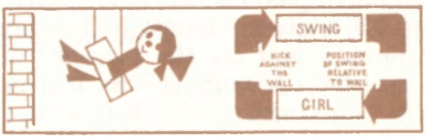


Figure 18

This is just the sort of arrangement we find in an electric generator—a source of periodic oscillations. In the generator an electric circuit, which, as we have already seen, is an oscillator, takes the place of the swing; and the girl is replaced by a regulator that converts the output oscillations into a series of pulses that are fed back into the input.

So control has a universal quality that is independent of the physical nature of the controlled object. That such a generalized approach to control processes is possible for objects of varying physical make-up was first stated by Norbert Wiener, who is rightfully called 'the father of cybernetics'.

Until cybernetics made its appearance, control processes in an electric generator were investigated by electrical engineering, control of the motion of a clock pendulum (in effect a swing) was dealt with in mechanics, and control of population dynamics in biology. Norbert Wiener was the first to point to the universal nature of control and to show that the organizing of an object (the lowering of its entropy) could be achieved by means of standard procedures, that is, by applying the methods of cybernetics independently of the physical characteristics of the object.

The development of these universal methods of control has only just begun. At the present time a process that may be called *cybernetization* is taking place in science, a process of increasing application and use of universal control methods. These methods are being worked out by cybernetics and are finding their application in various branches of science and technology for the purposes of acquiring knowledge and achieving control.

This last stage in the history of control opens up such staggering prospects for the development of science and technology that Norbert Wiener has justifiably called it the second industrial revolution.

But let us get back to chance interference and man's struggle against it. Man has had some notable successes in this field. He has developed and put into operation a wide variety of methods both for doing battle with interference and for peacefully coexisting with it. In the latter case man has developed measures that lighten the burden of coexistence.

4. THE BATTLE WITH CHANCE INTERFERENCE

The war man wages against chance has two fronts. On one front the chief weapons are various means of crushing and annihilating chance—for example, soundproofing is a defensive measure designed to prevent noise from penetrating your flat. The second front seeks peaceful means of coexistence with chance interference. These diplomatic methods allow us to develop such patterns of behaviour as tend to prevent interference from disturbing us to so great an extent. A simple example of this is the way we raise our voices on the telephone when the line is bad, and repeat individual words and phrases, and so on. Here the interference is allowed to remain at the same level, and it is we who adopt a special procedure that enables us to maintain contact despite its presence.

To illustrate this let us examine the working of a very simple communication channel between two people talking to each other as shown in Figure 19.

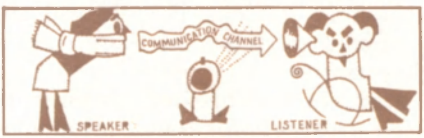


Figure 19

A system like this is liable to be acted upon by three types of interference.

Interference Number One is that due to the transmitter. It manifests itself as incorrect pronunciation of words, perhaps as a result of lisping, rhotacism, swallowing the ends of words, stammering—anything in fact that comes under the heading of bad diction.

Interference Number Two originates in the external medium: background noise, the clanging of trams, the conversation of other people, the laughter or crying of a child, and so on.

Interference Number Three characterizes the effectiveness of the receiver, here—the listener. It may result from poor hearing, poor knowledge of the language, poor eyesight (it is known that speech is easier to understand if you can watch the speaker's gestures), poor condition of the listener's nervous system (ringing in the ears, alcoholic intoxication, and so on).

All three types of interference have an adverse effect on the communication process (the conversation); and if they reach a high enough level they can actually prevent communication altogether. We shall not concern ourselves with the first and third types of interference—those inherent in the transmitter and the receiver—because they are highly specific and depend on the physical nature of the transmitter and the receiver, respectively. In order to overcome interference due to stammering, for example, it would be better to engage the services of a good speech therapist rather than a specialist in cybernetics. Similarly, cases of deafness should properly be referred to a physician. If we strike analogous problems in setting up a radio link, we should call in a radio technician. And in studying the relations between domestic animals interference is analysed by a veterinary surgeon.

Interference Number Two is the one that interests us, that is, the interference that arises in the communication channel itself. This interference has the same physical structure as the signal carrying the message; it wears the same dress as the signal, so to speak—otherwise it would not interfere with us. For example, when we are trying to carry on a conversation in a tram it is the clanking of the wheels and the cries of the conductor that get in our way, rather than the sparking at the current collector: here an acoustic channel is obstructed by acoustic interference. Similarly, an optical communication channel will be obstructed by optical interference, and a radio channel by radio interference.

Figure 20 shows two examples of useful information-bearing signals interacting with chance interference.

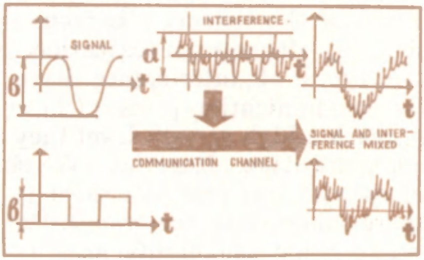


Figure 20

As we can see, the interference seriously distorts the useful signal in each case.

Every communication channel is characterized by a certain level of noise, that is, of distortion of the signal by interference (noise). In order to characterize the effectiveness of a particular channel it is convenient to introduce some means of measuring this noise level, some number that will indicate how badly the channel transmits information as a result of interference. For this purpose it is common to use, as an index of noise level, the signal-to-noise ratio obtained by dividing the amplitude of the noise (a) by the amplitude of the useful signal (b):

$$K = \frac{a}{b}$$

This quantity defines the noise level of the channel. To give us an idea of the practical significance of this quantity we may note that when $K=1$, conversation is impossible: interference stifles the signal to such an extent that the listener cannot make any sense of the message at all.

We can see from the above formula that there are two possible ways of reducing the signal-to-noise ratio:

(a) by suppressing the interference a, that is, by lowering the noise level (this corresponds to the first front of the battle with interference);

(b) by raising the level of the useful signal, in other words by increasing the amplitude b.

In either case communication efficiency will be improved.

However, the second method (increasing the strength of the useful signal) has rather severe practical limitations. It is a simple matter of fact that you cannot continue a conversation for very long if you have to shout all the time—your voice will just not stand up to it; and shout as you will, without a telephone you will not be able to hold a conversation with someone on the other side of town. And with radio, the strength of the signal is limited by the power of the transmitter and cannot be increased beyond a certain value. In other words, we may as well forget about this particular method.

But there is another way of overcoming a high level of interference. This is by increasing the redundancy of the message, that is, by repeating it and by asking the transmitter to go over places where we suspect that we have not received the message correctly, and so on. These methods constitute the second front of the battle with chance interference.

Let us examine the weapons we can deploy on each of these fronts.

A. THE FRONT FOR STIFLING CHANCE

The first and most effective measure against chance is feedback, with which we shall now begin.

FEEDBACK

We first encountered the concept of feedback in connection with Maxwell's demon. The demon represented a control system functioning so as to lower the entropy of the controlled object (the box together with the molecules of gas) by sorting the molecules according to their speeds. Here feedback consisted in the demon's observing the behaviour of the molecules and acting on his observations to operate the shutter and so to assign molecules to one side of the box or the other. As we saw, this feedback made the system 'livelier'.

In general, feedback consists in arranging for a certain action to be applied to the controlled object, this action being itself based on information received about the object's behaviour.

Feedback is extremely common in living nature. It is safe to say that animals and plants depend on it for their existence.

Let us have a look at an example of the kind of feedback that man has created for the purpose of doing battle with chance.

Man's activity is chiefly directed towards securing independence from the capricious and fortuitous whims of nature. Living in such chancy surroundings, man's first thought is to stabilize his immediate environment at the appropriate level—to procure the constancy of his microclimate regardless of the state of the weather, if you like. This is why he builds houses and kindles fires in them, the kindling of a fire for warmth being a manifestation of feedback because it is man's defensive reaction to cold aimed at providing an equable temperature within his microclimate.

In a modern dwelling the function of maintaining a constant temperature is performed by an air-conditioning system. How does it work?

Our habitation comes under the influence of two factors. One of them is the elements themselves—chiefly the temperature of the air surrounding the building, and also wind, which intensifies the severity of the weather, and humidity. All these elemental forces affect the temperature inside the house. If the elements constituted the only factor, the inside temperature would change in the same way as the outside temperature, apart from a certain time lag. But feedback serves to maintain the inside temperature at a constant level. It works as follows: the temperature in the building is measured and compared with the desired temperature, and the heating element of the air-conditioner is switched on or off as appropriate. And this constitutes feedback.

Here feedback is effected by a regulator capable of giving instructions to the air-conditioner. The regulator receives information about the temperature in the room and processes this information by comparing it with the desired temperature. Having processed the information, the regulator completes the control cycle by sending the heater an instruction. As a result, the inside temperature is kept at the desired level regardless of the chance behaviour of the elements.

In this way feedback enables us to overcome the chance factor.

Another measure for suppressing chance is the cumulative method.

SEVEN TIMES MEASURE ...

The Russian proverb 'Seven times measure, then cut' illustrates splendidly the application of the cumulative method for the purposes of combatting random interference. It is a well known fact that any process of measurement involves error. Error is essentially a chance phenomenon that interferes with precise measurement. Every measuring device, be it a rule, a clock, a thermometer or anything else, measures to within a definitely statable accuracy determined by the quality of its workmanship. The higher the quality, the more accurate the device. An ordinary wristwatch, for example, can measure a period of twenty-four hours to an accuracy of about one minute; a chronometer does the same thing to within about a second; and the most accurate and expensive timepiece we possess—the atomic clock—does it with an error of a millionth part of a second. From this we may conclude that precise measurement requires expensive equipment.

Is it, then, impossible to obtain accurate measurements with inaccurate instruments? Can we make measurements finer than those that the most accurate of measuring devices will allow?

Yes, we can.

Here, we do battle with random errors by making a large number of measurements and then averaging them. This average differs from the true value by an amount smaller than that of any of the actual measurements. In other words, the average of several measurements is always more accurate than a single measurement.

You can test this yourself by means of a simple experiment. (The writer has often performed this experiment in lectures, and invariably with complete success.) Ask your guests to estimate by eye the length of some object that happens to be handy—a pencil, for example. Write down all their estimates and work out the average. The value you obtain will turn out to be surprisingly close to the actual length of the pencil. Why is this so?

The point is that although each person gives a very rough figure for the length of the pencil, their errors can each be equally either positive or negative: some estimates will be too high, others too low. When the errors are combined by adding the estimates, they tend to cancel each other out. So when we divide the sum by the number of estimates to get the average, the result is more accurate than any single estimate.

Obviously, the accuracy obtainable by such a procedure improves as the number of measurements increases. It is therefore theoretically possible to achieve any desired degree of accuracy by repeating the measurement a sufficient number of times.

In practice, however, it is difficult to achieve a very high degree of accuracy by the cumulative method because it involves a square-root law: the gain in accuracy is proportional to the square root of the number of measurements. So that whereas four measurements are sufficient to double our accuracy, to improve it by one order—that is, by a factor of ten—we would require one hundred measurements.

Let us examine another example of the cumulative method.

In certain chemical engineering processes the concentration of a continuously prepared solution fluctuates all the time. Since these fluctuations are due to a large number of complex factors, they may be regarded as being random. The problem arises of how the average concentration of the solution should be determined if the requisite chemical analysis takes a long time. At first glance the answer appears to lie in taking frequent samples, analysing them, and averaging the results over a whole shift. This would give us a correct result, but it would be very time-consuming. There happens to be a simpler and more elegant method that allows us to obtain the same result at the expense of only a single analysis.

Here it is. We take strictly uniform samples of the solution periodically; and instead of sending them one by one to the laboratory we collect them together in a single tank. During the shift a definite number of samples accumulates in this tank. At the end of the shift the contents of the tank are thoroughly mixed and then subjected to chemical analysis. The results of this analysis will give the average concentration of the solution during the shift.

In this example averaging is achieved by stirring the contents of the tank. A single accurate analysis then gives a numerical value for the concentration of the averaged solution.

So chance factors in measurement can be successfully dealt with by the cumulative method—a general-purpose method that considerably reduces the effects of chance on the final result.

Contemplation of the cumulative method leads us immediately to consider another method for dealing with chance interference: the method known as *filtration*.

In everyday life filtration means the separation of a liquid from a mixture containing the given liquid together with non-liquid foreign matter, the latter being equivalent to interference. This is done by passing the mixture through a filter that takes the form of a fine-mesh gauze that retains the solid phase. In other words, filtration is the separation of a mixture into its two constituent phases, liquid and solid.

And it was this perfectly ordinary, everyday operation that strayed across into the field of radio telecommunications, and from there into radar. On the way, of course, the object of filtration changed from apple juice to radio signals.

We are all aware that communication channels, be they telephone wires or, for radio, simply the atmosphere, are prey to stray electrical interference that gets into the channel and mingles with useful signals. This interference is both natural and artificial in origin.

Natural interference is born of atmospheric electricity, lightning in particular. Anyone who has ever switched on a radio-set during a thunderstorm will be familiar with the dry crackle of the lightning that it picks up. Artificial interference results from the electric sparking associated with various electrical appliances, both industrial and domestic—electric welders, faulty electric motors, trams, trolleybuses, and so on. All of these produce miniature lightning discharges that clutter up our channels of communication.

All this random interference combines with the useful signal to form a highly 'unappetizing' mixture in which the useful signal is often totally submerged. It was for just this reason that science's battle against interference began with the birth of radio.

What do we mean by the 'battle against interference'? In order to understand a given communication we have to distinguish the useful signal and filter it out from the random interference. This task is performed by an electrical filter that separates the useful signal from the mixture of signal and interference. This mixture is fed into the input of the filter, and the output should be the filtered signal. And that is the basic idea of filtration.

But until the advent of radar the need for isolating the useful signal was not so sharply felt. For the purposes of ordinary communication we could rely on redundancy—repeating the message several times, for example—to increase the reliability of our communication channel in conditions of interference. Radar—the location of an object by means of a radio signal reflected from it—immediately made filtration of paramount importance, because the reflected signal is always millions of times weaker than the pulse sent out by the transmitter. The reason for this is that the original signal is reflected in all directions at once, which means that attenuation of the reflected signal takes place very rapidly. This is why radar receiving antennas are made as large as possible.

You can get an idea of the effect of very rapid attenuation from the following simple experiment. On a sunny day take a large metal billiard ball, or a ball from a ball-bearing, and a mirror of the same size, and place them side by side. (If a suitable mirror cannot be found, take a larger one and cover it with black paper, leaving a hole in the paper of the same size as the ball.) The ball represents a radar target, the sun a radar transmitter, and your eye a radar receiving antenna. We still have to account for the mirror: it represents an ideal reflector, reflecting the entire signal in a single direction without scattering (actually it does scatter light to some extent; but the amount of scattering is very small and can be ignored for the purposes of this experiment).

Suppose we set the ball and the mirror at eye level and arrange the mirror so that its reflected beam is horizontal and is therefore always easy to see. Then we step slowly backwards away from the ball and the mirror, keeping the beam from the mirror always in view. After a few steps we notice that the patch of light from the mirror is still nearly as bright as ever, whereas the sun's reflection is hardly visible in the metal ball at all: a few more steps and it disappears from view altogether, while the mirror goes on shining as brightly as before.

A radar signal reflected from an aeroplane fades out in exactly the same way. And the smaller the reflected signal gets, the more easily is it swallowed up by the interference that abounds in both the atmosphere and the receiving apparatus itself. But this is not the only difficulty. When the enemy discovers that he is in the beam of a radar transmitter, he takes countermeasures by simulating interference with the object of making the radar monitor's task all the more difficult.

As a result, the reflected signal picked up by the receiver is so small, and the interference is so great, that only very reliable filtration can enable the radar to function effectively.

Yet how can we filter out the signal?

There are several methods for filtration; and we shall now examine some of them.

The smoothing filter

This type of filter makes use of the averaging procedure that we have already described for taking an average over a certain time interval T. It works as follows.

Suppose a continuous signal is fed into the input of the filter. At an instant of time t_1 the filter outputs a signal equal to the average value of the input over the time interval from $(t_1 - T)$ to t_1. In Figure 21

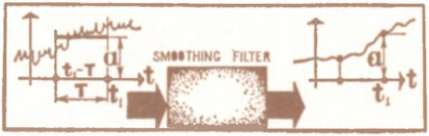


Figure 21

we see that the effect of such a filter is to smooth out the incoming signal. This is what we would expect, because any averaging process smooths or levels out the factual data: it is this that enables us to overcome any interference that takes the form of random vibrations. In radio engineering, vibration interference is known as 'white noise'. It consists of a mixture of various oscillations, just as white light results from the blending of different colours.

However, although averaging enables us to overcome white noise interference, it also distorts the basic signal. The effect of this distortion is to 'smear' the signal out along the time axis, as it were. The shorter the signal the worse the distortion, because the main effect of the filter is to eliminate vibrations—and a short signal is very similar to a single interference vibration.

Consequently, in separating the useful signal from the random interference, the smoothing filter also distorts the signal.

This defect is absent in the correlation filter.

But first we shall have to explain what we mean by *correlation*.

Correlation

The word 'correlation' signifies the presence of an interrelation. If two phenomena are mutually united by something and if they are somehow interconnected, we say that they are correlated. By defining the correlation in a particular case we can protect ourselves to some extent from chance.

Let us examine the concept of correlation in the light of the following example. We all know that when we are invited to look through an album of family photographs we are expected to delight our hosts by guessing their close relatives; and that it is easy enough to oblige because close relatives usually resemble one another: their faces correlate. We shall express the resemblance between two faces in the form of a number K having values between zero and unity. Zero indicates that the faces are completely unlike; and unity that they are absolutely identical, like twins. Intermediate values of our 'coefficient of resemblance' indicate appropriate intermediate degrees of similarity. And since similarity is a reciprocal relationship, we may call K a correlation coefficient.

The question is how to determine values for this correlation coefficient. We shall do this by using the cumulative method we discussed earlier.

We take three photographs—one each of a son, his father, and his grandfather (taken at the same age for each, if possible, so that grandfather's whiskers and grandson's tousled hair do not complicate the issue)—and ask our friends to estimate the degree of similarity between them, that is, to give their values for the correlation coefficients for each possible pair of the three photo-graphs according to the following scale:

Identical	$K = 1$
Very similar	$K = 0.75$
Similar	$K = 0.5$
Scarcely similar	$K = 0.25$
Not similar	$K = 0$

We record the results in tabular form as follows:

Serial	Son—father	Father—grandfather	Son—grandfather
1	0.75	0.5	0.5
2	0.75	0.75	0.25
3	0.25	0.5	0.5
4	0.5	0.75	0.5
5	0.75	0.75	0.5
6	0.5	0.5	0.25
Average:	0.58	0.64	0.37

From these results we can conclude that the similarity between the son and the father is the same as that between the father and the grandfather because 0.58 and 0.64 are near enough to being equal. This is what we would expect since in either case we are dealing with the relation between a father and a son (father is to grandfather as son is to father).

Now we want to find out how the correlation coefficient varies from generation to generation. Let N be the generation number, positive for future generations and negative for past generations. So $N=0$ is my generation, $N=1$ is my children's generation, $N=2$ is my grandchildren's generation, $N=-1$ is my parents' generation, and $N=-2$ is my grandparents' generation.

We shall now denote the coefficient of correlation between generations zero and N by $K(N)$. In other words, $K(N)$ expresses the degree to which I resemble the Nth generation. Obviously, $K(N)$ has the following property:

$$K(N) = K(-N)$$

which means that I resemble the Nth generation to the same extent as the Nth generation resembles me.

We shall begin our deliberations with respect to the son, the son being myself. Since I represent the zeroth generation ($N=0$), the correlation coefficient will in this case be equal to unity (the maximum possible value) because I am more like myself than anybody else. My father has a correlation of 58 percent with me ($K=0.58$), my grandfather 37 percent ($K=0.37$). Therefore, the correlation graph for me and my forebears takes the form shown in Figure 22.

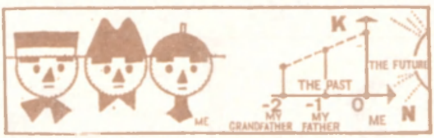


Figure 22

Now we shall examine the situation with respect to the father, so that I am now the father. I correlate with my son to 58 percent, and with my father to 64 percent. My relationship with neighbouring generations is therefore as depicted in the graph of Figure 23.

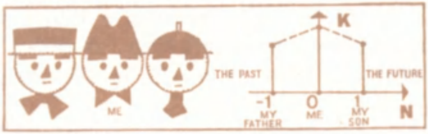


Figure 23

Now suppose I am the grandfather. The correlation graph then takes the form shown in Figure 24.

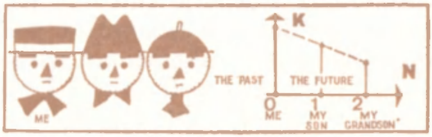


Figure 24

If we take a good look at these three graphs, we can see that they are segments of a single, larger graph for a succession of generations. This is the graph shown in Figure 25. It is a symmetrical graph, which indicates

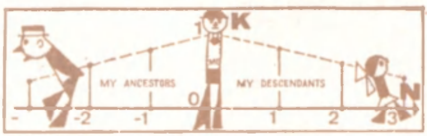


Figure 25

that the process of inheriting external characteristics is the same for both the past and the future. For very large values of N, that is, in the distant future, K will be equal to zero. In other words, my distant descendants will not resemble me at all, which is a fairly logical conclusion that nobody could object to. For large negative values of N the graph behaves in exactly the same way, again revealing the undeniable truth that my distant forefathers did not bear the slightest resemblance to me.

The relationship that we have been looking at between the correlation coefficient and time is called a *correlation function.* Such functions are extremely common and are very useful for studying the phenomena of our chancy world because they show how chance processes are related to time.

The whims of fashion, for example, are found to be not matters of chance at all, but to have a definite correlation. We can set up a correlation graph linking present fashions with those of the past. The graph we obtain has the characteristic shape shown in Figure 26, containing a number of peaks. The peaks of the

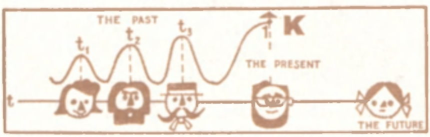


Figure 26

graph indicate points of resemblance between current fashions and past fashions at times t_1, t_2 and t_3 years ago. Experienced fashion designers are well aware of this relationship and use old vogue magazines as a source of inspiration, working on the principle that 'new' means 'old and long forgotten'. And this gives rise to similar oscillations in fashions.

An interesting example of the correlational method is its application to the analysis of interdependence between letters in a word and between words in a phrase. Let us determine the correlation between letters in individual words and between words in separate sentences.

Suppose we perform a simple experiment. We take a dozen or so words at random and ask someone to guess each letter of a word in succession. If there were no connection at all between the letters, the correlation coefficient (the proportion of correct guesses) would be close to zero. But experience shows that the proportion of correct guesses is large—on the average more than fifty percent of all attempts. This means that the letters of a word exhibit an obvious correlational interdependence; this is graphed in Figure 27.

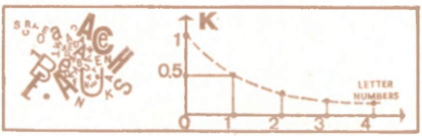


Figure 27

The presence of correlation makes our language redundant and enables us to guess words quite easily even when there are a large number of mistakes and misprints. This redundancy proves to be a sound defence against chance interference. For instance, if we receive a telegram that concludes with the words 'munch love', we can make out easily enough that the sender wishes to convey the abundance of his love rather than its edibility.

If we carry out a similar experiment to determine the correlation between words in a phrase, the result (the probability of a correct guess) is not quite so impressive; but it is sufficient, nonetheless, to indicate some degree of correlation. The probability of correctly guessing a word in an average Russian text is about one tenth, that is, about one word in ten can be guessed correctly.

True, this figure varies widely depending on the nature of the text. Technical literature is characterized by a high level of redundancy, which facilitates rapid scanning of technical material and also enables one to read technical articles in an unfamiliar language. Particularly high redundancy characterizes the exchanges that take place between an airline pilot in the air and the airport flight controller on the ground. In this case the use of redundancy is prompted by the severe and possibly tragic consequences of a mistake; so the probability of mistakes is made extremely small by using a high level of redundancy. The lowest level of redundancy, that is, the least correlation between words, is found in the language of creative literature; non-conformity, expressiveness and unexpectedness are, after all, part of a creative writer's stock-in-trade.

It is interesting to compare the redundancy levels of written material and oral speech. The latter is found to have the greater redundancy. Indeed, in ordinary conversation we tend to indulge in a good deal of repetition with little care for niceties of style and use a great many superfluous words whose object is to give the speaker time to think about what to say next.

On the other hand, living speech possesses possibilities unknown to written language. It contains such additional aids to understanding as stress, intonation, and the peculiarities of the individual voice. The information contained in these devices alone may be as much as fifty to seventy percent of the basic lexical content of the spoken message, and this tends to decrease the redundancy of the spoken word. Thus the sentence 'What have you done?' has different meaning depending on which word is stressed.

Now let us see how correlation can be used to filter a useful signal from a background of interference.

The correlation filter

The chief peculiarity of the correlation filter is that it uses information about the form of the received signal (whereas the smoothing filter does not). Details regarding the form of the useful signal contain an unusually large amount of information and enable us to separate out the useful signal with a fair degree of reliability.

We all know that it is pointless to go into the woods looking for mushrooms and berries at the same time; yet quite a lot of enthusiasts try to do just this—and finish up finding neither mushrooms nor berries. Whenever we decide to look for something, that something is a particular object having particular characteristics. Knowledge of these characteristics enables us to find what we are looking for quickly. But if we start looking for several things at once, all having different properties, our search will be ineffectual. It is better to look for one thing and find it, and then go on to look for the second thing, and so on.

My wife informs me that her searches for new apparel in fashion magazines have to be prosecuted in accordance with the above principle: first you flick through the pages looking for a suitable evening dress; then you flick through a second time in search of a new bathing costume; and so on. But if you try to find everything you need at once, you stand a fair chance of finding nothing at all however many times you go through the magazine.

But to return to the correlation filter. The idea of correlation is both simple and elegant. It involves determining the correlation coefficient for the signal we receive and the signal we expect. The latter we may call the standard signal: information concerning this signal is pre-loaded into the filter. If the correlation coefficient is large, it means that a correlation exists between the received signal and the standard signal—in other words, they are related, and the communication contains a useful signal; if the coefficient is small, there is no useful signal.

How is the coefficient of correlation between the actual and the standard signal determined?

All the correlation filter has to do is to multiply them together and average the result. This it does by passing the product of the two signals through a smoothing filter. The smoothing filter outputs the required correlation coefficient. We then have only to decide whether to regard the coefficient as large, thereby acknowledging the presence of the useful (standard) signal in the received pulse, or small, meaning no useful signal. This last step is performed by a decision-making device; we shall examine how this works later.

Figure 28 shows a block diagram for a correlation filter. Here the product block multiplies the received

Figure 28

signal by the standard signal, and the decision-maker only outputs a signal if the corresponding correlation coefficient is large.

In order to get a better idea of what the correlation filter does, let us look at the examples sketched in Figure 29. From them we can see how the trick of

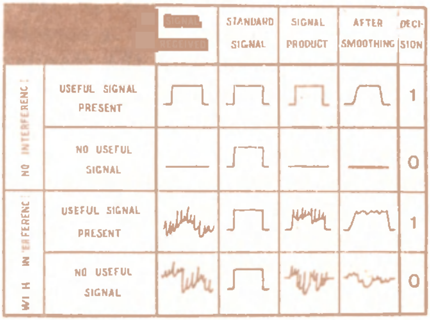


Figure 29

multiplying the signals enables us to overcome the difficulties associated with interference to a significant degree. And this is the advantage of correlative reception.

Correlative reception of information is extremely common in ordinary life and we use it widely. Indeed, knowing the signal we are supposed to receive is the same thing as expecting to receive it, in which case it is almost impossible to pass over it without recognizing it. The importance of internal organization for the reception of a given piece of information is well known. For example, if we are looking for a particular friend in a crowd, we may bump into dozens of acquaintances without even recognizing them; but we shall see the person we are looking for while he is still a long way off. This is the essence of correlative reception in which interference is filtered out to make way for the expected information. In this case our acquaintances constitute interference that complicates the business of finding the person we are looking for; but our internal filter blots them out.

B. PEACEFUL COEXISTENCE WITH CHANCE INTERFERENCE

As we have already seen, no method of diminishing the effects of chance interference can eliminate it entirely. We are always left with a certain amout of random distortion which we have to take into account. Even after all our attempts to suppress chance, our thrice chancy world remains chancy—to a lesser extent, perhaps, but still chancy all the same.

It is natural, therefore, to ask whether in such an incorrigibly chancy world we will ever be able to obtain very precise information or to perform very precise actions. To put it in a nutshell: can we ever act in circumstances involving interference without incurring even a very small number of errors? Or are we doomed to live joyless lives made bitter by a consistently high proportion of blunders?

The history of man's development has already provided answers to these questions. This is particularly true of the history of our means of communication, namely: speech and writing. It should be obvious that in a world subject to random interference (non-random interference presents no problem—we can always adapt ourselves to it and thus shut it out) man was *forced*—elementally, in the process of evolution—to devise reliable means of communication that would permit of almost error-free social intercourse. The universal means for doing this is *redundancy*, which we have already touched on above.

What do we mean when we talk about redundancy in connection with organizing a communication channel? Here, we are chiefly concerned with such a system of coding a message as will enable us to correct any errors that arise during either transmission or reception.

There are two ways of approaching this problem. One approach is to add to the transmitted message a particular control sign that enables us to check whether the message has been received correctly. If there is an error and it is detected by virtue of the control sign, the addressee can query the transmission by asking the originator to repeat the segment of the message containing the error. A communications system that operates on this principle is called an *originator referral system*.

An example of such a system is the telegraph. An ordinary telegram always indicates the number of words it contains. This number functions as a control sign because it can be used to check, albeit roughly, whether the telegram is correct. Indeed, by counting the number of words in the telegram and comparing it with the control number we can quickly check the reliability of the telegraph. If the actual number of words is less than the control number, then obviously some words are missing from the telegram. True, this constitutes a very crude test because it fails to distinguish between two different telegrams having identical numbers of words. Nevertheless it is still a check, however crude.

The other approach to the problem of ensuring reliable communication depends on the use of a special code that not only shows up errors as they occur, but at the same time enables us to correct them independently without referring them back to the originator. Codes having this property are called *self-correcting* or *error-correcting codes.*

Ordinary human language provides an example of such a code, because we can generally correct spelling mistakes in a message without referring back to the person that sent the message (and possibly made the mistake). Thus we can confidently amend a received 'sope' to read 'soap'; but to do this, our corrective mechanism has to be familiar with the grammatical rules and exceptions of the English language.

Let us examine each of these approaches separately.

THE ORIGINATOR REFERRAL CHANNEL

The basic problem here is the detection of error, because the actual business of referring back to the originator does not in itself present any new difficulty. How can we code a message so that it will reveal to us any errors it contains?

We shall examine a code that is known as binary because it makes use of only two symbols: 0 and 1. Suppose 0 corresponds to the absence of a signal in the communication channel (a pause) and 1 corresponds to a signal. (We may note in passing that the familiar Morse code—the famous Morse alphabet—is a typical example of a ternary code, a code that employs three symbols: dot, dash, and pause.)

Every code consists of blocks, each containing the same number of symbols. An alphabet code, for example, might consist of blocks of five symbols, as follows:

$$A = 00001$$
$$B = 00010$$
$$C = 00011$$
$$\cdots\cdots$$

$$X = 11000$$
$$Y = 11001$$
$$Z = 11010$$

This code is completely devoid of redundancy. A message such as BY CAB, for example, which in code is: 00010 11001 00011 00001 00010, loses its meaning if there is a single mistake in any of the signals. In other words, this type of code does not show up errors. How can we increase the redundancy of this code so that it will allow us to correct any errors that may arise?

The first and simplest idea that comes to mind is to duplicate each signal, that is, to transmit each signal twice. The above message would then look like this:

0000001100 1111000011 0000001111 0000000011
0000001100

Such a procedure would certainly enable us to detect errors whenever we noticed that the signals in any would-be pair were different. However, all our messages would then contain twice as many symbols; and this is too high a price to pay for the resultant gain in communication reliability. Suppose we try to assign numerical values to the pros and cons of duplication.

The effectiveness of deliberately introduced redundancy is characterized, naturally enough, by means of the following two quantities:

1. The number (or percentage) of undetected, and therefore uncorrected, errors.

2. The percentage increase in message length.

Obviously, a good code will be one for which both these numbers are sufficiently small.

Straight-out duplication gives us an increase in message length of 100 percent because each message is doubled.

Now let us determine the number of undetected errors passed by a duplicated code.

Suppose interference in the communication channel completely reverses, on the average, one transmitted signal in a hundred (reversing a signal amounts to introducing an error). The channel is then said to operate with a probability of error of one percent (1/100).

With duplication an error will only remain undetected if both *signals of a given pair*—the basic signal and its duplicate—are changed by interference. The probability that *one* of them will be changed is 1/100; but any single error will be speedily corrected by referring back to the originator: the referral command will appear as soon as a basic signal fails to coincide with its duplicate. But if both signals of a pair are the same (both wrong), referral will not take place and the error will pass unnoticed.

Obviously, this event—two errors in succession—occurs much less frequently than a single error: one hundred times less frequently in this case. Consequently, duplication means that one error in a hundred will pass undetected on the average; the remainder will all be rectified. In this example, then, duplication reduces the number of undetected errors by a factor of one hundred. This is very good going; but the price we have to pay—doubling the length of the message—is too high.

So although simple duplication has its uses, it is not the best method for increasing redundancy. Let us now consider another, more economical, method.

Suppose we add one more symbol to each code block thus: a 1 if the sum of the five original symbols is an odd number; and a 0 if it is even (zero is also to be regarded as an even number). From our original code we then obtain:

For the letter A: $0+0+0+0+1=1$, an odd number, so the new block is $A=000011$.

For the letter C: $0+0+0+1+1=2$, an even number, so the new block for C is $C=000110$.

Following this procedure right through, we obtain our new code as follows:

$$A = 000011$$
$$B = 000101$$
$$C = 000110$$
$$\cdots\cdots$$
$$X = 110000$$
$$Y = 110011$$
$$Z = 110101$$

This code now contains what we call a control sum (the last symbol in each block). This sum enables us to check the transmission accuracy of the message by checking each block for evenness and comparing the result with the control sum.

Interference in the communication channel or in the receiver or the transmitter can reverse a signal, that is, turn a 0 into a 1, or a 1 into a 0. The evenness test obviously enables us to spot single errors because a single error in a block makes an even block odd and an odd block even. But if two errors occur in the same block, there is no change in the evenness or oddness of the block and the error passes undetected. Three errors in the same block will again trigger the originator referral signal, and the mistake will be corrected. Four errors in one block will pass unnoticed. And so on.

Consequently, the evenness test does not eliminate all errors. A proportion of errors still gets through undetected. Naturally, we want to know how large a proportion, that is, what percentage of errors do we fail to detect if we use this method of introducing redundancy.

Let us therefore determine the percentage of undetected errors characteristic of the control sum method. As in the previous case, we shall again assume that the communication channel has a probability of error of 1/100, that is, on the average one mistake occurs for every hundred symbols transmitted correctly.

Let us consider a single block. It consists of six symbols now, rather than five, the extra symbol being the control sum, which also has to be transmitted correctly. Suppose one symbol in the block has been transmitted incorrectly. If all the rest, including the control sum, are correct, there will be a discrepancy in the evenness calculation. This will immediately trigger the control mechanism that gives the command for originator referral, and the error will be eliminated as a result.

But if one of the remaining five symbols is also wrongly transmitted, there will be no discrepancy in the evenness figure and the error will pass unnoticed. How often does this happen?

If the probability of error for one symbol is 1/100, then for five symbols it is approximately five times as great: the number of possible occasions on which errors may occur increases; hence there is a corresponding increase in the overall probability of error, to 1/20 in this case. This means that one block containing two errors occurs for every twenty blocks containing only one error (on the average).

In other words, this method of introducing redundancy decreases the number of errors by a factor of twenty.

Consequently, the evenness test makes the code twenty times more reliable. Correspondingly, the effects of random interference are twenty times less severe. At the same time it is clear that we have made no attempt to reduce the level of interference in the communication channel itself. We have achieved this magnificent result solely by virtue of an efficient method of coding, that is, by incorporating an evenness check in our code blocks. In the process the length of our messages has increased by only twenty percent (one additional symbol to each original block of five).

These methods of introducing redundancy make for a considerable improvement in the reliability of a code at the expense of a certain increase in code length. They can therefore be used to good effect for transmitting messages in a communication channel that is subject to random interference.

The advantage of using this kind of redundancy is obvious in the case of the telegraph. However, this method can also be used in other situations which, on the face of it, have nothing in common with the telegraph. One such application of redundancy is found in connection with the modern high-speed computer.

Computers offer tremendous scope for the use of redundancy. The need to incorporate redundancy is dictated by two vital factors: (1) the human being tending the computer; (2) the unreliability of the computer itself.

In order to understand how the fallible human being in charge of a computer influences the effectiveness of its operation we need to know how information is fed into a modern computer.

To solve a given problem the computer must be provided with information telling it *how to calculate* and *what to calculate*.

The first requirement means writing a programme of operations that the computer has to follow in order to solve the problem and then introducing this programme into the computer in the form of a numerical code similar to those whose redundancy we have just been examining.

For example, if we wish to solve the equation

$$ax^4 + bx^3 + cx^2 + dx + e = 0$$

using a general-purpose computer, we have to compile a programme that will tell the computer how to solve the equation, and then we have to feed this programme into the machine. The second type of information fed into the computer tells it what to calculate and consists of all the initial data needed for the calculation. For solving the equation above, these data would be the values of the coefficients a, b, c, d, and e.

All this information can be fed into the computer in various ways: either the computer operator can do it himself, or it can be done by means of punched tape or punched cards.

The operator sitting at the console constitutes the least efficient means of feeding information into the computer because he is far too slow.

Punched tape consists of a paper or celluloid ribbon punched with holes (or perforations) that carry the necessary coded information into the computer. Each hole in the tape corresponds to a 1 in the binary code, and the absence of a hole to a 0. The tape is placed in the special input unit of the computer where it is passed at high speed between rows of lamps and photoelectric cells. As each hole passes between a lamp and a cell it allows a pulse of light from the lamp to impinge upon the cell, which in turn delivers a pulse of electric current to the storage banks of the computer; and this corresponds to a 1 being entered into the computer. The absence of pulses like these corresponds to a 0.

This is a very efficient way of feeding information into a computer because the tape can be run through the input unit at high speed.

Finally, there are punched cards. These are made of ordinary paper card and are about three times the size of ordinary playing cards. They, too, have holes carrying the coded information. A 'pack' of punched cards contains all the information needed for a given calculation. At present, punched cards represent the most efficient way of getting information into a computer, the reason being that the input unit reads the entire contents of each card 'at a single glance', and this considerably improves the rate of information input. In addition, a programme punched on cards is easy to alter, because it is simply a matter of replacing one or two cards with new ones. Punched tape, on the other hand, has to be cut and spliced—and with great precision too, because of the high speed with which it passes through the input head.

Punched cards are excellent in every way. But—

There is alas! one 'but' ... .

A SAD STORY WITH A HAPPY ENDING

The fact is that the circumstances in which the holes in the cards are punched are pervaded by the 'human factor'. More often than not, this work is done by nice, quiet girls who have generally just finished high school and are interested in everything that should interest young girls in the tender springtime of life. They sit at their perforators (card-punching devices that are capable of making the holes, but not of deciding where to put them) looking at a programme devised by a programmer (also a sinful mortal) and jabbing away at the keys in front of them, thus producing the requisite holes in the cards.

Since this is rather monotonous work, the girls generally chatter among themselves about every subject under the sun. From time to time they remove from the perforator a finished card containing essential information for the computer together with ... errors.

If we were to take these cards and feed them straight into a computer (as in fact is done), the lives of the computer operator and the programmer would be immediately clouded over with gloom (as in fact happens). Everything would seem to be in order: the programme has been checked several times; the computer is in top form and passing all its tests without the slightest error—and yet it still refuses to proceed with the calculation.

And who is to blame for all this but the nice girl at the perforator with other things on her mind. So she hastens to oblige with a new card, discussing the latest film all the while and making fresh mistakes.

If we examine this situation more closely, it is easy to see that the perforator girl constitutes a communication channel between the manuscript programme and the computer (Fig. 30). This channel is beset by a

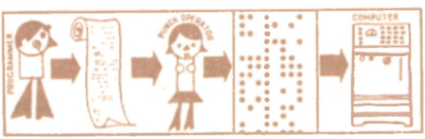


Figure 30

peculiar kind of chance interference all of its own, which is very difficult to suppress. If you bawl the girl out, she bursts into tears; if you neglect to give her a bonus, she gives notice to quit. So we are forced to reconcile ourselves to this source of random errors in our punched cards and to make provision for their detection while the programme is being run on the computer. And this is a very difficult problem.

But if the only difficulty is a communication channel with random interference that is impossible to filter out, why not use a code containing an evenness test such as the one we have just been discussing? After all, this would take care of nearly all the errors arising within the channel.

And so it was realized that an evenness test had to be introduced on the cards themselves. To get an idea of how this is done, let us see what a punched card looks like.

A punched card usually consists of eighty-eight vertical columns. Each column contains a number of holes carrying coded information, the exact code that is used being unimportant (Fig. 31).

Figure 31

The bottom row of the card is set aside for control purposes. A hole is punched in this row if the sum of the holes in the corresponding column is odd; no hole is punched if the sum is even. The control row is worked out beforehand by the programmer and is written into the programme by him.

The perforator incorporates a simple device for checking the evenness of the columns against the indication in the control row. If the check reveals an error, that is to say, if the sum of the holes in any column is even and there is a hole punched in the control space, or the other way round (an odd sum and no hole), a bell rings to indicate that someone, either the programmer or the punch operator, has made a mistake.

It is easy enough to find out who is responsible by doing a little counting; and then either quietly repunching the card or taking the programmer to task for not knowing the difference between odd and even numbers.

And everybody is happy. The programmer is happy—no more poring over a programme looking for other people's mistakes. The punch girl is happy—people stop cursing her and she even gets a chance to give the programmer a talking-to. The computer operator is happy—no longer does he have to content with a machine that is forever 'running wild'. And the maintenance engineer is happy—there are fewer insults hurled at his computer, whereas before, at the slightest hint of trouble, everyone gathered round and insinuated all sorts of nasty things about his precious machine.

CAN WE PERFORM CORRECT CALCULATIONS ON A COMPUTER THAT MAKES MISTAKES?

I once heard a most interesting opinion expressed in connection with this question. My colleagues were discussing high-speed computers and their inevitable aberrations—aberrations that sour the natures of all who come in contact with these machines. Let us consider, as an example, a computer that goes astray about once per working hour. In real life, however, the problems we have to solve require very often tens, if not hundreds, of hours of computer time. As a result, the solution to any problem is bound to contain an error. From this the conclusion is often drawn that we must either forget all about such 'superproblems' or else construct machines that err only once in a hundred to a thousand hours of continuous operation (such machines are very expensive—one has to pay a great deal of money for reliability).

Such machines are doubtless necessary; but are we correct in thinking that superproblems cannot be solved by computers that make mistakes 'only' a hundred times more frequently? Let us see whether we can work out how to go about solving a lengthy problem with an unreliable machine.

The usual method in such a case is to repeat the calculation. It seems at first as though all we have to do is to run the calculation through the computer several times until we get a pair of solutions that agree, which we should then regard as the correct result.

But suppose the probability of getting a correct solution is small—which is indeed the case with very involved problems. The computer would have to spend far too long a time repeating the calculation before it came up with two identical solutions.

Let us assume that the computer makes one mistake per hour on the average, and that we have a problem that will take five hours to solve (five hours assuming faultless operation of the computer, that is). The probability of obtaining five error-free hours of operation in succession is 1/32. (We reckon this as follows: the probability of no error in the first hour is one half; the probability of no error in the first two hours is half that, in other words a quarter; and so on.) This means that to get one correct solution to our five-hour problem the computer would have to repeat the calculation thirty-two times on the average. The time it would take to do this would be roughly $32\times\times5=160$ hours.

Working seven hours a day, it would take the computer over a month to solve one five-hour problem. Here, indeed, is food for thought. Perhaps the sceptics were right in saying that problems like this could not be solved on error-making machines.

If we look into it, however, this figure of 160 hours turns out to be an unjustifiable and uneconomical waste of time brought about by mental laziness. Do we have to repeat the entire calculation? Perhaps all we need to do is to repeat small bits of it that are likely to contain an error. And this is the key to the whole problem.

We break the calculation up into a number of consecutive stages. We begin with the first stage and repeat it until we get two identical results for it. We can then be certain that the first stage has been solved correctly because the probability of identical errors is practically zero and can be safely ignored. Then we proceed to the second stage of the calculation, repeating it in the same way until we have two identical results; then to the third stage; and so on.

We shall illustrate the effectiveness of this approach, using the same five-hour problem and the same computer making the same fatal mistake on the average once per hour. Suppose we can break the problem up into five stages each requiring one hour (of faultless operation) for solution. During each stage the probability of error is one half; consequently, we expect to have to run each stage through the computer twice, on the average, in order to obtain one correct result. Therefore, for two correct results we would have to perform the calculation four times, so that the total time spent in obtaining a guaranteed correct solution would be $4 \times 5 = 20$ hours, which is but one eighth of the time that we needed in order to obtain one unverified solution to the same problem by the first method.

But this is still not the best method.

Unreliable machines can be used to solve much larger problems than this. It is only a matter of knowing how to break the calculation down into the optimum number of stages.

And so, the application of special methods for solving lengthy problems enables us to get reliable results from unreliable computers. This is another example where chance factors are overcome not by suppressing chance (the computer may continue to make just as many mistakes as it ever did), but by organizing the computer's work in a special way.

The problem of getting reliable work from an unreliable machine may be likened to the problem of communication by means of a very 'noisy' channel. For the latter we require a high level of redundancy in the transmitted message; but it would be most decidedly not to our advantage to repeat the entire message over and over again until two identical versions of it were received through the clatter of interference, as we have already seen. It is much better to break the message up into a number of blocks (stages), each of which is transmitted as often as is necessary to ensure correct reception, that is, until two identical versions of each block have been received. Obviously, for each message there is an optimum number of blocks and an optimum block size that will guarantee the correct transmission of the message in the minimum of time. We must remember, however, that such a high level of redundancy is only needed when we are dealing with a communication channel plagued by a particularly high level of interference.

Let us now have a look at error-correcting or self-correcting codes. These codes have an undoubted advantage over those that require an originator referral system. Indeed, for the latter we have to have a reverse connection—a sort of feedback system—in order to refer queries back to the originator; and this is rather an expensive luxury because it involves duplicating all our transmitting and receiving equipment, as well as interruptions throughout the transmission to ask and to answer questions.

'SELF-HEALING' CODES

Self-correcting codes provide a typical example of self-restoring redundancy because they contain information about how to restore the parts of a message that are faulty as a result of random interference.

As before, we shall deal with a code consisting of separate blocks of symbols, bearing in mind that we only need to examine the self-resorting properties of a single block.

Suppose the block contains k symbols which can be written out as a sequence thus:

$$a_1, a_2, \ldots, a_k$$

where each symbol a_i represents, as usual, one of two values—either 0 or 1:

$$a_i = \begin{cases} 0 \\ 1 \end{cases}$$

for all values of i ($i=1, 2, \ldots, k$).

For our originator referral channel we would add an additional symbol b to this code to describe the evenness or oddness of the block. In terms of the block symbols, b has the following values:

$$b = \begin{cases} 1, \text{ if } a_1 + a_2 + \ldots + a_k \text{ is an odd number} \\ 0, \text{ if } a_1 + a_2 + \ldots + a_k \text{ is an even number} \end{cases}$$

and the new block can be written thus:

$$a_1,\ a_2,\ \ldots,\ a_k,\ b$$

The evenness symbol b has simply been added to the right-hand end of the block.

As we have seen, this measure enables us to detect the presence of errors in the block so long as the number of errors is odd. But to find out exactly whereabouts in the block the error lay, we had to refer back to the originator; and this meant we had to stop the transmission and spend time correcting the error—and to do all this we also had to have additional equipment for a reverse connection.

Now, to develop a code that will allow us to correct errors without referring back to the originator via a reverse connection, we shall start with a basic code block of a particular size: $k=12$, say.

$$a_1,\ a_2,\ \ldots,\ a_{11},\ a_{12}$$

Next, we rearrange this basic block in the form of a table or, as it is more properly called, a rectangular array.

a_1	a_2	a_3	a_4
a_5	a_6	a_7	a_8
a_9	a_{10}	a_{11}	a_{12}

And now we determine the oddness or evenness of each row and each column of this array, and add the appropriate symbols along the right-hand and lower edges, thus:

a_1	a_2	a_3	a_4	b_1
a_5	a_6	a_7	a_8	b_2
a_9	a_{10}	a_{11}	a_{12}	b_3
b_4	b_5	b_6	b_7	

Here b_1, b_2, and b_3 indicate oddness or evenness of the rows; b_4, b_5, b_6, and b_7—that of the columns. In symbols:

$$b_1 = \begin{cases} 1, & \text{if } a_1 + a_2 + a_3 + a_4 \text{ is an odd number} \\ 0, & \text{if } a_1 + a_2 + a_3 + a_4 \text{ is an even number} \end{cases}$$

$$b_2 = \begin{cases} 1, & \text{if } a_5 + a_6 + a_7 + a_8 \text{ is an odd number} \\ 0, & \text{if } a_5 + a_6 + a_7 + a_8 \text{ is an even number} \end{cases}$$

. . . .

$$b_7 = \begin{cases} 1, & \text{if } a_4 + a_8 + a_{12} \text{ is an odd number} \\ 0, & \text{if } a_4 + a_8 + a_{12} \text{ is an even number} \end{cases}$$

Now, by writing out the rows of this supplemented array one after another in the form of a sequence, we obtain our new block with redundancy:

$$a_1\ a_2\ a_3\ a_4\ b_1\ a_5\ a_6\ a_7\ a_8\ b_2\ a_9\ a_{10}\ a_{11}\ a_{12}\ b_3\ b_4\ b_5\ b_6\ b_7$$

For example, the basic block

1011 0100 0111

becomes the array

1	0	1	1
0	1	0	0
0	1	1	1

to which we add the oddness-evenness symbols:

1	0	1	1	1
0	1	0	0	1
0	1	1	1	1
1	0	0	0	

and the new block with redundance is:

10111 01001 01111 1000

And now, if we use this last array supplemented by the seven symbols $b_1, b_2, \ldots, b_7$ as our code block, we can correct errors without referring back to the originator.

Indeed, let us suppose there is an error in the basic block a_1, a_2, ..., a_{12}. The erroneous symbol will then break the evenness correlation in both the row and the column corresponding to that particular symbol. Consequently, a single error in the basic block breaks two of the evenness conditions. These pinpoint the incorrect symbol precisely and enable us to restore it to its correct value. All we have to do then is to take the symbol standing at the intersection of the row and the column containing the oddness-evenness discrepancies, and change its value to the only other value possible (remembering that each symbol can only be either 0 or 1). And that is all there is to it.

For example, suppose the following block has been received as the result of interference in the communication channel:

1	0	1	1	1
0	1	1	1	1
0	0	0	1	0
0	1	0	1	

Figure 32

Checking shows that the evenness condition has been broken in the first column and the last row of the basic block. This means that the symbol a_9 standing at their intersection has been received incorrectly and should be reversed, that is, changed from $a_9=0$ to $a_9=1$. The block is then correct.

Errors may occur, however, among the control symbols b_1, b_2, ..., b_7 as well. If this happens, only one control condition is broken, and this points up the erroneous control symbol immediately.

Suppose, for example, we have received a block that looks like this when it is written in the form of an array:

1	0	1	0	0
0	1	1	0	1
0	1	1	1	1
1	0	1	1	

Figure 33

Here the control condition has been broken only in the second row. This means that the control symbol b_2 has been received incorrectly and should be changed to $b_2=0$.

In our example we have examined the simplest case, the case where the block contains a single error. It is possible, however, to devise codes that will correct two, three, and more, mistakes. These codes incorporate additional evenness tests—along the diagonals, for example, or by taking symbols in groups resembling the knight's move in chess, and so on.

Using various oddness-evenness tests, we can devise self-correcting codes to any degree of reliability we please. In the process, however, the size of the code block increases steadily. In the example we have just been looking at, the basic block grew from $k=12$ to $k=12+7=19$ with the introduction of redundancy —an increase of sixty percent. Is this perhaps too much?

No, it is not. As the number of symbols in the basic block is increased, this percentage decreases. For a basic block of $k=100$ (a 10×10 array) we need twenty control symbols; the self-correcting block will then contain 120 symbols, representing an increase of only twenty percent. For larger k this percentage will get smaller still.

From the point of view of code economy, then, it is more convenient to have large blocks. In that case, suppose we code not individual letters, which amount to only twenty-six in all, but entire words, of which there are a far greater number. The size of each block will be increased as a result.

The above considerations enable us to assert (this was first demonstrated by the famous Claude Shannon) that for any communication channel with any level of error it is always possible to construct a self-correcting code capable of ensuring transmission reliability as close to ideal as we care to choose. Naturally, any improvement in the reliability of transmission is associated with a slowing down in the rate of transmission because the size of the code block is increased.

In conclusion, we may remark that self-correcting codes provide a brilliant example of self-restoring systems that contain sufficient redundancy to ensure their stability even in the face of considerable random interference. They constitute an extremely effective means for doing battle with chance, not by suppressing it, but by circumventing chance interference with a rational course of action that enables us to obtain reliable results despite the most 'unreliable' conditions.

5. ALTERNATIVES, RISK AND DECISION

'To be, or not to be, that is the question.' The popularity of this quotation rests on the fact that we have all asked ourselves this question on more than one occasion and have somehow or other come up with an answer after a good deal of agony and with varying degrees of success. We are only ever racked with the doubts of a Hamlet when we are faced with a choice between alternatives (to be or not to be) each of which is associated with the risk of unpleasant consequences. We can imagine how ridiculous Hamlet's position would have been if the circumstantial evidence had been coincidental and his uncle had not been his father's murderer after all. Had that been the case, the tragedy would have turned into a farce—and Shakespeare could hardly have allowed that to happen. This writer does not insist on such an interpretation either: he merely wishes to emphasize that one of the alternatives threatened Hamlet with unpleasant consequences.

The point is that often only one among all possible alternatives is correct; but lack of information prevents us from knowing which one. If we always had the necessary knowledge at our fingertips, Hamlet's question would be just as ridiculous as a man pondering deeply the problem of twice two. Since Shakespeare's time, however, our thirst for information has, if anything, become more acute. It is true that information is somewhat more accessible than it was (we can even get a few scraps such as time, weather, sports results, and the like for the price of a telephone call); but the number of questions has grown at a much faster rate.

On top of all this, the inescapable presence of random interference distorts the information available to us and robs it of much of its value. (In Hamlet's case the distorting 'interference' came from his former friends Rosencrantz and Guildenstern.)

If we give due regard to all we have been saying, we shall be forced to admit to a profound respect for the brilliant way in which Hamlet provoked and resolved a most complicated situation and succeeded in unmasking his uncle. But neither Hamlet nor Shakespeare left us with a recipe for making decisions in conditions of interference. In one particular instance we learn how to unmask a villainous uncle; in real life, however, the same situation hardly ever repeats itself. What does 'to be' amount to if the circumstances change?

So: *to be* or *not to be?* Alas! Shakespeare does not provide the answer. The answer was provided by statistical decision theory.

THE TALE OF THE FAIR KNIGHT AND THE MILESTONE

Ever since our childhood days we have all been familiar with the classic fairy-tale situation in which the difficulties of making decisions are to be felt with particular acuteness. We can re-create this situation somewhat as follows (without pretending to reproduce all the details with complete accuracy).

The fair knight mounted on his mighty charger rides up to a cross-roads where the road branches three ways. There are neither policemen nor passers-by about, no one he can ask the way. And instead of an ordinary road-sign, there is a milestone bearing the following inscription:

'Go to the right, and thy horse flees in fright. Go straight ahead, and thou losest thy head. Go to the left, and thou'lt be bereft.'

The fair knight involuntarily pushes back his helmet and scratches the back of his head—a portion of the anatomy where good people often seek, and not infrequently find, answers to the most abstruse questions. He has to make a choice from among four alternatives—to use the modern idiom: he has four courses of action open to him.

Course No. 1: To take the first road and, perhaps, lose his horse.

Course No. 2: To take the second road and, perhaps, lose his head.

Course No. 3: To take the third road and, perhaps, be stricken with grief.

Course No. 4: To turn back.

The adoption of any one of these courses presents not the slightest difficulty: a jab of the spurs, and he is off. But how is he to make the right decision—in fact, what decision should we call the right one, the optimum one, in such a situation? If only there were some hope of pleasant prospects—an encounter with a beautiful princess, for example, or, at very least, a sleeping beauty that he would have to wake—in but one of the proffered alternatives! But here, which-ever road he takes, he faces nothing but unpleasantness. What is he to do?

Now in addition to the traditional trappings (steed, spear, sword, long-bow and arrows), our fair knight is also armed with common sense. Drawing on this redoubtable resource, he reaches the conclusion that he is in a spot and that it is up to him to make the best of a bad job; in other words, he must take the road that offers the least unpleasant consequences. The wise old saying about accepting the least of one's misfortunes forms the basis of the knights intuitive—and perfectly correct—approach to making the optimum decision. At this point he has already achieved a great deal: to begin with, he has chosen his decision rule, that is, he has determined how to go about making the best decision; and secondly, he has established that the best decision will be the one that minimizes the unpleasantness.

But before he can proceed any further he has to work out how he proposes to measure the unpleasantness associated with each of the possible alternatives: he has to determine what units he should use for measuring unpleasantness and how many of these units he should allot to the consequences of each particular choice.

And so the knight reflects that *if the inscription on the stone is to be believed*, the situation he has got himself into can end in only one of four ways:

1. He can lose his horse by taking the first road.
2. He can lose his head by taking the second road.
3. He can be stricken with grief by taking the third road.
4. He can bring shame and dishonour upon himself by turning back.

'What I shall do,' says the knight to himself, 'is reckon my losses according to the number of enemies

I shall be unable to defeat in each case. In battle my charger would trample four foes: so without him I stand to lose by four units. I myself can dispatch seven: consequently, without my head my losses will amount to eleven units (seven for me personally, four for my charger, because without me to guide him he would not trample anybody—he has not reached that stage in his training yet).'

Grief has different effects on different people. Our knight decides that in his case grief would unsteady his hand to the extent that the number of enemies he could vanquish would decrease by three. His loss in taking the third road is therefore three units.

Turning back would be a sign of cowardice and result in a loss of prestige together with his knighthood —which for our fair knight would be tantamount to losing his head anyway. Consequently, this alternative also scores eleven units of loss.

Thus the knight estimates his losses, basing his figures on the assumption that the stone is telling the truth. But suppose, to put it bluntly, the stone is rather overdoing it. Things like that can happen—and not only in fairy-stories. (Here comes our interference again!) If it is, then his expected losses will be rather different.

Our knight, however, is a seasoned warrior. Many a long year has he spent in quest of adventure, and he has learnt many things about the world. So he is quite capable of evaluating the reliability of any information that comes to hand. His experience testifies to a tendency among fairy-tale milestones to exaggerate the dangers that lie ahead and tells him that the more ominous inscriptions are to be taken at only about half their face value. At this point, then, he is forced to consider his attitude to the stone's credibility.

After due consideration he decides to assign the value zero to anything he totally disbelieves, and unity to anything he can believe absolutely; intermediate values will correspond to appropriate degrees of credibility. This procedure will give him an opportunity of determining the degree of certainty that any particular prediction will in fact be borne out in practice.

So our knight decides that the credibility figure for the first alternative (loss of horse) is 0.6, for the second (loss of head) 0.4, for the third (grief) 0.9, and for the fourth (retreat) 1.0 (because he is absolutely certain to lose prestige if he displays cowardice).

We have now arrived at a most important concept: the concept of *risk*, as it applies to the making of optimum decisions.

The magnitude of the risk associated with a particular decision is determined both by the possible loss associated with the decision and by the degree to which it is obvious that this loss will in fact result. If there is little chance that the stated loss will actually occur, the risk is low; it is likewise low if the probability of loss is high, but the loss itself is small.

For example, have you ever wondered why people leave a building by the door rather than by a window? The answer is to be found in the concept of risk. The risk of breaking your neck clambering out through the window is much higher than the risk of the same thing happening as you walk through the door. Of course, on the way to the door we might fall down the stairs and break a leg: but we know that this is a trifle compared with breaking your neck, and also that the chance of its happening is small.

It is possible, of course, to climb through a window unscathed; but it is hardly likely, and the damage one could do to oneself is too severe. Certainly, one could safely refuse any invitations to dinner as one tumbled past one's neighbours' windows because the risk would be so high that although it is permissible to reckon on a happy outcome, it would obviously be impractical to do so.

These are the basic considerations underlying our choice of the best way of getting on to the street in one piece. We understand all this intuitively—and put the mat for wiping our feet on the doorstep, rather than on the window-sill.

It should be obvious by now that risk is equal to the product of the possible loss multiplied by the probability that the loss will occur. For example, if unit loss occurs on only half of the appropriate occasions (the probability of loss equals one half), the risk associated with the decision is one half:

$$1 \times \frac{1}{2} = \frac{1}{2}$$

Risk is thus equivalent to average possible loss.

Let us return to our fair knight. In order to determine the risk associated with each of his four alternatives he has to multiply each loss by the degree of certainty that it will occur. Figure 34 shows the losses together with the corresponding degrees of certainty (probability) and the risk values for each of the four courses of action open to the knight. Obviously, *the best (or optimum) course of action is the one that carries the minimum risk.*

COURSE OF ACTION	No. 1	No. 2	No. 3	No. 4
LOSS, ACCORDING TO THE STONE	4	11	3	11
DEGREE OF CERTAINTY OF LOSS	0.6	0.4	0.9	1
RISK	2.4	4.4	2.7	11

Figure 34

Indeed, by minimizing the risk we ensure that on the average our losses will be minimal. This does not mean that the actual loss on any particular occasion cannot be greater than the average value; but, by the same token, it can also be smaller. Therefore it makes good sense to base our decisions on the average expected loss, and to try to minimize it.

Now the problem of choosing which road to take reduces to a determination of the course of action carrying the minimum risk, which in this case is Course No. 1: to take the road to the right and risk losing the horse. This represents the optimum decision for the knight because it subjects him to the minimum risk. We must not be misled into thinking that he is certain to lose his horse: nothing could be further from the truth. In fact, judging by previous experience, he has only a sixty-percent chance of losing the horse; and his faith in his own prowess gives him the right to reckon on a favourable outcome.

And so our fair knight chooses the best possible course. As we have seen, he is helped in this by his experience, without which he would have been unable to evaluate the trustworthiness of the milestone. But what if he had not had the requisite experience? What if he were sallying forth for the very first time? If this were the case, there would be nothing for it but to treat all the inscriptions as being equally believable or equally unbelievable. The former would be the attitude of a sadly pessimistic hero for whom the optimum course would be the most cautious one: Course No. 3 (grief), as shown in Figure 35, because it would guarantee him minimum risk in accordance with his pessimistic reverence for the milestone's credibility. For the latter, devil-may-care attitude that 'nothing can go wrong', all possibilities other than the fourth carry zero risk and therefore present equally good prospects, so that optimum behaviour would consist in an arbitrary choice of one of the three roads; and this is the sort of recklessness that is characteristic of the optimist.

All the steps in the knight's method of solving his problem are completely simple and natural. They also

COURSE OF ACTION	N1	N2	N3	N4
LOSS, ACCORDING TO THE STONE	4	11	3	11
PESSIMISTIC VALUE FOR CREDIBILITY	1	1	1	1
RISK	4	11	3	11
OPTIMISTIC VALUE FOR CREDIBILITY	0	0	0	1
RISK	0	0	0	11

Figure 35

lie at the heart of statistical decision theory, which constitutes yet another means for overcoming the chanciness of our world.

But now we shall turn from our magical fairy-tale to a subject of severely adult interest, namely: criminology.

THE SAME OR NOT THE SAME? (A THRILLER)

Inspector Maigret shuddered. A large, cold drop of autumn rain had fallen down the back of his neck. A second drop ran along the barrel of his pistol, pausing for an instant as it negotiated the fore-sight, and hung suspended precariously from the muzzle. Maigret shook it off and thrust the pistol back into his pocket with a sigh. He nodded to the sergeant to continue watching the shed and set off wearily for the car.

'The devil take this fellow!' Maigret was thinking; 'making us hang about in this filthy weather when a man could be nice and snug in front of a roaring fire with a cup of hot coffee and a magazine.'

The policeman at the wheel of the car handed him a thermos. Maigret winced at the thought of what the coffee would be like by this; but he swallowed the tepid liquid and grunted his thanks.

For the umpteenth time that day he placed two photographs on his knees and began studying them. One was from a police dossier. It was dominated by the broad, self-satisfied smile of a man already showing signs of aging, with bold, insolent eyes and a rock-like jaw. 'His type always shoots first—and last,' Maigret reflected. This piece of information was not actually visible in the man's features: it was simply that the Inspector knew him rather well and had been following his activities for several years. A self-confessed fascist, member of an officer terrorist organization, former collaborationist, connected, so it was rumoured, with the Gestapo, and so on, he was well aware that he was destined for the guillotine, which was probably why he only became more audacious with each passing year.

The other photograph was taken from a police helicopter while police were pursuing the unknown who took refuge in this shed. It is not a good picture. It has been enlarged a good deal, so that the grain of the film is quite visible and the outlines are blurred and diffuse. But the hunted look of fear on the face half-turned towards the pursuers is unmistakable.

Maigret has to decide whether the two photographs depict the same man or not: for any further plan of action—and possibly the lives of a good many people—depends wholly and entirely upon this decision.

If both photographs show the same man, he will have to cordon off the shed with especial care because they could expect anything from this fellow: sniper fire, machine-gun bursts, grenades. He had hardly headed straight for this shed for nothing once he saw that escape was impossible: obviously, he had a whole arsenal in there.

But if the photographs show different people, life takes on a rosier hue. They will be able to come calmly to some agreement with the chap in the shed and persuade him to stop resisting and making things worse for himself. As a matter of fact, his shooting is pretty odd—as if he were only trying to frighten, not to kill.

So: 'the same or not the same?' This is the question that Inspector Maigret has been asking himself all morning, and still he cannot decide. Everything about the two photographs is different: size, angle of approach, clarity, facial expression—none are remotely similar. And yet it could still be one and the same man.

None of the experts that the Inspector has called in will give a straight answer—the photographs are too utterly different. True, he has heard that problems of this sort can be solved by computers; but as usual there has not been time to drop in on the cybernetics people, and anyway, he is getting too old to go back to the schoolroom now.

'And yet that is obviously what I should do,' thinks Maigret; and he decides to get in touch with the computer centre... At which point we shall bid the famous detective farewell.

We can also say good-bye to the shooting and pursuing: the thriller opening was merely a useful way of posing the problem more effectively—a necessary device in any popular book that sets out to popularize rather unpopular matters.

Let us consider a completely realistic example. We are talking about no less serious and responsible department of criminology than that of identifying the criminal.

At almost every stage of his work the detective is faced with the problem of establishing, on the basis of known facts about the criminal, whether a suspect is the criminal or not. For the sake of simplicity we shall examine the case where we have two photographs: one of the criminal and one of the suspect. The question we have to answer is: do the photographs show the same person or different people?

And so: the same or not the same?

We should not be in too much of a hurry to answer this question, for it is not as simple as it seems. Let us begin by analysing the implications of the situation.

The detective has to make a decision and choose between two alternatives: 'the same' and 'not the same', that is, 'suspect and criminal are the same man' and 'they are different people'. After due analysis of the photographs the detective has to come to a decision one way or the other. And it is desirable, of course, that this decision be in a certain sense the best possible one.

But what does 'the best possible one' mean? The process of investigation, like any real process, is subject to the effects of random interference which impedes our efforts to select the correct alternative. Here, interference takes the form of poor quality in the photographs we have to compare, distortions of the subject due to the optical systems of the cameras, different views of the faces, different facial expressions, and so on. Obviously, these distortions cannot be eliminated from the prints and constitute interference that has to be reckoned with, because it can lead to mistakes being made in the investigation. These mistakes will fall into two main classes.

Mistakes that result in the acquittal of the guilty we shall call type one mistakes. The two photographs are of the same person, but the level of interference is so high that the faces in them appear to the detective to belong to different people, and he makes a mistake. As a result of this mistake, the criminal goes free.

There is also another kind of mistake. The photographs are of different people, but they are so similar that the detective mistakenly concludes that they represent the same person. In this case the innocent man who is taken for the criminal will suffer. Mistakes of this kind we shall call type two mistakes.

Both kinds of mistakes are undesirable because both of them involve individuals, the law-courts, and, finally, society itself in certain losses. In the first case (acquittal of the guilty) the losses consist in the fact that the crime goes unpunished and the criminal, remaining at large, is free to commit fresh crimes. In the second case (punishment of the innocent) the criminal, once again, gets off scot-free, but, worse still, an innocent person is made to suffer. Obviously, this kind of mistake is the more serious because it involves society in graver losses. (This accords with the humanitarian dictum that it is better to let a guilty man go free than to condemn an innocent man.)

The detective is well aware of these facts, so he endeavours to make such a decision as will keep society's losses to a minimum should it prove wrong.

Suppose we assign the numerical value A to the losses associated with acquitting the guilty, and the value B to those associated with condemning the innocent. It is then obvious that a type one mistake results in losses A, and a type two mistake in losses $A+B$.

It is difficult to say what units these losses should be measured in; however, on closer analysis we realize that this is of no particular consequence and that it is sufficient to determine by how many times one set of losses exceeds the other, that is, to find the ratio $q=\frac{A+B}{B}$. In the simplest case it is sufficient to set $A=B$ ($q=2$), that is, to regard the total losses associated with condemning an innocent man as being twice as great as those associated with acquitting the guilty party. The absolute values of A and B are then of no further interest.

Suppose the detective makes use of a particular rule (or algorithm) for comparing the photographs—a rule that enables him to calculate the degree of non-correspondence of the two faces. Suppose we designate this quantity by a number Q. The greater Q is, the more different are the two faces in the photographs; conversely, the smaller Q the more they are alike. If there were no interference, the problem would be quickly solved: 'the same' if $Q=0$; 'not the same' if $Q>0$. But interference complicates the whole picture. It may lead to the result $Q=0$, when in fact the faces are different; and vice versa. So how is the detective going to make his decision?

The answer is that he has to devise a *decision rule*. Such a rule takes an extremely simple form: if the non-correspondence index Q is greater than a certain number τ, the faces are to be considered different and the suspect innocent; if Q is less than τ, the photographs show one and the same person.

But how is the value of τ to be determined? This is important because the success of an investigation depends, in many ways, on this number. Suppose τ is small (or actually equal to zero). Then in accordance with our decision rule we shall hardly ever condemn an innocent man. But if the suspect happens to be the criminal, we shall certainly let him go, and thus commit a type one error. So if τ is too small, we free the innocent and almost certainly free the guilty as well.

Suppose τ is large. The criminal, provided he is suspected, will not escape punishment. But if the suspect is innocent, our decision rule will compel us to convict him, and the criminal will remain at large: and the losses will be still greater $(A+B)$.

It is obvious from the above that τ must have some intermediate value if it is to minimize the losses that would result from a wrong decision.

All the same, how are we to determine this value? It is here that statistical decision theory comes to our aid.

We have to construct a *risk function*. This function takes the following simple form:

$$R = Ap_1 + (A + B)p_2$$

where A and B are, as before, the losses entailed in acquitting the guilty and condemning the innocent, respectively; p_1 is the probability of acquitting the guilty, that is, the degree of certainty that a type one mistake will occur; p_2 is the probability of condemning the innocent, that is, the degree of certainty that a type two mistake will occur.

The risk function thus provides a measure of the average losses that would result from a wrong decision.

The values of the probabilities p_1 and p_2 depend on the value of τ, as is shown in Figure 36. Clearly, for $\tau=0$ a type one mistake is certain to occur $(p_1=1)$, because the criminal will be acquitted; whereas if τ is too large, a type two mistake is bound to occur, because an innocent man will be condemned $(p_2=1)$.

If, now, we substitute in the risk formula expressions for p_1 and p_2 in terms of τ, we obtain an expres-

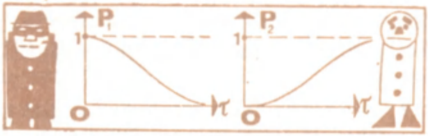


Figure 36

sion showing how the risk varies with τ. This is graphed in Figure 37.

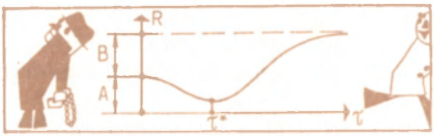


Figure 37

The graph shows that the risk has a well-defined minimum at τ^*. This is, therefore, the optimum value of τ. Consequently, we shall optimize the investigation if we take $\tau=\tau^*$ because the risk associated with the investigation will then be a minimum. If the detective does this he can be certain that the resulting losses in the event of a mistake will, on the average, be the smallest possible.

It is worth noting that the absolute value of the risk does not interest us here. The important thing is that it should be a minimum. The precise value of this minimum is immaterial to the determination of τ^*, as we have seen. This fact simplifies the problem considerably because it relieves us of the necessity of assigning precise values to the quantities *A* and *B*—which, alas!, we are at present not only unable to calculate, but about which we cannot even advance any rational considerations.

For example, suppose $A = B$, that is, the losses are the same. The risk formula then becomes:

$$R = A(p_1 + 2p_2)$$

which shows quite clearly that the position of the risk minimum is independent of A. (The value of A will determine the magnitude of the risk, but not the position of its minimum.)

And so, minimization of risk by appropriate choice of the parameter τ enables the detective to overcome the effects of chance interference inherent in the process of identifying a criminal.

Now let us tackle something that we are more familiar with: soft drink machines—and the uncertainties that surround their operation.

THE TRUTH ABOUT SOFT DRINK MACHINES

The simplest example of a mechanical device that makes decisions is the ordinary coin-in-the-slot machine that sells soda water, lemonade, and the like. When a coin is inserted in the slot, the machine has to decide whether in fact it is a coin of the realm or not. (To be, or not to be.) It has two courses of action open to it:

Course No. 1: 'to be'—to accept the coin as satisfactory and give its owner a drink.

Course No. 2: 'not to be'—to reject the coin as unsatisfactory and return it to its owner.

In order to make this decision the machine has to perform an experiment with a view to determining whether the coin is satisfactory. Let us assume that experiment consists of measuring the coin's diameter. The machine possesses two limit gauges for this purpose, an upper and a lower one. The coin should pass freely through the former, but not through the latter: only then will it be accepted as satisfactory.

The upper limit gauge tests whether the coin is larger than the specified standard. If it is, then it simply will not go into the machine, in which case the machine rejects it as unsatisfactory and refuses to serve the customer.

The lower limit gauge sorts the coins that reach it into two classes. The first class embraces coins that are larger than the gauge: these are stopped by the gauge and are thus acknowledged as satisfactory. The other class embraces coins that are smaller than the gauge: these pass through the gauge and are returned to their owners as being unsatisfactory.

The designer of the machine has to decide the dimensions of these gauges. This is easy enough for the larger gauge: it has to be equal to the diameter (d) of a new coin. Why a new one? Well, no coin gets bigger with the passage of time; therefore all good coins will have a diameter not larger than the initial diameter of a new coin.

It is far more difficult to establish a dimension for the smaller gauge. If it is too close to that of the larger gauge, coins that are old and worn, but otherwise worthy, will be rejected by the machine as unsatisfactory. On the other hand, if the gauge is too small the machine will accept counterfeit coins and other substitutes along with the real article. In either case the machine will suffer losses: in the first—losses of patronage and prestige; in the second—direct loss of income because the drinks are not being paid for—more precisely, are being paid for with dud coins, washers, and so on.

Obviously, there is an optimum, in a certain sense a best, dimension that should be used for the smaller, lower limit, gauge. This dimension will be such as minimizes the average losses due to errors of the first and second types: it has to minimize the risk. A type one error in this case consists in the rejection of a good coin: a coin happens to be rather worn, say, so the machine rejects it, even though in other respects it is a satisfactory coin (compare: condemning the innocent). A type two error occurs when the machine accepts a bad coin or a non-coin (compare: acquitting the guilty).

Let d be the standard diameter of a coin of the appropriate denomination, and $d - \tau$ the dimension of the lower limit gauge. The probabilities of the two types of error will then depend on the value of τ in the same manner as is shown in the graphs of Figure 36. If $\tau=0$, both gauges are exactly the same size and the machine will not accept anything: type two errors (acceptance of bad coins) will not occur at all; type one errors (rejection of good coins) are bound to occur. If τ is made sufficiently large, type one errors will almost certainly be eliminated and the machine will accept good coins; but at the same time it will also accept bad coins, so that the probability of type two errors will be increased.

In order to determine the optimum value for τ the designer must construct a risk function and choose τ so as to minimize the risk. For this purpose he needs to introduce two quantities, q_1 and q_2, corresponding to the losses associated with type one and type two errors, respectively. The risk formula for the machine then takes the form:

$$R = q_1 p_1 + q_2 p_2$$

For the sake of simplicity let us suppose that $q_1 = q_2 = q$, that is, both types of errors result in the same losses. The formula then becomes:

$$R = q(p_1 + p_2)$$

The graph of Figure 38 shows how the risk varies with τ. Once again, it is obvious that the value τ^* mini-

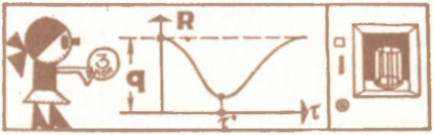


Figure 38

mizes the risk. This value is therefore the one that the designer must specify in his design. Only then will the risk of a wrong decision by the machine be minimal, and only then will the machine make optimum decisions.

It is interesting to note that τ^* is independent of the magnitude of the losses q. Different values of q will result in different risks, but the location of the minimum will remain unchanged. Consequently, the designer can use the following formula for risk:

$$R = p_1 + p_2$$

which considerably simplifies his task.

In concluding our discussion of decisions and risk we shall only remark that the idea of introducing risk has proved extremely fruitful not only in criminology, but also in physics, biology, economics, and other sciences. Whenever we are looking for an optimum decision in circumstances involving chance, we have to evaluate the risk associated with the decision and endeavour to keep it to a minimum. This will ensure that our decisions are as sound as we can make them despite the presence of chance interference; in other words, it enables us to overcome chance and to lessen its destructive consequences.

We live in a world of chance, a world in which nothing can be stated with one hundred percent certainty. Every judgement must begin with the words 'in all probability' because any categorical statement runs the risk of proving false. The noisy background created by chance interference produces conditions in which errors are difficult to avoid.

In the preceding chapters we have been examining the ways and means at our disposal in the struggle against chance interference. This struggle, like any other, involves certain sacrifices and losses, in particular—that most precious and irrecoverable thing—time. As we have seen, the best way of overcoming chance interferences is to use a cumulative method; and any cumulative method requires the passage of time, which is therefore lost.

Let us examine the following extremely common situation. Suppose we are faced with a number of alternatives and we have to make a decision, such as where to spend our holidays, for example: Odessa, Yalta or Sochi. Before we make up our minds, we ought to gather as much information as we can about how these places rate as holiday resorts. As a rule, this information will be subject to interference of every kind. If, for example, you go round asking various acquaintances about the life and comforts to be found in these three places, you are liable to hear some highly contradictory opinions. One would have got off to a bad start for some reason—and then met a girl on the beach, fallen in love, and finished up enjoying himself immensely. Another would have started his holiday in splendid form with a beautiful beachfront flat—and then had such a bad quarrel with his wife that she wanted to pack her bags and leave before the holiday was half over. Naturally, the first will have nothing but the highest praise for wherever he spent his holiday; the second will be left with a poor impression of what may have been an excellent resort.

If you want to make the right decision, you will have to filter out all the extraneous matter in the information you receive. One way of doing this is as follows. Set aside three pages in your notebook for recording information about the three places you have in mind. Mark any favourable opinions with a plus, and any unfavourable ones with a minus, making sure, of course, that all these opinions are from different people—and from people whose judgement you respect. Then, before you rush off to buy your tickets, have a look at your notebook and tabulate the results of your enquiries. You will possibly finish up with something looking like this:

Place	Odessa	Yalta	Sochi
Plus	8	5	16
Minus	4	2	7
Total	12	7	23

What next? Probably the first thing to do is to agree upon a decision rule. Here, the natural choice for such a rule is: to take the place that has the highest percentage of plusses. After a little arithmetic the table gives us:

Odessa ... 66%
Yalta ... 71%
Sochi ... 69%

So Yalta has the highest percentage. Does this mean that you cannot go wrong if you opt for Yalta?

No, of course not.

The point is that the percentages we have just obtained are approximations that we must not attach equal weight to. Indeed, the significance of any averaged result depends on the number of individual items that the average represents: the more items there are, the more accurate is the result. Therefore, the most accurate evaluation of merit is the one for Sochi (twenty-three opinions), and the least accurate the one for Yalta (seven opinions). Consequently, you could easily be wrong if you decided on Yalta, because it is quite possible that with sufficient additional information to bring the total number of Yalta opinions to twenty-three the proportion of plusses for Yalta might drop, say, to 67%. And this would alter your choice in favour of Sochi, with 69%.

What are you going to do, then? Strictly speaking, if you want absolutely reliable results, you will have to collect so much information and spend so much time in the process that the whole exercise becomes pointless—unless you are prepared to postpone your holiday until the following year. This is precisely the reason why, whenever we have to make a decision, we limit ourselves to a reasonable period of time, being always aware that the decision may prove wrong.

But can we, perhaps, determine the optimum amount of information we should collect in order to solve a given problem? As it turns out, we can. First we have to work out our losses; and they will be of two kinds: those incurred in the process of collecting information (by using a cumulative method, for example) and those associated with making a wrong decision.

For the sake of simplicity let us suppose that the collection losses W are proportional to the time T spent on collecting information (compare the saying: 'Time is money'). These losses are shown in Figure 39 as an oblique straight line, indicating that the

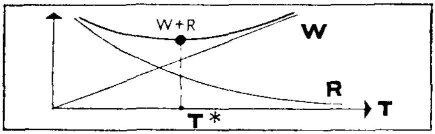


Figure 39

losses involved in gathering information depend directly on time spent.

Now let us consider the losses associated with a wrong decision. The likelihood of making a wrong decision depends inversely on the amount of information we have at our disposal. The less the information the greater the probability of error and the greater the risk (remembering that risk is the average loss resulting from error). The risk graph R in Figure 39 is a decreasing function. It shows that the probability of error and, correspondingly, the risk both decrease as the amount of information increases.

The total losses involved in solving the problem will be equal to the sum $R+W$ of these two kinds of losses. The graph of this sum shows a well-defined minimum at the point T^*. This is therefore the optimum time we should spend on collecting information.

If we act in accordance with these principles, we guarantee ourselves minimum total losses on the average even though we may occasionally make mistakes. These mistakes, however, will cause us less trouble in the long run than would the collecting of the large amount of information needed to avoid them.

This, then, forms the theoretical justification for our right to err. It is perfectly permissible to make mistakes; however, we must always attempt a sober evaluation of the losses we incur in making them.

We can go further than this and assert that mistakes are necessary. If a person—or a decision-making device of any sort—never makes the slightest mistake we can be quite certain that he or she—or it—is not working at optimum efficiency. Faultless operation can only mean one of two things: either the rate of work is extremely slow because an excessive amount of time is being spent on filtering and purifying the information; or there is an unrealistically large amount of redundancy being employed to guarantee reliability—for example, by solving the problem by a number of methods simultaneously and then taking a vote. In both cases the losses involved are enormous, and quite unjustifiably so.

At the same time, we should avoid rushing to the opposite extreme: that of forcing ourselves to make mistakes. We should bear in mind the concept of risk, remembering that it is based on two factors: the cost of the mistake and the probability of making it. If the mistake costs us but little, the risk is small and we can tolerate a relatively large number of such mistakes. On the other hand, if the mistake is an expensive one (one likely to lead to an accident, for example), we should try to make its probability of occurrence as small as we can by improving the reliability of the system.

It is fitting, I think, to conclude this chapter with a quotation from the well-known cyberneticist R. Ashby. In one of his papers he declared:

'It is much cheaper and much easier to construct a cybernetic machine that does not have one hundred percent accuracy, but that has an accuracy of up to, say, ninety percent; and then, when using this machine, to evaluate its possible operating error on the basis of probability theory. The gains we achieve by doing this—in terms of the cost and the simplicity of construction of the machine—are very large indeed.

'Very often people who attempt to build machines that are one hundred percent accurate devote an incredible amount of effort to the task. This effort never pays for itself. It is far simpler to have a machine that is less accurate, but at the same time much easier to construct—and is also just as usable.'

PART II

Welcome Chance

> Never disregard any special or remarkable incident: often it will be a false alarm; but occasionally it will conceal an important truth
>
> *Flemming*

1. SHERLOCK HOLMES SPEAKS HIS MIND AT LAST

'Ah, my dear Watson!' Holmes breathed, stretching his legs inside the warm plaid that enclosed them and sending a smoke ring to the ceiling. 'It is such a pleasant evening that I feel like being myself for a change.'

He gazed thoughtfully at the glowing embers in the smouldering hearth, the shooting flames of the dying fire casting a flickering light across his cold, aquiline features. Whether it was the wine he had had or the sumptuous dinner prepared by the doting hand of Miss N., or whether it was the enchantment of the dying hearth—whatever the cause, Holmes's face had lost its customary frown of concentration and become soft and kindly. Doctor Watson had never seen his friend in such a condition. It was so unlike the famous detective's decisive, resolute nature. Holmes looked as though he would break into a cheery smile at any moment and start joking about his nephew's successes.

Suddenly Holmes chuckled to himself and said:

'You cannot imagine, my dear Doctor, the extent to which the work of a detective depends on good fortune. I myself arrived at a proper understanding of this fact only after I had read the latest works on cybernetics. I was particularly impressed by a piece from the pen of that eminent cyberneticist, Ashby. The fellow has gone so far as to construct a machine he calls the homeostat: seeks out its target in an entirely fortuitous manner. There's food for thought for you, eh, Watson?'

'Holmes, I am always at a loss to follow the train of your thoughts. What possible connection can there be between cybernetics and the work of a detective? And what can you mean by uttering those tender sentiments about Dame Fortune, when the truth is that chance represents nothing but a hindrance to the detective on every conceivable occasion. Surely a detective's conclusions must be the result of a process of logical reasoning, rather than an indulgence of the idle whims of chance?'

'That is perfectly correct, Watson,—but a little too orthodox. It is true that one has to be logical in one's conclusions; the question is: how does one arrive at those conclusions? Logic merely permits one to verify the truth of a conclusion; it does not enable one to reach a conclusion. If you will remember, a certain philosopher has said that logic does not teach us to think logically, any more than knowledge of the digestive processes improves one's digestion.'

'You amaze me, Holmes. Have I not heard you speak most earnestly, time and time again, of the need for logic in the detective's work? Am I to understand that your opinion in this regard has changed?' Watson concluded in some confusion.

'Not changed: deepened,' Holmes replied thoughtfully, sending another smoke ring to the ceiling. 'All our attempts to detect crime by the application of analytical methods are only of value as rules of thumb for Scotland Yard's police-school novices.'

'Really, Holmes!' the agitated Doctor expostulated. 'And does that remark apply also to your famous deductive method? Is it not true that this method, coupled with logical reasoning, has enabled you to solve the most baffling mysteries with ease?'

'Alas!' Holmes observed sadly. 'The deductive method is a very powerful instrument; but before one can use that instrument one has to have at one's disposal an immense quantity of initial information, the kind of information that no detective possesses while he is actually conducting his investigations. In this regard he is forced to work in circumstances of the cruellest privation. Of what possible use is the deductive method here? To be quite frank'—his voice sank to a whisper—'I do not use the deductive method at all.'

Watson was dumbfounded. Opening his mouth with an effort, he gasped: 'But what about all those stories of yours—the ones that I wrote up and had published under that pseudonym—you remember—"Conan Doyle"? In those you gave splendid, most convincing descriptions of the process of solving a crime; and they most certainly did include the deductive method.'

'That is really the whole point,' Holmes observed wearily: 'they were *descriptions*. It is easy enough to describe a crime; but to solve it, aha! that is much more difficult. There, the so-called deductive method is virtually useless. One has to resort, as a rule, to an inductive method—provided, of course, that the crime is not a trivial one. Just between ourselves, to confuse deduction with induction is an unforgivable error even in the greenest freshman. At the same time, I confess that I once sincerely believed that I acted according to deductive principles; but when I examined the situation more closely, I found the exact opposite to be the case. Deduction and induction are, after all, directly opposed to each other. Deduction is a process of reasoning leading from the general to the particular; induction is quite the reverse—from the particular to the general. So you see: I was preaching one thing and practising quite another. This I understood only after the appearance of works on the heuristic method of problem solving, a method, mark you, that is very similar to the inductive method.'

'What on earth are you talking about!' Watson gasped in still greater alarm. 'How could a man of your intellect confound such elementary matters?'

'The whole point is this: I described the process of solving a crime *after* it had been solved, rather than *while* it was being solved.'

'What difference should that make?'

'All the difference in the world. Once the crime is solved, everything seems so simple and natural. When one is describing the affair all one's thoughts tend to be one-sided: the pieces all fit into place pointing infallibly to the already known solution. The situation can be depicted, if you like, in the form of a chain of deductions something like this.' And Holmes proceeded to draw a diagram (Fig. 40).

'In actual fact,' Holmes went on, 'during the investigation—until the crime is solved—something quite different takes place. The solution is not at all obvious, and one has no idea of how to set about

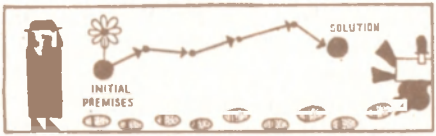


Figure 40

finding it. The real chain of deductions is highly reminiscent of the behaviour of a blind puppy setting out to look for a dish of milk.'

Another sketch appeared on the sheet of paper (Fig. 41).

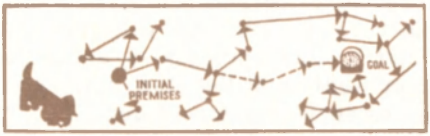


Figure 41

'Here you see a mass of wrong guesses that subsequently are not substantiated and are therefore discarded; the actual path taken to reach the goal is an extraordinarily confused tangle. Chance plays a decisive role here. Once the target has, by chance, been found, anybody can trace the shortest logical path (I have indicated it by a broken line); but as you can see, it was not the path taken to reach the goal.'

'Could one say, then, that a description of a process of investigation is always deductive, but the process itself requires an inductive method?' Watson asked timidly.

'Perfectly correct. My dear Doctor, you always catch my meaning splendidly—'

'—and put my foot in it,' Watson concluded: 'it is just as well I had the forethought to publish your notes under a pseudonym.'

'Pray do not upset yourself, my dear Watson,' said Holmes with a smile. 'Your reputation will remain unsullied if you will but extend the list of errata to include a remark to the effect that "deduction" is to be read as "induction" throughout; and everything will turn out splendidly.'

Watson was deeply touched. 'I have always known that for you, my dear Holmes, there is no such thing as an insoluble problem.'

I devised this conversation in order to show that the creative process of searching for truth cannot be described strictly in terms of a series of logical deductions. Every such process is accompanied by a chance factor that supplies the variety and 'divine inspiration' essential to the search.

We shall now examine various ways of utilizing chance, that is, various control methods that make use of an element of chance. One of these is the *Monte Carlo method.*

2. THE MONTE CARLO METHOD

Monte Carlo? We usually associate this name with the gambling houses of Monaco, the tiny principality hidden away somewhere in the South of France and consisting entirely of the city of Monte Carlo itself.

How is it, then, that the name 'Monte Carlo' has recently started to appear in the pages of serious technical and mathematical journals?

Let us have a closer look at a roulette wheel. The wheel takes the form of a shallow circular dish with a raised edge, and its inner part is provided with one hundred shallow holes. A light-weight ball is released into the dish at high speed. The ball bounces off the raised edge of the dish repeatedly, until at last it falls into one of the holes. Is it possible to predict precisely the exact hole into which the ball will fall?

Of course, it is possible. If we determine the exact initial direction of the ball's motion, taking into account the slightest tremble of the thrower's hand; if we calculate the exact direction of rebound for every collision of the ball with the edge of the dish; if .... In a word, if we know all the conditions governing the ball's motion exactly—that is, if we know the motions of all its molecules—we can predict where it will come to rest.

It is perfectly clear, however, that we shall never succeed in determining all these factors exactly. One fundamental reason is contained in the uncertainty principle that we discussed some time ago, according to which it is impossible to measure the exact motions of the ball's molecules. Moreover, the factors involved are exceedingly numerous and change so rapidly that we would be quite unable to keep track of them; as a result, the ball would still fall with equal probability into any one of the holes even if the uncertainty principle could be disregarded.

In nature and in technology and in ordinary life there are a great many processes that can only be described in terms of probability: a mountain rock-fall, for example; the flight of a bird hunting for swarms of midges; the number of people travelling in trains, trams, and aeroplanes; the number of bankruptcies during a financial crisis in the capitalist world; the numbers of young and adult fish in a given body of water; the number of children that will be born in a five- or a ten-year period. There are millions more examples like these. Each of them contains an element of uncertainty and a question to which there is no *exact* answer. But there are many questions of this kind that need to be answered. For example: what should be our annual output of aircraft, railway rolling stock, shipping, and tramcars? How many factories should we build in the next few years and how large should they be to meet the needs of the population? And what will those needs be?

To answer questions like these, we use methods based on probability theory. These methods do not give us precise answers; instead, they enable us to determine with sufficient accuracy the limits within which the quantity that we seek may be expected to vary, or the probability that a particular event will occur. One of these methods is called the Monte Carlo method.

In order to get an idea of the essence of this method, let us examine a simple, but instructive, example.

AN EXPERIMENT AND ITS MATHEMATICAL MODEL

Suppose we wanted to find the area of a circle of radius equal to one centimetre. We could work it out using the familiar formula πr^2, giving $3.14 \times 1^2 = 3.14$, that is, the number π.

But have you ever wondered how you can determine a value for π, even if only an approximate one? You can do this rather easily, in fact, by means of the Monte Carlo method.

We take a grain of sand and throw it a large number of times on to a sheet of paper marked with a series of circles, all of a radius equal to one centimetre, as in Figure 42. The grain will fall either inside the circles or in the spaces between them (C_1 is an example of a point inside a circle, C_2 a point outside the circles). The greater the total area taken up by the circles, the more often will the grain fall inside them.

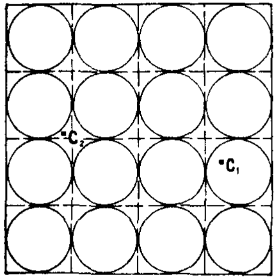


Figure 42

Suppose we take a square of area equal to 100 cm^2 and inscribe it with circles of the above-mentioned radius (there will be twenty-five of them); and suppose that in 1000 throws, the grain of sand falls 700 times within the circles and 300 times in the empty spaces. It is then natural to assume that the area occupied by the circles is given by the ratio of the number of falls (n) within the circles to the total number of throws (N), that is: $n/N = \frac{700}{1000} = \frac{7}{10}$ of the total area of the square; that is, 70 cm^2.

Dividing this result by 25, we obtain the area of one circle, and from this we can easily calculate a value for π. The value we obtain will become more accurate as we throw the grain of sand a greater number of times. And we observe that an experiment of this nature takes up a great deal of time.

Suppose we try to speed up the experiment by constructing a mathematical model of it. Figure 43 shows a part of the field on to which we threw the grain of sand. In view of the symmetry of the whole field, we need only consider a part of it containing a single quadrant of one of the circles. You can easily verify that the whole field consists of one hundred of these elements. To simulate random throwing of the sand, we choose two random numbers, A and B, lying between, but not equal to, nought and one, thus

$$0 < A < 1; \; 0 < B < 1$$

Starting at the origin O (Fig. 43), we set off a length corresponding to the number A along the ordinate,

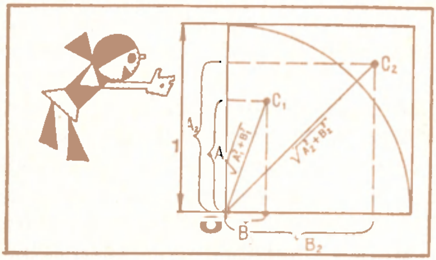


Figure 43

and a length corresponding to the number B along the abscissa, and then obtain a point C at the intersection of the perpendiculars. Thus the numbers A_1 and B_1 determine a point C_1. the numbers A_2 and B_2 a point C_2. The points C_1 and C_2 represent different points of impact of the grain of sand in the actual experiment. If $A^2+B^2 \leqslant 1$ (equal to or less than unity), the grain falls within the circle; if $A^2+B^2>1$ (greater than unity), it falls outside. Therefore, in order to determine whether any point C falls within a circle, we only have to check whether the inequality $A^2+B^2 \leqslant 1$ is satisfied.

Now we are in a position to define relations between the essential features of the experiment and those of the model.

The actual experiment	Its mathematical model
1. Throwing a grain of sand on to the field (Fig. 42)	1. Choosing two random numbers A and B greater than 0 and less than 1
2. Checking whether the grain falls within one of the circles or on its circumference	2. Checking whether the inequality $A^2+B^2 \leqslant 1$ is satisfied

The mathematical model is ready for use.

Now, instead of performing the actual experiment (throwing sand on a marked field), we can calculate the results of the experiment directly. All we need is a table of random numbers, a pencil and a piece of paper.

Therefore, it is possible to replace a physical experiment by a mathematical one, that is, by a calculation.

Let us compare the advantages of these two methods of evaluating π in regard to two particular factors: chance interference, and the time spent for the same number of 'throws' in either case.

The physical experiment, like any other real process, is highly susceptible to chance interference. Interference makes its presence felt, for example, in inaccuracies in determining where the sand grain falls, in the unevenness of the field surface (the sand tends to 'prefer' depressions in the field), and in inaccuracies in drawing up the field itself (non-ideal circles, and so on).

All these inaccuracies are absent from the mathematical model. To be sure, there are, strictly speaking, interference factors at work here as well (errors in calculation and rounding off, for example); but it is possible to reduce these to insignificant proportions, so they may be ignored.

Consequently, from the point of view of error the mathematical experiment is preferable to the physical one.

Now a word about time. Here the palm of victory should go to the physical experiment. For indeed the calculation and summing of two squares on paper will always take longer than simply throwing a grain of sand. Here the most ham-fisted experimenter will always beat the nimblest mathematician.

But what about our high-speed computers that can do sums like this in a matter of seconds? Could we use one here? Suppose we try.

A modern, high-speed, general-purpose computer can easily deal with a problem like this in a very short time indeed, but... But it has to be programmed to do it, and somebody has to write the programme. Just as any human calculator has to be told at some stage what numbers to insert in his calculation and how to go about it, so a computer has to 'know' what to do and how to do it; otherwise it cannot do its job. This is the function of a computer programme.

Suppose we take a look at how a computer programme is written. First, we draw up a programme of the experiment itself. This is shown in Figure 44. Each

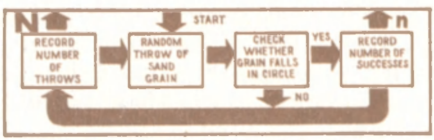


Figure 44

rectangle in the diagram represents an essential operation in the experiment. Arrows emerging from each rectangle point to the operation that is to be performed next after the one represented by the given rectangle. Where a rectangle has two arrows emerging from it, the conditions under which the process follows either arrow must be indicated.

The block diagram of the experiment contains two elementary operations: a record of the total number of times the sand is thrown, and a record of the number of successful throws. These rather obvious procedures are included because the programme should correspond to the actual experiment exactly. And it does: when we throw the sand grain we count the number of times it lands inside a circle and the total number of throws. If this detail is omitted, the whole experiment becomes pointless because the numbers N and n constitute the information it is designed to provide us with.

The computer programme is drawn up in an exactly analogous manner. The form it takes is shown in Figure 45. Here the counters for N and n correspond to the recording operations of the physical experiment.

If we examine both diagrams carefully, we see that they are very similar generally, and have identical structures. Both have the same number of rectangles and the same number of arrows pointing in the same

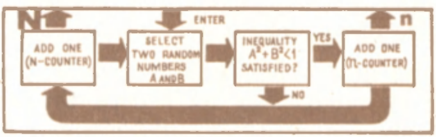


Figure 45

directions and so indicating the same set of relationships between the various operations. This is hardly surprising since both diagrams describe essentially one and the same information process aimed at determining the numbers N and n. Therefore we would expect the same result in either case: whether we throw sand grains, noting where they land (working according to the programme for the physical experiment); or whether we do the same experiment mathematically by means of a series of calculations (working according to the programme for the mathematical model).

And so we have two methods at our disposal for obtaining the required result (the number n): the physical experiment, on the one hand; and a computer calculation based on a mathematical model of the experiment, on the other. Suppose, now, we draw up a table showing all the pros and cons for both methods.

	The physical experiment to determine n	The computer calculation for n
For	1. Simplicity (no computer needed)	1. Rapidity
	2. Demonstrability	2. Large N
Against	1. Large time losses	1. Need to write a programme
	2. Small N	2. Need to have a computer

This shows quite clearly that provided we are not frightened by computers and have the opportunity to use one—and practically anybody can have access to one nowadays—the calculation is undoubtedly superior to the experiment. All the same, we must recognize clearly that a mathematical model of the experiment is an essential pre-requisite to the calculation. Only if we have such a model is it possible to run the experiment on a computer and take advantage of its tremendous speed by completing a large number of trials.

And it is this simulation of physical experiments on computers that is called the Monte Carlo method.

The chance element is absolutely essential to this method because it alone reflects the randomness and uncertainty that are inherent in the physical experiment we wish to simulate. In this connection the Monte Carlo method is often known as the *statistical* (or *probability*) *simulation method.*

Chance in the computer is specially generated by devices known as random number generators and reflects the random processes that take place in the actual experiment. These devices enable us to replace an extremely lengthy and tedious physical experiment with a convenient mathematical model by means of which we can repeat the 'experiment' hundreds of times in the computer.

The randomness simulated by the machine is an example of necessary and useful chance. It reflects the role of probability in real experiments; and this may be useful, or it may be harmful and appear as interference. In the above experiment enabling us to determine the number π by means of random throws, chance is useful. In most other cases chance rears its ugly head as a source of interference—but one that can also be appropriately simulated by the Monte Carlo method.

MONTE CARLO AND THE BALLISTIC MISSILE

We shall now see how the Monte Carlo method can be applied to calculating the point of impact of a ballistic missile.

The trajectory and the point of impact of a missile could both be computed with perfect accuracy if all the parameters governing its flight were known exactly. We would need to know the exact all-up weight of the missile and its fuel, the exact wind strength and direction in various layers of the atmosphere along the entire flight path, the exact temperature, pressure and density variations of the air at all points through which the missile would pass, and much, much more besides.

But in practice it is absolutely and utterly impossible to determine accurate values for all these parameters. They change—and they change rapidly. Painstaking study and observation enable us to do no more than to establish the limits between which these parameters vary, and to determine their statistical properties. Where do we go from here?

To determine the accuracy of the missile's landfall we can adopt a procedure similar to the one that we used in the experiment with the grain of sand. We can devise a programme for calculating the trajectory of the missile. This programme will contain parameters the values of which are unknown to us. What we do is to select random values within the appropriate limits for each of them, and then perform a calculation on the basis of these values to determine where the missile would land. Then we perform a second calculation using a second set of random values chosen again within the limits for each of the unknown parameters; then a third calculation using a third set of values; and so on. When we have completed a whole series of such calculations, we obtain a set of landfall points for the missile. These are random points since each is obtained as the result of a random calculation; but a sufficiently large number of them will mark out an area known as a *scatter ellipse*.

This ellipse contains extremely valuable information concerning the quality and the effectiveness of the missile. It enables us to pin-point the area in which the missile is most likely to land, and to gauge the accuracy with which we can expect to be able to aim the missile, and so on.

And so we see how with the aid of the Monte Carlo method we can obtain exceedingly valuable information regarding a missile's point of impact without the need for a series of extremely expensive test flights of the missile. We are therefore able to effect huge savings in terms of both time and material resources.

The Monte Carlo method finds a highly original application in the solution of a number of problems of mathematical physics to do with thermal conduction. We can illustrate this with the following simple example.

A DRUNK SOLVES A PROBLEM

A drunk? Certainly! And not just a little merry, but so blind drunk that when he finds himself at an intersection it is a matter of utter indifference to him which road he should take. And it is in just this state of alcoholic stupefaction that the drunk is best able to assist us in solving one of the most complex problems of mathematical physics—the problem of thermal conduction in a continuous medium. Surprised?

Don't be too hasty. Let us examine the problem of heating a plane disc, or, as the physicists say: the two-dimensional problem of thermal conduction. This is a problem that anybody who has dealings with the casings of a variety of heating appliances (such as hair-dryers, ballistic missile warheads, smelting furnaces and so on) may be called upon to solve. With any heating appliance it is extremely important to be able to determine the behaviour of the thermal stresses arising within the appliance: disaster may follow if these stresses become too great.

For the sake of simplicity let us consider a rectangular disc whose edges are maintained at a definite, known temperature. The problem is to calculate the temperature at any point on the disc chosen at random.

The reader will probably be a little bewildered at this juncture because there is evidently not the slightest element of chance involved in the problem—yet we have been saying all along that the Monte Carlo method is only applicable to experiments in which chance is a key factor. Such bewilderment, if such there was, is quite uncalled for. The laws of thermal conduction are determined by thermodynamics; and thermodynamics is bound up in the closest possible way with statistical (random) processes.

It is now known that heat is not conducted continuously, but in separate small amounts, or quanta, as they are called. We can regard the motion of heat quanta as being chaotic; that is to say, the quanta move in random directions. If a large number of them gather at a particular point, that point will be heated to a greater extent than one where fewer quanta have gathered.

To find the temperature at a particular point on the disc we have to determine how often energy quanta arrive at the point from various parts of the disc. Each quantum traverses a random path across the disc, beginning at one edge and ending at another. The particular route taken by a quantum is entirely a matter of chance. Obviously enough, each point on the disc must have several of these random paths, originating at different edges of the disc, passing through it; and the temperature at each point will be equal to the average value of the temperatures brought to it. For example, if the point forms the intersection of quantum paths originating chiefly at an edge having a temperature of 50 °C, the temperature at the point will be close to fifty degrees. Again, if the intersecting quantum paths originate at edges having different temperatures, some being at 50 °C, some at 20 °C, and there are twice as many of the former as of the latter, the temperature at the point in question can be determined, as you might have guessed, by a simple calculation:

$$t^{\circ} = \frac{2 \times 50 + 20}{3} = 40 ^{\circ}\mathrm{C}$$

This phenomenon, however, also permits of an 'alcoholic' interpretation (such as we began this section with) as follows.

Let us imagine for the moment that there is a city populated entirely by alcoholics (the author tenders his apologies in advance for so freely invoking this fine breed of men). And let us suppose that this city is built on a rectangular plan similar to the shape of our rectangular disc, and that all its wine shops are disposed around the perimeter and are specialized to the extent that each sells wines or spirits of a particular potency only and is designated by a number corresponding to the proof strength of its merchandise. So wine shop No. 40 deals only in sherry, No. 60 only in rum, No. 12 only in dry Georgian wines, and so on. Let us suppose, further, that the numbers of the wine shops correspond to the temperatures at the edges of our heated rectangular disc: so that, point for point, degrees of temperature at the edges of the disc, degrees of proof strength of the beverages in the wine shops and the numbers assigned to the wine shops around the borders of our imaginary city all coincide.

The inhabitants of this merry city lurch into a shop that happens to be handy, equip themselves with a bottle, and thereupon commence their drunken, and of course random, wandering about the streets. Whenever they meet up with other fellow citizens, each similarly equipped with the inevitable bottle, they mix themselves a cocktail by combining the contents of the various bottles in equal proportions. The potency of the cocktail will contain information as to where the bottles have come from and what the 'temperature' is at the point where the cocktail is enjoyed. In other words, to determine the temperature at any point on the disc we have only to taste a cocktail mixed at the corresponding point in our city of alcoholics.

However, a little observation of the erratic routes taken by our tipsy friends is enough to convince us that our desire to sample and grade the cocktails at any point will be thwarted by at least two circumstances.

The first is that, having picked out the particular intersection where we propose to do the sampling, we shall sometimes have to wait for very long periods between successive alcoholics for the simple reason that in their aimless shambling about the city, they do not arrive particularly often at the point where we are waiting for them. Moreover, most of them do not pass through the point where we are waiting at all. Consequently, any attempt to determine the temperature at the point by this method involves us in a mass of unnecessary losses.

The second difficulty is that we shall obviously have to wait still longer for several citizens to meet at the intersection simultaneously and start mixing drinks. (After all, this is the event we are really waiting for.) However, we can get round this one by collecting a tribute from each passer-by, instead of waiting for two or more to arrive at once, and placing them all in a single bottle. A swig from this bottle will then enable us to determine with ease the potency of the resulting mixture and the temperature at the corresponding point of the heated disc.

But how can we arrange matters so that the revellers will appear at our intersection more often? We cannot call out to them or beckon to them without upsetting the randomness of their wanderings. What can we do, then?

The answer lies in making use of a cunning device based once again on taking advantage of chance. If you look at a random path that comes suddenly to an end at the edges of the disc, you will find that you cannot tell where it begins and where it ends. This is because the direction of motion along a random path is irrelevant. Consequently, we need only examine the paths that depart from the point in question, and follow them to wherever they end on one of the edges of the disc. Then we simply 'reverse' the motion and treat the whole situation as if it were the other way round: that is, as if these paths had brought to the point in question the temperatures of the edges where they terminate (or 'from which they had come'—in accordance with our reversed view of the situation).

We can apply this method to our merry city as follows. We select any completely sozzled citizen at random and tag him; then we release him at the particular intersection that interests us. At the same time, we ask all the wine-shop proprietors to telephone us if the tagged man ends his wanderings in their particular shop. We do the same thing with several other devotees of the bottle. All we have to do then is to man the telephone and note the shop-numbers as the proprietors telephone in. For example, suppose we get a set of calls from shops Nos. 40, 40, 60 and 20: the strength of the cocktail at the point under investigation will then be

$$\frac{1}{4}(40 + 40 + 60 + 20) = 40° \text{ proof}$$

And in accordance with our analogy, this will be the temperature of the corresponding point on the heated disc (40°C).

It follows that the problem of determining the temperature at a given point can be solved very simply by tracing a number of random paths from the point in question and reading off the temperatures at the points where they meet the edges of the disc, remembering that these temperatures at the edges are included in the data for the problem and may be regarded as known. The temperature at the selected point will then be equal to the arithmetic mean of the temperatures at the ends of the random paths leaving the point.

This provides the basis for applying the Monte Carlo method to problems in thermal conduction.

But where in this analogy is the much-needed model of the phenomenon?

Actually, what we have just been doing is developing a model for thermal conduction in a disc. Since the individual heat quanta move randomly, we chose as the basis for our model a situation in which we have objects moving at random, namely: a collection of drunks following random routes through a city. And this model enabled us to solve the problem.

A MODEL FOR THE DRUNK

So far we have been talking about a random path mapped out by a drunk as though it were the actual substance of the model. We even went to some trouble to make sure we had a drunk because we wanted someone who would meander haphazardly about the city.

In order to solve the problem on an electronic computer, however, we have to be able to simulate random paths without recourse to the services of a crowd of alcoholics. How can we do this?

One way of mapping random paths is as follows. We take a sufficiently fine rectangular grid upon which we shall move from node to node. We select any node at all on the grid as our starting point (this corresponds to the street intersection in our drunken city) and then choose one of four directions—up, down, to the right, and to the left—and move to the next node in the chosen direction.

Since the path has to be absolutely random, each of the four directions has to be equally probable. In order to randomize the direction of motion we can use the simple expedient of tossing two coins simultaneously. For each double toss there are four possible results: HH, HT, TH, and TT, where H means head and T tails. We then assign a meaning (a direction) to each of these four results as follows, say:

$$HH = \text{up};\quad TT = \text{down};\quad HT = \text{to the right};\quad TH = \text{to the left}$$

Obviously the tossing of the two coins gives a series of completely random commands. The procedure, then, consists of tossing the coins, reading off the direction, and moving to the adjacent node on the grid in the direction indicated; and then tossing the coins again, reading off the new direction, and moving to the next node in that direction; tossing the coins again,—and so on. The path we eventually obtain, made up of small segments of straight lines parallel to the axes of the grid, simulates random wandering on a plane surface.

To sum up: the essence of the Monte Carlo method is the mathematical simulation of physical experiments that contain an inevitable element of chance. Multiple repetitions of the physical experiment are replaced by multiple repetitions of a calculation based on a mathematical model of the experiment. The only thing that presents any difficulty is the devising of the model. Once this has been done, any problem can be solved by the Monte Carlo method with little further effort, because then it is only a matter of writing a programme and running it on a computer. For this reason the Monte Carlo method may be called the *method of mathematical experimentation* or the *method of statistical trialling*, thus emphasizing the iterative nature of the method.

In conclusion we may observe that the Monte Carlo method was conceived and developed only after high-speed computers appeared on the scene. Manual application of such a method would be completely beside the point because the essential feature of the technique consists in a large number of calculations of a single type. Just as no man, no matter how strong, could have built the Great Pyramid of Cheops unaided, so no man, working alone, could make use of the Monte Carlo method without the help of a computer. The Monte Carlo method is for large, high-speed computers alone.

3. CHANCE IN GAMES

Games provide a rich field for the study of chance. By a 'game' we mean a situation in which two opposing sides having opposing interests interact within a framework of definite rules. The theory of games has recently begun to attract growing interest because of its usefulness in solving problems that arise in many important situations involving conflict, in particular, those that relate to military operations.

It has been found that children's games such as blind man's-buff, hide-and-seek, and others, and also adult games—card-games, for instance—constitute simple, unpretentious models of the relationships that may exist between two countries, let us say, or among a number of firms, and so on.

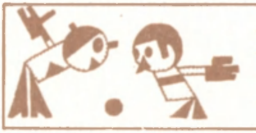
The most essential and specific features of any game are the rivalry and aggressiveness of those taking part.

It is easy to beat an opponent who is inexperienced and unimaginative: you simply pursue a course of action that takes direct advantage of his guilelessness. But suppose he is only pretending to be a simpleton and is all the time devising a subtle trap that will get you into a hopeless position. Inexperienced chess players often lose, for example, by being tempted into rash moves and hoping that their opponent will overlook the threat. The conclusion is obvious: it is useless to rely on the mistakes or the passiveness or the inattention of your opponent, and it is essential to base your strategy on the assumption that he is clever, cautious, and is thirsting for victory just as passionately as you are.

The simplest type of situation that the theory of games can be applied to is really fantastically simple. The game requires two players with exactly opposing interests. Suppose we have a rectangular array like a chess-board with a particular number printed in each square. The rules of the game specify that the first player (A) choose any row of the squares, and the second player (B) any column. The result of these two moves, taken together, is the number lying at the intersection of the row and the column so chosen. We assume, of course, that each player makes his move in ignorance of that of the other player.

If the number so selected has a plus sign in front of it, A wins; if a minus sign, B wins; the number of points won in either case being equal to the absolute value of the number in the square. A simple scheme such as this is representative of the overwhelming majority of conflict situations we encounter in our daily lives.

Let us use the array shown in Figure 46 to see how this sort of game works.



A \ B	B_1	B_2	B_3	B_4
A_1	-2	-1	4	-3
A_2	3	-2	-3	1
A_3	2	1	2	3
A_4	1	-4	-2	5

Figure 46

Studying the numbers, player A soon notices that his fourth row, A_4, offers him his maximum win of five points if B plays his fourth column, B_4. But he realizes that B, being no fool, might answer the move A_4 with B_2, thus winning himself four points. A similar study of the move A_1, promising a best return of four points, leads to the sorry conclusion that B would win with every reply but B_3.

Player A realizes the futility of dreaming about two birds in the bush and resolves to be content with the one in the hand. So he starts looking for a move that will guarantee him a win, even if it is only a small one. And this more sober approach to the matter is immediately rewarded.

His study of the array complete, A sees that he has at his disposal one row, A_3, that gives him a certain win no matter what his opponent does. B notices this also, and realizes that he will lose every time A plays A_3 and that there is nothing he can do about it except to try to keep his losses as low as possible by playing B_2. And so in this particular game one of the players is certain to win, the other (sad but true) certain to lose.

After this detailed analysis of the game the players may decide not to play because the outcome is already clear, or player B may contest the rules because they are clearly biassed against him. The reason for this conclusion is that the optimum strategy for both players is specific and consists in A moving A_3 and B moving B_2, which determines the result of the game without the slightest ambiguity. The determination of optimum strategies is, in fact, the whole point of the theory of games.

The case we have just been examining was a straightforward one in which each player had only four alternatives. But suppose there were a great many more than four—consider, for example, the enormous number of possible moves available to a chess player, particularly in the middle game, or the number of permutations that have to be considered by an industrial planner in deciding how to distribute a work-load among several factories. Nowadays we use high-speed computers to check out all possible courses of action in complex situations and determine the optimum strategy. Computers like these are able to perform hundreds of thousands of arithmetical operations every second. But before we can set the computer to work we must make a correct analysis of the situation and compile a table similar to the one we have just been examining.

And here we run into a new complication. Very often the situation we have to deal with is such that no matter how hard we try we cannot come up with an optimum strategy. Situations of this type are extremely common in games. In fact the overwhelming majority of games do not contain any optimum move at all. The reason for this is plain: if a game contains an optimum move, there is no point in playing it (as our friend *B* observed in the example above). Any interest the game might have had would disappear as soon as it was discovered that there was a single best move. It is for just this reason that no worthwhile game admits of any best moves. A simple coin game will serve as an example.

HEADS OR TAILS?

The game we are going to analyse requires two players, each armed with a coin. Each player places his coin on the back of his hand, choosing freely and independently of the other player's choice whether to show heads or tails. The players then uncover their coins and compare results. If both coins show the same side (both heads or both tails), player *A* wins and player *B* awards him one point; if the coins show different sides (a head and a tail), player *B* wins and takes a point from *A*. So far, it would seem, the game is perfectly simple. The table (Fig. 47) shows,

A \ B	HEADS	TAILS
HEADS	1	-1
TAILS	-1	1

Figure 47

in all, two possible choices for each player. But, as we shall see, the simplicity is deceptive. If it was dead easy to pick the right strategy in the last game we looked at, in this game the players will have to do some thinking.

Suppose player *A* decides to adopt a particular strategy, for example: heads-heads-tails, heads-heads-tails, and so on. Clearly, once *B* notices the pattern, he will immediately adopt a counter-strategy, namely: tails-tails-heads, and so on. *B* will then be a certain winner.

If *A* employs a more complicated strategy, it will be more difficult for *B* to discover what it is; but as soon as he succeeds in working it out, he will start winning again. In the meantime, while he is trying to discover the pattern in *A*'s play, *B* will have to bide his time and choose moves that will result in no particular advantage, trying only not to lose consistently. To this end, *B* need only ensure that his moves are random by tossing his coin in the air and entrusting his choice to chance. His wins and losses will then occur at random and will cancel out on the average. At the same time, he must pay close attention to his opponent's play in order to discover the latter's strategy. As soon as *B* determines *A*'s strategy, he can immediately formulate his own counter-strategy and then go on to win all the time.

Consequently, by adopting a specific or, as we say, determinate strategy, player *A* always finishes up getting the worst of it. While *B* is kept busy deciphering his play, *A* will be winning and losing more or less equally; but once *B* is on to him, *A* can only lose.

But why? Why is this so? Why is it player *A* that has to lose all the time? Surely both players have exactly equal opportunities, have they not?

The point is that A tried to make certain of victory by adopting a determinate strategy. He played according to a strictly determined rule that B was able to work out. Until B finally discovered what A's rule was, both players had equal chances; but once the rule was discovered, A had lost the game.

Player A may try changing his strategy more often, thus compelling B to spend a greater proportion of the time on working out the pattern. But this does not help either. The situation is essentially the same as before: while B is biding his time and simply observing A, both players have equal chances; but B becomes a sure winner again as soon as he discovers a pattern.

The reader has probably noticed that there are periods in this game when both players have equal chances and that these occur when winning or losing depends on moves made entirely at random. And, strange as it may appear to the casual observer, the strategy of completely random moves is the best (optimum) strategy for both players in this particular situation.

Player A is punished with certain defeat for the very reason that by trying to use a determinate strategy he adopted a procedure other than the optimum. The best procedure in this particular game calls for a strategy that relies purely and simply on chance; for although it does not guarantee that you will win, it does not entail certain defeat either, because your opponent will never be able to tell what your next move is going to be, by virtue of your random strategy. In situations involving conflict, therefore, an element of chance often acts as a kind of smoke screen that confuses the enemy and paralyses his attempts at purposeful action.

The decoding of secret ciphers affords a good example of the sort of conflict situation we are talking about. The position here is exactly the same as it was for the game we have just been examining. Any code can be cracked if regularities can be discovered among the symbols of the coded message. An intricate code may similarly be regarded as a smoke screen that allows one side to gain the time that the other wastes on deciphering it. Here, once again, there are two opponents involved: the coder trying to conceal the meaning of his message, and the enemy decoder trying to discover its meaning.

At first glance it may appear that this 'game' is grossly unfair to the coder because he has to transmit, willy-nilly, a meaningful message, which his opponent can always decipher. The decoder derives his confidence of being able to do this from the following well-known postulate of information theory: 'Any coded message can be *deciphered provided that* it is (1) of sufficient length and (2) meaningful.'

The coder is also aware of this postulate and tries to code the message in such a way as to force the enemy decoder to spend as much time as possible deciphering it. Any secret service whose function is the transmission of important messages usually operates in this way.

The most effective way of coding a message is, paradoxically enough, to omit the message from the transmission. This can be done either by using a purely random code or by transmitting a meaningless collection of letters or words. True, such a procedure conceals the meaning from its intended recipient just as effectively as from the enemy. For this reason a random code is only used from time to time mixed with a meaningful one. This procedure causes the opposing side the greatest distress.

When a decoder is confronted with a message in code, the first thing he has to decide is whether it contains any meaning, and, if it does, which parts of the message carry the meaning and which are simply meaningless jumbles of symbols. This is really a most baffling problem calling for the greatest expenditure of time. The actual deciphering of the meaningful parts does not take very long because it is usually done by a high-speed computer.

We can conclude, then, that in conflict situations an element of chance plays a decisive role. It constitutes a reliable means of hampering the enemy's activity and preventing him from gaining the advantage. This makes the introduction of a chance factor imperative in any situation where determinate behaviour leads to defeat.

The most typical application of a chance element for this purpose is to be found in the conduct of war. Any confrontation of the opposing sides involves a search for the optimum mode of behaviour; and this often turns out to be the one that relies on chance.

In economics there is a whole series of situations—particularly where capitalist competition is involved—in which those who emerge victorious are those who adopt a random strategy.

The widespread application of chance in conflict situations arises from the sober realization, and the firm conviction, that a random strategy is the most expedient one and the one most likely to lead to optimum decisions. Those who refuse to accept the optimizing role of chance will be the losers.

4. LEARNING, CONDITIONED REFLEXES, AND CHANCE

We shall now examine a living organism as it lives, develops, acts upon and reacts to its environment. This environment is at once its wet-nurse, nurse, teacher, friend, foe, and judge. 'A man is the child of Mother Nature and Father Chance' wrote S. Lem. Naturally, if an organism is to live in a particular environment it must be adapted to that environment. This means that it must acquire such habits and develop such skills as will make its life more or less bearable.

We shall approach our examination of the learning and adaptational processes of living organisms, not as physiologists, but as technologists whose aim is to apply the principles of living nature to mechanical devices.

It is extremely difficult to construct a mechanical system that will fulfil a complicated function. For example, although it is no easy matter to set up the production line for a car such as Volga, it would be still harder to modify it to produce the Chaika car: we have to construct a completely new production line for the more sophisticated car. And in a few years' time when Chaika is out of date and has been replaced by a more advanced design, we will have to build yet another, still more complex production line.

But perhaps we could use another approach. Perhaps we could make do with only one production line by constructing it in such a way that it could easily be 're-trained' to produce new models.

A ridiculous fantasy?—Not at all: theoretically, such a solution is entirely possible.

But if this is so, why are there no such 're-trainable' production lines in existence?

The reason is that nobody knows at present exactly how they are to be constructed. And this is why engineers have been compelled by sheer technical necessity to study the behaviour of living organisms possessing properties and abilities far in advance of the cleverest machines. The ability to adapt, to learn and to re-learn characteristic of all living organisms will have to be incorporated in the machines of the future as the very basis of their design.

What do we mean by 'learning'? To answer this question we shall have to deal with a number of related topics, in each of which chance plays an important and sometimes a decisive role.

As we have already said, learning, like adaptation, occurs as the result of the interaction of an organism with its environment or *of a pupil with his teacher*. In determining the relationship between pupil and teacher during the learning process, we shall describe one of the most perfect learning systems to be found, for example, in human society. We are referring here to the schoolroom system of learning.

Another kind of learning is based on the imitative abilities of the pupil. Here the pupil tries to duplicate the skilled actions of the teacher, and the latter corrects the pupil's attempts without indulging in explanation.

There is a third kind in which the teaching function is performed entirely by the environment without the aid of any special teacher. This form of learning we may call *self-instruction*.

Let us examine each of these kinds of learning processes separately.

LEARNING IN THE SCHOLROOM

This takes place in two stages. During the first stage (Fig. 48) the teacher communicates to the pupil the information that has to be learned. The pupil receives the information, digests it, and memorizes it; otherwise he learns nothing. During this stage the teacher does not know whether the pupil is actu-

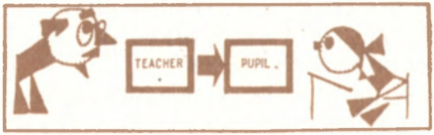


Figure 48

ally receiving the information or whether he is only pretending to do so. Consequently, effective teaching must make provision for testing and assessing the level of comprehension and retention of the information he transmits to the pupil.

This constitutes the second stage of the learning process and is depicted in Figure 49. The teacher asks

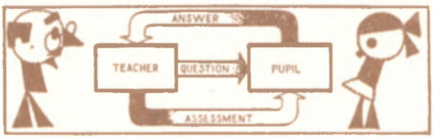


Figure 49

the pupil a test question to find out how much of the material the pupil absorbed during the first stage. The pupil comes up with an answer that enables the teacher to assess how much the pupil has learnt. Next, the teacher informs the pupil of his assessment, thereby indicating how much of the given information he has understood and memorized correctly.

The learning process ends with the teacher's drawing his conclusions and expressing them by rewarding or punishing the pupil. The teacher has at his disposal a large number of methods that have been developed throughout the entire history of pedagogy for exerting his influence over his pupil. The purpose of this influence is to stimulate the pupil's powers of receiving information correctly.

Our treatment of this system of learning is highly superficial and does not deal with such questions as how does the pupil receive and memorize the information, and what is involved in digesting it? These questions belong to the field of educational psychology and will not be discussed here. We only need to emphasize that this system of learning presupposes the existence of ready-made pieces of knowledge that can be transmitted to the pupil and that may be of use to him in the future.

It is for this very reason that the schoolroom system is the one adopted by the State for all schools, colleges, and universities.

In sum, for the schoolroom system to operate there must be—apart from pupils—the following:

1. Teachers well-informed on a particular subject and capable of expounding it.

2. A methodology for testing and assessing the pupils' progress, this being the feedback component in the teacher-pupil relationship.

This system has both advantages and disadvantages. Among its advantages is the fact that it enables the student to master such abstract disciplines as mathematics and physics, as well as to learn the large numbers of facts required in subjects such as history and geography.

All the same, the schoolroom system is a long way from being suitable for teaching everything. Knowledge of the laws of logic, for example, does not necessarily enable one to think logically; analysis of the creative process cannot teach one to be creative; and a study of the origins and characteristics of humour does not always enable one to cultivate a sense of humour. This category of human knowledge requires a different kind of learning process. We shall call it imitative learning.

DO AS I DO

Whereas schoolroom learning is found only within human society, imitative learning is much more widely distributed and occurs throughout the entire animal kingdom.

Imitative learning takes place as follows. Teacher and pupil, finding themselves in a particular situation, each perform certain actions independently of, and without reference to, the other. The consequences of their actions are then compared, and the teacher admonishes the pupil if the latter has chosen a wrong mode of behaviour. In effect (and often in actual fact) the teacher says to his pupil: 'Do as I do.' And that is the essence of imitative learning, the block diagram for which is shown in Figure 50.

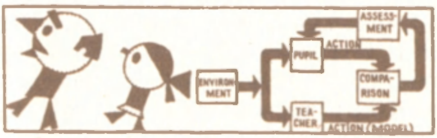


Figure 50

Here, a certain situation within the environment acts upon both teacher and pupil. The teacher either cannot or will not explain to the pupil what he should do in the given circumstances in order to achieve the best results, but provides a model of optimum behaviour instead. The pupil's behaviour is then compared with that of the teacher, and any discrepancies are corrected.

It is characteristic of this form of instruction that the teacher must be able to perform the correct action himself; whether or not he knows in his head how he does it, is less important. Instruction takes the form of a demonstration of the most sensible (optimum) behaviour, rather than a lecture on how to behave in a given situation.

Learning by demonstration is the basic method of instruction for the tradesman and the craftsman. Imitating the best available examples is an essential stage on the way to mastery of the particular trade or craft.

The mirror in a gymnasium, to take another example, is not put there to indulge the vanity of the gymnasts, but is an essential tool enabling self-assessment of proficiency at the moment of doing the exercises. The sporting enthusiast goes along to watch the big match not simply to have his passions roused, but to observe the example of first-class players for later imitation. Creative literature is not just a source of entertainment, but provides a fund of examples of behaviour from which the reader can learn—it is well-known that reading has an enormous influence on the moulding of personality: our favourite literary hero is essentially an example we hold up to ourselves and try to imitate.

This method of learning is widely distributed throughout the animal kingdom, particularly among birds and mammals. Canary-lovers are well aware of this fact. In order to get the young bird to sing he has to have the benefit of training with a master of the art. Both birds are put in a single cage and the pupil starts imitating the maestro. Given a capable pupil, the little school soon turns out a first-rate songster; but if the teacher has the voice of a rooster, his assiduous little pupil will start to crow.

'Birds of a feather flock together' might well become the motto of learning by demonstration.

SELF-INSTRUCTION

Self-instruction is the commonest kind of learning process. It is superficially more simple and is shown schematically in Figure 51. With self-instruction both

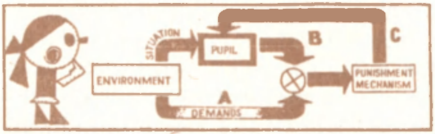


Figure 51

the teaching and the testing of the results are carried out simultaneously by the environment. The pupil is surrounded by the environment; he comes into contact with it and interacts with it. Every environment has its own peculiarities and makes its own special demands on everything within its confines. The pupil must comply with these demands. In the diagram we see how environmental factors act upon the pupil via channel *A*. The pupil's behaviour either does or does not fulfil the demands made by the environment, and he 'informs' the environment of this via channel *B*. This means that to deal with the various situations that he encounters, he develops patterns of behaviour that are controlled by the environment.

We should note that the channels A, B, and so on denote logical relations linking the pupil with the environment and forming a single learning system. Of course, there are no sharply defined channels here, such as there were in the schoolroom system; but links there are, nevertheless, and quite strong ones at that.

Looking at the diagram again, we can say that the environment, as it were, prods the pupil towards a certain code of behaviour. If he completely satisfies all the demands of the environment, that is, if he acts in accordance with the required code of behaviour (and this is tested by comparing the pupil's behaviour with the code, the test being symbolized on the diagram by the circumscribed cross), the punishment channel (C) is not brought into play. But if the pupil comes into conflict with the environment by not satisfying its requirements, some sort of punishment mechanism will operate and notify the pupil via channel C of his departure from the code of behaviour that the environment presupposes.

The probability that the punishment mechanism will operate and the severity of the punishment if it does, both depend on the extent of the pupil's infringement of the environment's rules. It is a curious fact that it is more or less a matter of chance whether punishment—and hence learning—actually occurs. For example, if the pupil breaks the traffic rules, punishment does not invariably follow. The more often he breaks them the more likely he is to be run over or run in and so punished, but in different degrees according to his luck. On the other hand, if he is thoughtless enough to leap from a tenth-storey window, the punishment will be both inevitable and severe.

The result is that even though nobody shows the pupil models of correct behaviour, he nonetheless begins to behave correctly. The pupil travels this unroyal road to learning by himself and without the slightest assistance, but only by virtue of the cuffs, blows, and canings meted out to him in abundance by his environment.

As we can see, the pupil undergoing self-instruction learns perforce from his mistakes; his wishes have nothing to do with it. The more mistakes he makes, the more quickly he learns. This method of learning from one's mistakes is obviously the basic form of learning to be found in nature.

Let us note that self-instruction in no way excludes the possibility of learning without punishment. Learning through positive experience occasionally gives better results than those obtained by learning through punishment.

But what is the actual basis of learning?

From what we have been saying it is obvious that memory plays the leading role in the learning process. But memory alone does not make learning possible.

THE MEMORY MANUAL

Let us imagine that we know a person who is so absent-minded that he forgets absolutely everything. At the same time, he is able to read and to understand what he reads. To enable this person to live and to work in his surroundings, somebody has thoughtfully compiled a most detailed set of instructions and shown him how to use them. These instructions specify the appropriate behaviour for every conceivable life situation (let us suppose for the moment that such a thing is possible). Our absent-minded friend is thus provided with a memory manual that he can carry about under his arm. Could we say that such a person is capable of leading a normal life? No, of course not. Life will require him to answer a multitude of questions in rapid succession; and this he will not be able to do because to find the answer to any one question he will have to start leafing through the whole of his instruction manual from the beginning. He will not be able to take a single step without stopping and reading all the instructions—even if only to find out how to take the next step.

The moral of this story is that mere memory is not enough for learning and developing correct behaviour. We also have to know how to use our memory. And the ability to use one's memory has nothing to do with thumbing one's way through all the instructions one after another (or 'scanning', as this method is called in technical contexts), but consists in being able to switch on the appropriate section of the memory immediately. When we cross the road, for example, we have to remember the traffic rules, and not what we had for yesterday's lunch.

But how can we account for the remarkable ability of the human memory not only to store and to catalogue vast quantities of information, but also to select almost instantaneously any required section within the maze of instructions it contains?

For the answer to this question we shall have to turn to the subject of conditioned reflexes.

WHAT IS A CONDITIONED REFLEX?

A conditioned reflex is the habitual reaction of an organism to a particular stimulus. This habit is acquired as a result of the organism's being placed repeatedly in the same situation. For example, an adult human being never consciously allows his flesh to approach too close to fire, despite its tempting beauty.

Small children, on the other hand, seize burning things eagerly into their little hands because they do not as yet possess the protective conditioned reflex. Only after they have experienced a few painful burns do they acquire the ability to refrain from playing with fire. And this ability we call a conditioned reflex.

Does the child learn anything in the process? Yes, he does; in fact he learns a great deal. He acquires a protective mechanism that consists in a fear of being burned, and so he learns to keep away from fire.

However, an organism can develop conditioned reflexes in other than, let us say, natural conditions: they can be deliberately induced by artificial means.

The reader will have heard, no doubt, of the famous experiments conducted by the great Russian physiologist I. P. Pavlov, in which he succeeded in developing artificial conditioned reflexes in animals. The procedure he used in these experiments was briefly as follows.

Food was placed before a dog, whereupon the dog began to salivate. (It should be noted that salivation in the presence of food is an unconditioned reflex—an inborn characteristic transmitted by heredity.)

The working of this reflex is illustrated in Figure 52. The input 'food' acts upon the system 'dog' (the

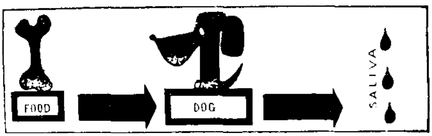


Figure 52

arrow on the left) to produce an immediate flow of saliva, which we mark in the diagram as the output 'saliva'.

At the start of the experiment the dog salivated only in the presence of food. Other stimuli—acoustic ones, for example—did not produce salivation (Fig. 53).

Figure 53

Next, Pavlov superimposed on the food stimulus an acoustic stimulus: a bell was set ringing while the dog was given food, so that both stimuli, 'bell' and 'food', were presented simultaneously. As is shown in Figure 54, the system 'dog' now had two inputs

Figure 54

'food' and 'bell'. One of them, namely 'food', was unquestionably the cause of the output 'saliva'. The other input, 'bell', produced of itself no reaction in the system at all; it merely accompanied the input 'food'. And so the dog salivated in the combined presence of the two stimuli.

After several sessions of this the dog began to react to the bell minus food in the same way as it did to the bell plus food or to food alone, that is, it salivated. And so a conditioned reflex that the dog did not previously possess had been created artificially (Fig. 55).

How did this come about?

Within the dog's nervous system a link was established between the signals 'food' and 'bell'. This link

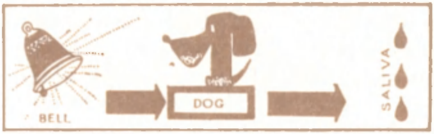


Figure 55

became so strong that one signal could be replaced by the other without producing the slightest change in the organism's reaction. It would be wrong to think that the dog ceased to distinguish between the two signals and that as far as the dog was concerned they merged into a single salivation-producing stimulus. That would be quite wrong. The dog simply established a connection between these two different signals, the connection being that the signal 'bell' was always accompanied by the signal 'food', the latter producing the salivation.

It should be noted that instead of the signal 'bell' Pavlov could have used for this experiment any other signal that was sufficiently distinct and did not frighten the dog, such as a light, say, or stroking, and so on. A conditioned reflex could have been established for these stimuli, too, despite the fact that they are substantially different in character and would be perceived by the dog via different channels in his nervous system, that is, by different senses: hearing, sight, touch, and so on. The dog would salivate in response to all these stimuli even though they have not the slightest physiological connection with eating or with digestion.

A fundamental ingredient of the learning process is the formation of conditioned reflexes that connect various nervous signals within the organism and thus ensure identical reactions to those signals.

But how are these conditioned reflexes formed?

In order to answer this question we shall have to examine the structure of the nervous system.

THE STRUCTURE OF THE NERVOUS SYSTEM

The nervous system of a living organism consists of vast numbers of special cells called neurons. The more complex the organism, the greater the number of neurons it possesses. The nervous system of man, for example, contains about ten thousand million neurons (a one followed by ten noughts). The structure of a neuron is shown schematically in Figure 56

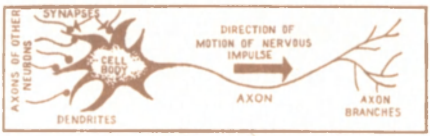


Figure 56

The neuron consists of a central body, from which a large number of short processes called dendrites project, and a long thread-like process, called the axon, terminating in a mass of tree-like branches.

Nervous excitation travels from the dendrites through the cell body and along the axon to the branches at its end. These branches usually lie close to the dendrites of other neurons. When a nervous impulse excites a neuron, the neuron passes it on to other neurons, which in turn pass it on to others, and so on. Obviously, contact between neurons takes place across the gap between the ends of the axon of one neuron and the dendrites of another. If these are sufficiently close together, so that this gap is small, a synaptic junction, or synapse, may be formed, connecting the two neurons at this point.

A synapse resembles resistance in an electrical circuit. If the resistance is high, the connection between the neurons is weak, so that excitation in one neuron may not produce excitation in the other, which means that the signal is not transmitted. If, on the other hand, the resistance of the synapse is low, a strong connection is formed and the second neuron is readily excited by the first one.

The excitation of neurons takes place in an 'all or nothing' fashion. In other words, a neuron may either be excited or not excited; either it transmits the nervous impulse along its axon to synaptic junctions with other neurons, or it does not. There is, as it were, a threshold of sensitivity for neurons: if the resistance of the synapse is above a certain value, the excitation is not transmitted.

But the resistance of synaptic junctions can change. The rule accounting for changes in synaptic resistance was formulated by D.O.Hebb, and is as follows: 'When two neurons having a common synaptic junction are excited simultaneously, the resistance of the exciting synapse decreases.'

In other words, if two neighbouring neurons are simultaneously excited several times for various reasons, the resistance of the synapse joining them decreases and in time may fall below a certain critical value; excitation of one neuron may then lead to excitation of the other. The explanation of this is that during excitation a stable substance that lowers synaptic resistance is formed in the synaptic junction. In time, and in the absence of further excitation, this substance may decay, so that the synapse 'forgets' the fact of simultaneous excitation. The result is that synaptic resistance serves as a memory-carrier within the organism, the elementary memory cell being the solitary synapse.

Naturally, the direction taken by nervous impulses through the nervous system is completely and entirely determined by the resistances of the synaptic junctions that they encounter. So for different synaptic resistances one and the same stimulus can produce different reactions in the organism.

This is shown schematically in Figure 57. Here the same stimulus is applied to the same organism, but at

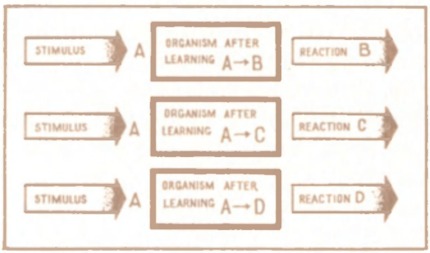


Figure 57

different times after it has been subjected to different training sessions. For example, the ringing of a bell may cause a dog to salivate if the bell was accompanied in the training session by food. The same bell may rouse the dog to anger if it was teased with a stick when the bell was sounded. In either case the bell provides the stimulus; but after the second training session the dog's reaction to it will be diametrically opposed to its earlier reaction. A jovial-genial mood in the first instance will be replaced by an access of rage in the second.

THE MECHANISM OF FORMATION OF CONDITIONED REFLEXES

We are now in a position to describe the mechanism by which conditioned reflexes are formed. Let us re-examine the 'bell-food-saliva' situation that we looked at a short time ago.

The nervous impulses 'bell' and 'food' each take a particular route through the nervous system and meet at a neuron that controls the apparatus for salivation. Since the event 'bell' did not produce salivation at first, the resistance of the corresponding synapse was high to begin with (Fig. 58).

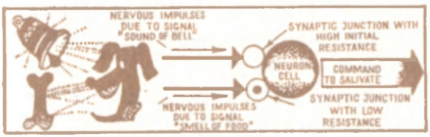


Figure 58

As the result of repeated simultaneous occurrences of the events 'food' and 'bell' this synaptic resistance decreased in accordance with Hebb's rule and the salivation neuron began to be excited by the signal 'bell'. As soon as this happened, the conditioned reflex was established.

From this we see that the formation of a conditioned reflex depends on two factors. One is the repeated si-

multaneous occurrence of the events required for lo
wering the synaptic resistance. The other is the pre
sence within the organism's nervous system of commo
intersection points where the nervous signals repre
senting the two events (and so participating in th
formation of the conditioned reflex) can be juxtaposed

Whereas the first condition is simply a matter o
having the appropriate learning situation, the secon
demands that the organism's nervous system posses
a certain kind of structure.

From this we can draw the fundamental conclusio
that not every organism is capable of developin
conditioned reflexes of any particular type. The refle
ability of a nervous system, that is, its ability to ac
quire conditioned reflexes and therefore its ability t
learn, is determined basically by the number of synap
ses it contains and by the presence of synapses of
particular kind. The greater the number of synapti
junctions, the greater will be the number of condi
tioned reflexes that the organism can acquire, in othe
words, the better will be the organism's ability t
adapt to its environment, the more it will 'learn' an
the 'cleverer' it will become.

We see then that it is not the number of convolution
of the brain that counts, but the brain's inner struc
ture.

We are now in a position to explain the automati
recall mechanism that compels an organism to remem
ber, without the exercise of will and without consci
ous effort, exactly what it needs to remember in
given life situation.

It is clear that automatic recall is associated with
a compound form of conditioned reflex. When an orga
nism encounters the same situation several times, i
develops a behaviour reflex linking the nervous sig
nals that define the situation for the organism with

those that produce the necessary behavioral reaction.

The development of the reflex of fear and avoidance of fire affords an excellent illustration of how the situation 'fire' produces, without reference to will or consciousness, the reflex reactions of 'fear' and 'avoidance'.

Automatic recall does not require any scanning of the memory whatsoever; the instruction manual of the memory does not have to be thumbed through because it opens of its own accord at the right page immediately, the mechanism that does this being triggered off by the external situation itself. All that the organism has to do is to react in accordance with the instructions on the particular 'page' of its memory.

Earlier, we established that the lowering of synaptic resistance is connected with learning. It is natural to ask what the synaptic resistances are like in the newborn organism. If all its synapses have very high resistances, the organism will not be able to *do* anything and, as we can readily appreciate, it will not be able to *learn* to do anything either, because for any conditioned reflexes to form at all it is essential that there be at least a few synapses with low initial resistance. It is found that the nervous systems of the newly born do contain a certain number of inborn low-resistance synapses inherited from the parents. These synapses determine the extremely simple behaviour of the infant: its ability to swallow, to cry, and so on, and these functions are known as unconditioned reflexes. These reflexes enable the infant to exist and, in time, to develop new low-resistance synaptic junctions that will form the basis for conditioned reflexes of ever increasing complexity.

We have already remarked that the acquisition of conditioned reflexes is associated with the structure of the nervous system and is determined by the number of connections between the neurons. The anatomical investigation of these connections among the thousands of millions of neurons in a human nervous system is obviously a colossal undertaking. Nevertheless some progress has been made in this direction. It has been found that there is considerable variation in the lengths of neurons: some neurons can only make contact with others lying close to them; other kinds of neurons reach out over quite large distances (up to fifty centimetres). There are neurons that make contact with only a few other neurons; and there are those that make contact with thousands of others. Within the brain there may be found almost any pattern of interconnection one cares to imagine.

The study of brain structure has led to a most interesting discovery: scientists have been unable to find a single case of identical neuron interconnection patterns in different individuals. Moreover it is now known that such things as ability and genius—and also their opposites—are not transmitted by heredity, and that the structure of the nervous system is not preserved in even the closest (parent-child) of biological relationships.

Apart from this it is clear that the hereditary information could hardly contain full instructions for each of the billions of neuron interconnections within the nervous system of a complex organism. Obviously the bearers of heredity transmit only a few neuron connections that determine the development of the nervous system in general and of the brain in particular; the actual networks of interconnected neurons are then formed subsequently and to a considerable extent at random. As a result, the actual system of neuron networks possessed by any individual is essentially a matter of chance and is therefore unique. The only exceptions to this are the synapses that form unconditioned reflexes, because these are inherited.

We may observe that it is apparently also possible for a few synaptic resistances acquired by conditioning to be transmitted by heredity. This would enable the offspring to benefit to a certain extent from the life experiences of their parents. The following simple experiment supports this belief.

A tale is told of how some white mice acquired a conditioned reflex to the ringing of a bell. The bell was rung for five seconds every time before the mice were fed. At first it took 298 repetitions of the combination of food and bell before the conditioned reflex was established. The same conditioning was repeated with the offspring of these mice, and this generation acquired the reflex after only 114 repetitions of food-and-bell. The third generation needed only twenty-nine repetitions, the fourth generation—eleven, and the fifth—a mere six repetitions.

This means that even if the conditioned reflex itself was not transmitted by heredity, the ability to acquire it—*susceptibility* to it—was.

The fact that the nervous system possesses a random structure is of profound significance. Indeed, in transmitting to the young organism the qualities it will need for a normal existence the parents must 'make provision for' unforeseen circumstances that it may encounter in the course of its life, but that were absent from their own. Consequently, it is essential that the structure of the young organism incorporate a 'chance factor' that will enable it to adapt to new and unforeseeable conditions that its parents could know nothing about.

And so we see that randomness in the structure of the brain and of the nervous system increases the adaptability of a species from generation to generation and guarantees it limitless possibilities of development.

5. CHANCE AND RECOGNITION

In the previous chapter we saw how the random nervous structure of a living organism is able, by interacting with an environment, to develop conditioned reflexes that determine the organism's purposeful behaviour in that environment.

There is a direct relationship between the mechanism of conditioned reflexes and the problem that the organism faces of recognizing any situation that it finds itself in. The solving of this problem forms an extremely important step in the process of an organism's adaptation to its environment.

The point is that no two situations are ever exactly alike. Even the most painstaking attempt to repeat the conditions of an earlier experiment will always contain individual differences distinguishing it from all other experiments. Under natural conditions the difference between similar situations will be all the greater. Every situation an organism has to deal with is essentially a new situation. But, as we saw earlier, in order to develop a conditioned reflex (by lowering a synaptic resistance to the required level) one and the same situation has to be repeated at least a few times.

Here we have an obvious paradox: identical situations are necessary, but they do not exist.

And yet, despite the approximateness with which a given situation repeats itself, the conditioned reflex almost always manages to develop. This bears witness to the fact that *there exists within the organism a me-*

chanism for recognizing situations; and it is this mechanism that permits the organism to regard similar situations as being identical, and to organize its behaviour accordingly.

During the process of recognition the situation is seen, as it were, as a whole, without regard for fine details. In other words, the organism forms a 'general impression' of the surrounding conditions and compares this impression with another impression stored in its memory, and so recognizes it.

The results of this process of recognition cause the organism to begin acting out a pattern of behaviour developed by conditioning in a similar situation. Of course, if the organism finds itself in a given situation for the first time, it has to develop a new pattern of behaviour.

So every action is preceded by recognition.

But what exactly is recognition?

Strictly speaking, by recognition we mean a process of assigning certain phenomena (images) to classes called forms. In other words, recognition involves establishing that a given phenomenon belongs to a particular class of phenomena that are similar to it in some way. For example, the process of sorting photographs of people into the two classes 'men' and 'women' constitutes recognition. Each individual photograph is an image that has to be recognized, that is, referred to one of two classes (or forms).

Recognition is perfectly straightforward if the features upon which it is based (the features that enable us to sort objects into classes) can be clearly set down. For example, we can distinguish between a captain and a lieutenant easily enough by examining the marks on their shoulders: recognition in this case depends entirely upon observing this single feature.

In actual practice, of course, we run into much more

complicated recognition problems than this, whe
the number of distinguishing features is large and, wh
is more, the features themselves are unknown. In situ
tions such as this, it is impossible to supply a simp
formula for recognition.

MAN OR WOMAN?

A couple of examples will show what we mean. Th
first concerns the problem of recognizing a person
sex. The division of all adult human beings into tw
classes—men and women—by means of external fe
tures is an example of recognition. A single glan
suffices to tell us whether a person is a man or a w
man. But how in fact do we tell which is which? Th
answer to this elementary question is not so simpl
Let us consider some of the most likely answers.

Answer No. 1: 'Men wear trousers, women wea
skirts. So if a person is wearing trousers, it is a ma
if skirts, a woman.' The inadequacy of this answe
can be demonstrated quite easily by pointing to th
fact that some women prefer to wear trousers (but re
main, nonetheless, women), and Scotsmen, of cours
like to get about in miniskirts they call kilts. All th
same, it is hard to mistake a woman for a man, eve
if she is wearing trousers.

Answer No. 2: 'Men wear their hair short, wome
wear theirs long. Therefore—' Clearly, if this is ou
criterion, any woman with a boyish coiffure will b
regarded as a man; and any long-haired, bearde
beatnik will go into the class 'women'.

If we examine a few more 'criteria' of this kind, w
arrive at the paradoxical conclusion that it is almos
impossible to state the external characteristics tha
distinguish men from women. And yet in practic

anyone can perform the distinction without the slightest difficulty. How do we account for it?

The point is that in each of these answers we tried to pick out a single, decisive feature that would distinguish men from women, women from men. Of course, the question would be much easier to answer, were it not for the burden of civilization that compels us to wear clothes. Civilization prevents us from indicating any such single, decisive, external feature. There are in fact many external features available to us; but none of them taken by itself is sufficient to resolve the matter.

Let us pose the more general question of how people distinguish visual forms. For example, how does one distinguish the letters of the alphabet without regard to their size, their slant or their style? Could we build a machine that could read?

The second example is not quite so graphic, but is of no less importance in the daily affairs of men. In fact, it was problems of this type that were basically responsible for drawing attention to the study of recognition in the first place. The problem concerns the diagnosis of diseases.

Before a physician treats a patient he must diagnose the patient's condition. In order to carry out this process of recognition he requires certain objective indicators of the patient's state of health, such as temperature, blood pressure, electrocardiogram, and the like. Using these 'input parameters', the physician is able to recognize the disease. How does he do it? Obviously, the business of arriving at a diagnosis calls for a great deal of experience because clear indications of the particular condition are often lacking. When experienced physicians make diagnoses in complicated and controversial cases, they do not use procedures set out in any textbook, but rely instead on intuition based on several years of experience in the field. But what exactly is intuition? Is it possible to recognize an illness by objective means? Could we build a diagnostic machine?

These questions are virtually the same as the ones that we asked ourselves in connection with the problem of recognizing forms.

Both of these problems have been solved in their essentials. Machines that read and machines that diagnose diseases do exist—and they do work. And that is not all: machines that can recognize spoken words have been built, and there are even machines that can distinguish various odours. Obviously, engineers have been applying themselves in earnest to the simulation of sensory organs of living organisms.

But we still want to know what the mechanism of recognizing a situation is really like. Let us consider a simpler example: the process of recognizing the figure '0'.

Since recognition necessarily implies the existence of a number of other forms (classes) that are discarded during the recognition process, let us use the figure '2' to represent such an 'other' form.

We hardly ever make a mistake in recognizing these two figures, hardly ever confuse one with the other however they are written. How do we do it? For the time being we can only suggest answers to this question. However, there are already in existence a number of methods for distinguishing forms—methods that form the theoretical basis of recognition (or reading) machines.

CODING AN IMAGE

The first thing we have to do is to code the images in question. This we do as follows. The object to be analysed is placed on a grid that is divided up into cells. Any cell that contains part of the outline of the object is assigned the value unity; any that does not is assigned the value zero. An example of how this works is shown in Figure 59(*a*) for the figure '2'.

Figure 59

The result is obviously an image formed in noughts and ones; this we transform into the number code for the image by writing out the completed rows of the grid in sequence thus:

$$\underbrace{1111}_{1}\ \underbrace{1001}_{2}\ \underbrace{0010}_{3}\ \underbrace{0110}_{4}\ \underbrace{1111}_{5}$$

rows

A different image would have a different code. Clearly, identical images (but not forms) will have identical codes, and vice versa.

In general, the code for any image drawn on this particular grid can be written in the form of the following sequence:

$$a_1,\ a_2,\ \ldots,\ a_{20}$$

where each *a* represents either zero or unity and is provided with a subscript number indicating a cell on the grid (Fig 59). Here we are considering grid of twenty cells, which codes an image rather roughly, as is shown in Figure 59(*b*), where the cells having the value unity are blacked out.

In order to code an image more finely we would have to use a grid containing a larger number of cells. Figure 59(*c*) shows the same image coded on a grid that is twice as fine as the first one. From this it is quite apparent that the finer grid permits a more accurate representation of the image.

HOW TO TELL A NOUGHT FROM A TWO

We shall now examine a concrete problem in which images belonging to the two classes 'noughts' and 'twos' have to be distinguished. Four representatives of each class are shown in coded form in Figure 60.

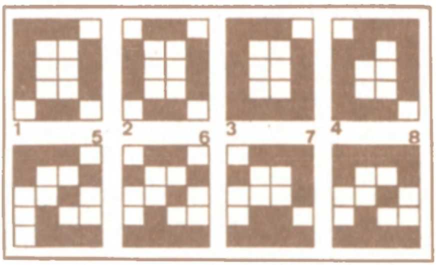


Figure 60

The corresponding codes are written out in the table of Figure 61.

The problem is to determine the features that distinguish these two classes. If we examine the images and their codes carefully, we find the following points of difference:

CLASS	IMAGE No. \ CELL No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
„0"	1	1	1	1	0	1	0	0	1	1	0	0	1	1	0	0	1	0	1	1	0
	2	0	1	1	0	1	0	0	1	1	0	0	1	1	0	0	1	0	1	1	0
	3	1	1	1	0	1	0	0	1	1	0	0	1	1	0	0	1	1	1	1	1
	4	0	1	1	1	1	1	0	1	1	0	0	1	1	0	0	1	1	1	1	0
„2"	5	1	1	1	0	1	0	0	1	0	0	1	0	0	1	0	0	0	1	1	1
	6	0	1	1	0	1	0	0	1	0	0	1	0	0	1	0	0	1	1	1	1
	7	0	1	1	1	1	0	0	1	0	0	0	1	0	1	1	0	1	1	1	1
	8	1	1	1	1	1	0	0	1	0	0	1	0	0	1	0	0	1	1	1	1

Figure 61

Features 1 to 3:

$$a_9 = a_{13} = a_{16} = \begin{cases} 1 \text{ for each image of '0'} \\ 0 \text{ for each image of '2'} \end{cases}$$

Feature 4:

$$a_{14} = \begin{cases} 0 \text{ for each image of '0'} \\ 1 \text{ for each image of '2'} \end{cases}$$

This means that cells 9, 13, 16 and 14 contain distinctive information about the classes, whereas the rest of the cells do not. Since we are only interested in information-bearing features, we can dispense with all the other cells and retain just these four. The images of Figure 60 then take the forms shown in Figure 62.

Figure 62

We can now see that any one of these features will suffice to distinguish between the images. The decision rule for recognizing noughts and twos can be written as follows in terms of the first feature:

$$\text{if } a_9 = 1, \text{ the image is a '0'}$$
$$\text{if } a_9 = 1, \text{ the image is a '2'}$$

Decision rules for the other three features can be formulated in exactly the same way.

Each of these decision rules will distinguish the images concerned with perfect accuracy—so long as there is no interference present. In conditions of interference there may be errors in the coding of the images because a 1 may by chance appear instead of a 0, and vice versa. When this happens, the decision rule will naturally have to take other features into account as well in order to raise the accuracy of the distinction. A rule that takes all four features into account can be written as follows:

$$\text{if } a_9 + a_{13} + a_{16} + \bar{a}_{14} > \delta, \text{ the image is a '0'}$$
$$\text{if } a_9 + a_{13} + a_{16} + \bar{a}_{14} < \delta, \text{ the image is a '2'}$$

where δ is an as yet unknown threshold value.

Here we have introduced the notation:

$$\bar{a} = \begin{cases} 1, & \text{if } a = 0 \\ 0, & \text{if } a = 1 \end{cases}$$

which is called negation or inversion.

We select the value for the threshold δ as follows: If there is no interference at all, we have:

$$a_9 + a_{13} + a_{16} + \bar{a}_{14} = \begin{cases} 4 \text{ for a '0'} \\ 0 \text{ for a '2'} \end{cases}$$

If interference is present, the sum will almost certainly lie between 0 and 4. It is natural, therefore, to define the threshold as the half-sum of these limits, thus:

$$\delta = \frac{1}{2}(4 + 0) = 2$$

Our decision rule then takes the final form:

$$\text{if } a_9 + a_{13} + a_{16} + a_{14} \text{ is} \begin{cases} \text{greater than 2, the image is a '0'} \\ \text{less than 2, the image is a '2'} \end{cases}$$

This rule will work without giving a wrong answer even in cases where interference distorts the information borne by *two* of the critical cells (the state of the non-information-bearing cells, as we have seen, plays no part in distinguishing these images).

Suppose, now, we try to use the above rule to distinguish new images that we have not met before. Can we be sure of being successful?

No, of course we cannot be sure. An image may be so badly distorted that the decision rule just will not work; however, this would be a rather artificial case unlikely to occur in practice. If we avoid trying to 'deceive' the rule deliberately (for nature never deceives intentionally; Einstein formulated this truth in the splendid aphorism: 'God (nature) is cunning, but not ill-intentioned'), it will be equal to the task.

In Figure 63 there are two new images of a '0' and a '2', which, as you can easily verify, are recognized

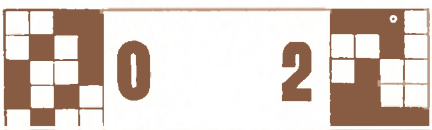


Figure 63

splendidly by our decision rule. The reader may draw several images like these for himself and apply the rule to them.

The fact that it is possible to recognize new, previously unknown images using a rule based on other images is of profound significance. It is palpable proof that features—or rather, combinations of features—singled out on the basis of only a few images contain information about the whole class. It is this that makes it possible to apply a decision rule to situations not previously met with.

We ought to point out, however, that we were able to derive our decision rule in this case by virtue of the fact that the distinguishing features were straightforward and were easily revealed by careful examination of the images. But what would we do if the images were so complicated that simple observation did not enable us to discern their distinguishing features? What then?

In cases like these the processes involved in learning come to our aid; and to deal with them we have to use special teaching machines that are designed to recognize images shown them.

Let us see how one of these machines works: we shall deal with the perceptron constructed by the American scientist Frank Rosenblat. (The name 'perceptron' is derived from the Latin word *perceptio* meaning 'understanding'.)

THE PERCEPTRON

The perceptron was born of attempts to simulate the process of seeing and distinguishing visual shapes as it occurs in the eye-brain system of a living organism. The perceptron also has 'eyes' that perceive a visual shape, it has 'nerves', and finally it has a 'brain' that consists of a device for carrying out analyses and making decisions.

Seeing and recognizing a shape means linking its appearance with a particular irritation pattern already existing in some part of the brain.

The 'eye' (or screen) of the perceptron, like the human eye with its retina consisting of a huge number of light-sensitive rods and cones, also consists of a large number of light-sensitive elements (for this reason it is sometimes referred to as the 'retina' as well). The light-sensitive elements are arranged to operate as follows: when light is incident upon them they register an output voltage ($a=1$); in the absence of light there is no voltage ($a=0$).

Every light-sensitive element is a transformer of light into electrical potential. Each is provided with two leads, one of which passes through a device called an 'invertor' (we have already seen what this means). If the light-sensitive element is illuminated ($a=1$), its invertor registers zero output voltage ($\bar{a}=0$); if there is no light and $a=0$, the invertor registers an output voltage $\bar{a}=1$. An invertor will be symbolized by a dash inside a circle (Fig. 64).

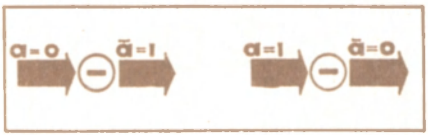


Figure 64

The other lead from the light-sensitive element carries information as to the actual state of the cell. These two leads (cell output and invertor output) terminate in bundles of wires the ends of which are joined to devices known as associative elements (A_1, A_2, ..., A_k in Figure 65). These elements perform the simple task of summing the voltages communicated to them by the leads. We shall see later why they are called 'associative'.

The connections between the light-sensitive elements and the associative elements are very unusual: they are random connections. This randomness is introduced during the wiring-up process. The leads from the light-sensitive elements and the invertors are simply soldered up to the associative elements in a completely haphazard fashion. (Anyone who has tried soldering complex circuitry will appreciate the advantages of the perceptron.) A chance component is thus introduced into the perceptron's circuit by random soldering of the leads.

The output voltages of the associative elements pass to an analyser that interprets them and so decides which class the image belongs to.

Now let us see how the perceptron works. The images

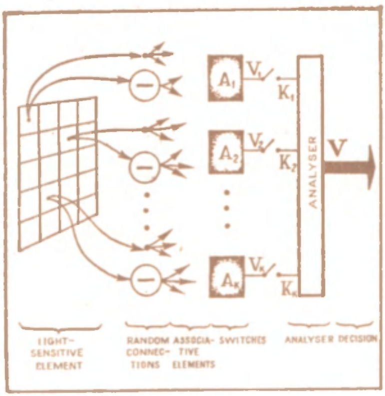


Figure 65

to be analysed—the same noughts and twos as we had before, let us say—are projected on to its retina, and the output voltages V_1, V_2, ..., V_k of the associative elements are observed. These voltages will generally be different for different images. By chance it may happen that the output of a certain associative element (which we shall denote V_i) has one value when the image is a two, and another when the image is a nought. This will only occur if as a result of the random connections the particular associative element happens to be connected to the retina in such a way as to satisfy the decision rule that we discussed earlier. The necessary set of connections is shown in Figure 66; and for the images of Figure 60 we expect:

$$V_i = \begin{cases} 0 \text{ for a '2'} \\ 1 \text{ for a '0'} \end{cases}$$

The working rule for the analyser is then simply:

$$\text{if } V_i \text{ is } \begin{cases} \text{greater than 2, the image is a '0'} \\ \text{less than 2, the image is a '2'} \end{cases}$$

Therefore, the analyser has only to keep track of the ith associative element and can simply ignore the remainder.

Figure 66

Generally speaking, however, we cannot rely on the chance of coming up with a set of connections like this.

The probability of forming a set such as this during the random soldering process is about one in one hundred million.

Well, in that case, how does this perceptron tell a nought from a two?

The answer is that for the time being it cannot. The perceptron that we have described so far is like a baby possessing only the *ability to learn* how to distinguish different shapes. Well then, suppose we 'teach' it.

TEACHING A PERCEPTRON

What does 'teaching' a perceptron mean?

Briefly, it means that we have to develop a decision rule for the analyser. If we knew the distinguishing features of a '0' and a '2', the analyser would be easy to construct. But we do not know what these features are. They will only become apparent during the learning process.

Suppose the process of teaching a perceptron is based on the following idea. We choose associative elements with their output voltages being high for a nought and low for a two.

The procedure is as follows. We project one of the images of a '2', say, on to the perceptron's light-sensitive retina and examine the output voltages of the associative elements. In accordance with the decision rule stated above, we then have to disconnect associative elements with maximum outputs (high voltages) from the analyser by opening the appropriate switches. Conversely, when we project a '0' on the screen, we have to disconnect elements having minimum output voltages. Obviously, after a few cycles of this, only those associative elements that give high voltages for a nought and low voltages for a two will be left 'in the game'. Summing the outputs of this group of elements, we obtain a quantity z.

The decision rule can then be written down as follows:

if z is less than 0, the image is a '2'

if z is greater than 0, the image is a '0'

The perceptron is now able to distinguish these two shapes.

You may wonder why we had to find the sum of all the output voltages of the associative elements: after all, the only elements left in the circuit are those that give high voltages for a nought and low voltages for a two. So what is there to stop us from using the results of just any one of these elements?

The point is that any associative element may make a mistake in determining the class to which a given image belongs. However, it is unlikely that most of them would make the same mistake for one and the same image. So each element can be expected to go wrong for 'its' particular image. Obviously, the overall sum of the associative element voltages gives an average result for all of them; and since most of them recognize the image correctly, the perceptron itself will not make mistakes.

We can understand this characteristic of the machine more clearly by considering the following parallel example.

Suppose there are some rather dull second-formers marooned on a desert island. Each of them has a rather imperfect knowledge of the multiplication table and each has his pet mistakes. One of them believes that twice two is five, another maintains that three threes are ten, and so on. Would these boys be able to compile a correct multiplication table? Of course they would—so long as they put every result to the vote and adopted the one that won the most votes.

The perceptron works in exactly the same way.

We have seen how the perceptron distinguishe geometrical forms, that is, figures, letters, and othe signs. Machines that can read are constructed on ex actly the same principle. They can be designed to rea either printed letters or handwritten ones. Before th machine starts working it has to be taught; and onl after it has done its lessons can it begin reading a text The learning process follows the same pattern a schoolroom learning with a teacher (Fig. 50).

We should point out, of course, that the machin we have been describing is a highly simplified versio of the perceptron. In actual practice the process o teaching the perceptron is far more involved, parti cularly if we take into account the fact that in it modern form it is able to distinguish considerabl more than just two shapes. On one occasion, for exam ple, the experimenters taught a perceptron all twenty six letters of the alphabet with ease, so that it wa able to recognize them in any script.

It is worth noting that the perceptron possesses th remarkable property of being able to generalize. Thi ability distinguishes it from other, less fortunate ma chines of the same general type and considerably en larges its scope. It can not only recognize similar ima ges that it 'sees' for the first time, it can also distin guish badly distorted images. It is also capable o making more involved generalizations.

In one experiment the perceptron was taught to dis tinguish between horizontal and vertical attitudes o a 20×4 rectangle that was projected on to differen parts of the light-sensitive field. It was then asked to distinguish these same two attitudes for various rec tangles of different proportions. The following results were obtained (in percentage correct attitude recog nition):

for a 20×2 rectangle—78%;
for a 20×7 rectangle—100%;
for a 20×15 rectangle—100%;
for a 15×4 rectangle—93%.

This shows that the perceptron learned to distinguish between horizontal and vertical attitudes of rectangles in general. The most remarkable thing is that it acquired this ability after being taught to recognize the attitudes of only one particular rectangle. And this constitutes a first step towards abstract thought.

THE PERCEPTRON AS PHYSICIAN

Still more interesting is the possibility of using the perceptron as a diagnostic machine. Suppose we connect each cell of the light-sensitive screen to one particular indicator of the patient's condition. For example, if the patient feels pain in the region of his heart, cell No. 23, say, of the screen is illuminated ($a_{23}=1$); otherwise the cell remains in the dark, and so on. The indicators of the patient's condition are thus fed in coded form into the perceptron. At the same time, a diagnosis is made by a highly experienced physician who has to decide as accurately as he can what the patient is suffering from. The diseases are then allotted numbers: 1, 2, and so on.

Teaching the perceptron to distinguish various diseases is carried out in the same way as teaching it to recognize visual shapes; that is, by switching off any associative elements that fail to distinguish between the diseases. Then by summing the outputs of all the remaining associative elements the perceptron can be used as a diagnostic machine. If, now, we feed information about a patient into the machine and it registers an output voltage greater than the threshold value δ, the patient is suffering from disease No. 1;

if the voltage is less than δ—disease No. 2. Consequently, the perceptron has learned to diagnose these diseases as accurately as the physician that taught it.

But the perceptron does not necessarily have to be taught by a physician. It can learn from written sources—to be precise, from descriptions of diseases. In this way the machine's memory can be made to retain data concerning a large number of diseases. Such a machine would then be capable of producing more competent diagnoses than even the most experienced physician.

What, after all, guides a physician in making his diagnosis? The answer is: his own experience, his own successes and failures, his sleepless nights, and the applause he gets at conferences. This is probably the most valuable resource available to the experienced specialist and distinguishes him from the novice physician. Moreover the experienced physician can always call to mind the great store of medical anecdotes that he has heard from his colleagues at one time or another ('I remember, once in Tyumen, they wheeled in a man that had...'). Finally, the specialist well remembers all that he has read in the medical journals. These three sources of information all serve the same end: that of making a correct diagnosis. And the greater the amount of information available, the more accurate will be the diagnosis. That is why physicians like to gather together and hold a consultation whenever they are confronted with a particularly difficult case. They do this in order to pool the experience of several physicians.

In the age of cybernetics a consultation like this is conducted in a different way and at a different level because thousands of physicians may 'participate'. Here is how it works.

The perceptron is taught to diagnose various diseases on the basis of material drawn from well-documented cases that have been thoroughly checked and of which a very large number can be collected because people with the same disease show much the same symptoms. This means that when the perceptron makes a diagnosis, it is armed with the experience of a very large number of physicians—so large, in fact, that to bring them all together into a single consultation would be plainly impossible. But the perceptron unites the experience not only of physicians living here and now, but also of those of earlier times and different lands. This makes the machine's diagnoses exceptionally accurate.

However, owing to the complexity of its circuitry the highly specialized diagnostic perceptron is not a practical proposition. And all that we have been saying about diagnosis can be made into a splendid programme for an all-purpose high-speed computer with its vast memory. It is into this memory that the case histories of the diseases are fed.

There is a diagnostic system of this type programmed to diagnose diseases of the heart already in operation at the Professor A.A. Vishnevsky Medical Institute in the Soviet Union. Its memory contains descriptions of heart diseases and their case histories gathered from nearly every corner of the globe. With this imposing 'erudition' at its disposal the machine is able to provide physicians with extremely valuable assistance in diagnosing various heart conditions.

The perceptron need not be restricted to diagnosis: it could also prescribe treatment. To this end it would only need to be taught case histories of successful treatments. In other words, it would have to be informed of the symptoms of the disease and the details of any treatment that resulted in rapid recovery without any unpleasant side effects. Given this training, the perceptron would not only give a correct diagnosis, it would also recommend the most effective treatment.

THE ROLE OF CHANCE IN THE PERCEPTRON

The random connections between the light-sensitive elements and the associative elements in the perceptron are of profound significance particularly in relation to the problem of distinguishing complicated forms. Indeed the chief difficulty with complicated forms is that it is impossible to say anything in advance about their distinctive features. Consequently it is found that more is to be gained by basing the recognition process on random associations and discarding any associations that do not discriminate between the forms.

The reason for making the connections between the retina and the associative elements random is that for any pair of forms there will always necessarily be found a set of associators that give high voltages for images of one form and low voltages for images of the other. If the connections were made according to some predetermined law, rather than according to a table of random numbers, there would always be at least one pair of forms that the perceptron could not distinguish. Consequently, the randomness in the perceptron's circuitry is its guarantee of being able to recognize any form.

As a machine with the ability to learn, the perceptron occupies a position intermediate between ordinary devices, such as cars, radio sets, and so on, on one hand, and biological systems, that is, living organisms, on the other. This peculiarity of the perceptron, together with its remarkable properties, led technologists to focus their attention on nature; the result was the birth of a new science known as *bionics*.

The rapid development of cybernetics from the year 1948 onwards was stimulated by the idea of the universality of control processes. Norbert Wiener, the founder of cybernetics, showed that as a means of attaining given objectives control has a universal character regardless of whether the controlled object is a machine, an organism, or a society. This momentous idea led to the creation of multipurpose control devices and their application in quite unexpected fields of human endeavour.

The invention of the general-purpose computer with its proven potential for almost unlimited application literally brought about a revolution not only in industry, but also in scientific thought. It seemed as though man had awakened to find in his grasp a Firebird that would help him to solve almost all the problems confronting science, technology and society.

But years passed and the Firebird of cybernetics began to lose its brilliant plumage. Something was obviously amiss. Hopes that seemed to be on the point of realization came to nothing. The age of robots and intelligent machines stubbornly refused to dawn. The reasons for this stubborn refusal were regarded at first as 'trivial'.—The available components were insufficiently reliable and could not operate continuously for protracted periods in a machine without requiring replacement or repair... The machines themselves lacked the capacity to solve the problems they were set... There were a few 'elementary' problems that had so far resisted all attempts at programming...

In time these 'trivia' grew into problems of such magnitude as to constitute a very real bar to the further development of cybernetics. By the end of the fifties it was clear that cybernetics stood in need of new ideas and new techniques. These ideas could not be born within cybernetics itself: they had to be looked for elsewhere.

And so living nature was seized upon as a likely source of inspiration. For indeed, our lesser brethren that jump, climb, croak and squeak all about us are capable of solving problems that are well beyond, alas!, the capacities of any computer. What could be easier than to borrow from mother nature the splendid ideas she had developed. The ideas and techniques that cybernetics needed had been right under our noses all the time in living nature—in the biological machines she had created with their extraordinary properties and possibilities.

Thus the new science of bionics came into being, proclaiming as its slogan: 'From living prototypes to engineering models.' Indeed, the raison d'être of bionics was to steal nature's patents.

But you will agree, I think, that one could hardly object to this particular brand of stealing.

Once again cybernetics donned its rose-tinted spectacles. Once again it seemed that the end was in sight—that it was simply a matter of becoming thoroughly conversant with the organization and principles of operation of biological systems and the key to the creation of similar machines would be at hand. Engineers and technologists, especially those with military interests, immediately became absorbed in problems of bionics; and to begin with they produced some promising results.

However, more careful enquiry into the functioning of biological systems showed that the principles on which they operated were, to put it mildly, unsuitable for technological application. An artificial neuron, shall we say, made in the image and likeness of a living neuron, was found to be less useful than standard computer components already in existence. Many more examples could be cited in which 'nature's patents', the substance of the great hopes of bionics, proved useless. Another crisis was coming to a head. The very foundations of bionics were subjected to closer scrutiny and it became apparent that all was not well with its basic assumptions. If you think about it, you will realize that a living organism is a most complex *chemical* device. The biological systems invented by nature operate on combinations of proteins. In a living organism information is carried not only by electrical impulses, but by chemical substances as well. Obviously, any attempt to reproduce the operating principles of biological systems by means of the electrical circuits of technology, as contemporary bionics would have us do, leads to a violation of those very principles. Herein lies the explanation of the extremely moderate successes of bionics.

All the same, how are we to explain the technologists' continuing interest in modern bionics? Here we are evidently dealing with a curious psychological fact: any scientist studying biological objects that are completely new to him knows that his problem is capable of being solved (because nature has already solved it) and is therefore not troubled by any psychological barrier of 'impossibility'. This barrier always represents a threat to the scientist's work. At the back of his mind there is the constant apprehension that the problem he is devoting all his energies to may in fact be insoluble. But the living prototypes remove this barrier. It was thanks to this that the perceptron was invented, to take but one example. According to its creator's original idea the perceptron was to simulate the functioning of the brain. In fact it does nothing of the kind; but it *has* proved to be a first-class inven-

tion that has for many years been providing mathe
maticians and technologists with new ideas.

Where should bionics go from here? Obviously it next step must be to explore the possibilities of usin living organisms themselves—or organs taken fron them—in technology. Nature possesses excellent de vices that are marvellously reliable and amazingl adaptable; and to neglect to use them would be shee extravagance. If we could adapt them to functionin as living machine components, actually using the bio logical systems themselves rather than just their prin ciples, and set them to work for the benefit of mankind a whole new realm of possibilities would open up fo development of cybernetics. The slogan of bionic would then become: 'From living prototypes to livin components.'

If we pursue the matter still further, we may fin ourselves entering the realm of fantasy that has al ready become a commonplace of science fiction. W may imagine computers of the future, not constructe out of living components, but actually grown an reared together with the necessary interconnections The theory of learning would be used in 'bringing up these machines. We could take the brain of any animal for example, provided we were able to make connec tions to its input and output 'terminals', and then relying on the ability of the living brain to form cross connections of the conditioned reflex type, we coul teach it to solve the problems we were concerned with

And so: 'From living prototypes through living com ponents to living machines'—that is the true sloga of bionics.

6. CHANCE, SELECTION, AND EVOLUTION

No sooner had Norbert Wiener founded cybernetics than arguments began as to who had really been the first in the field. Ostrogradsky was mentioned, so were Polzunov and Watt, and also Lomonosov. The writer was also a party to these disputes and, foaming at the mouth, would affirm the priority of none other than Kozma Prutkov. After all, it was Kozma who had uttered the immortal words: 'If you tap a mare on the nose, she will whisk her tail', which clearly expresses the functional relation that exists between a tap on a mare's nose and a whisk of her tail.

Similar logical transformations of the type 'tap nose-whisk tail' or 'tap nose-twirl tail' form the basis of present-day cybernetic devices. The professionals are well aware of this; but they preserve a modest silence about their organic link with Kozma Prutkov. The writer, however, resolutely defends the claim of that erstwhile Director of the Assay Office, that great poet and thinker. It is high time an injustice of history was put right and Kozma Prutkov was recognized as the true father of cybernetics.

The reader will have realized, of course, that one could use a similar argument to demonstrate any priority at all (even that of Buridan's Ass) in any sphere, not excluding radio astronomy. But not only is it right and proper to talk about the influence of this or that 'father' on the birth and development of a particular science, it is downright essential.

Charles Darwin, for example, the creator of the theory of evolution, has had an enormous influence on the development of modern cybernetics.

One could hardly imagine anything more natural and more complex than a living creature. But what exactly is life? Even today science is unable to provide us with a rigorous answer.

However, for the purposes of our 'cybernetical' discussion of life we may limit ourselves to its three fundamental characteristics.

THREE CHARACTERISTICS OF LIFE

1. Reproduction: the ability to produce an organism similar to oneself.

2. Heredity: the ability to transmit parental characters to the offspring. This conservative property helps to preserve in an organism the characteristics of its parents. It is easy to imagine the confusion that would result if this property were to be lost.

3. Variability: the ability to exhibit variation (mutation). This property guarantees the offspring the splendours of individuality and enables it to avoid being a carbon copy or an arithmetic mean of its parents.

It is difficult to overestimate the significance of these three factors in their relation to life. Without reproduction life would simply cease to exist. Without heredity there would be no continuity from generation to generation and the specific characteristics of the parents would not be transmitted to the offspring. And finally, without mutation there would be no variety and the development of life would never have progressed beyond its primordial state.

The chance factor so necessary to evolution is introduced by mutations that are responsible for the variability that endows us with such valuable and essential individual characteristics.

The tissues of a living organism consist of cells. Each cell contains a nucleus. The nucleus, in its turn, contains chromosomes—long, slender threads visible only under the most powerful microscope. The chromosomes carry all the hereditary information about the organism. The process of cell division begins with the chromosomes. Each chromosome, as it were, doubles and forms two identical halves, which immediately separate. When all the chromosomes have divided, that is, when the nucleus has divided, the remaining material of the cell divides. Each half then forms a new cell.

In this way one cell becomes at first two completely identical cells, then, after a second division, four cells, then eight, then sixteen, and so on.

The doubling of the chromosomes takes place with extraordinary precision: one could hardly find anything in the whole realm of technology to equal so rigorously faultless a mechanism. During the formation of a new organism millions of cells are formed out of a single cell, all of them having identical chromosomes.

Nothing in our world, however, is absolute: even such a faultless process as this has its limit of accuracy. Now and then, at very rate intervals—perhaps once in a million cell divisions—something goes wrong: there is a chance discrepancy in the new sets of chromosomes and the hereditary information that they contain changes somewhat. This occurs when some kind of chance interference happens to affect some chromosome, causing it to become slightly different (chromosomes also live in a chancy world). This process *of random change in chromosomes is called mutation*.

When a chromosome that has undergone mutation doubles again, it reproduces itself accurately, as before, repeating the mutated structure. Consequently the 'legacy' of a mutated chromosome is also mutated.

What is the result of mutations, then? Do they, perhaps, have no profound effect on an organism? How could an insignificant change in the structure of a chromosome affect the development of an entire organism?

There will no doubt as to the answer to these questions if it is remembered that a chromosome is essentially a system of commands that are issued during the process of an organism's development. These commands shape the organism. Clearly, the loss of one of these commands, or its replacement by another, will affect the development of individual organs and hence the organism as a whole. And since mutations occur in random fashion, they lead to the appearance of peculiarly individual characteristics in the developing organism. Mutations result in individual features that distinguish an organism both from its parents and from other members of its own generation. Because of the random nature of mutation these distinguishing features may appear in any part or function of the organism.

A mutation can have fatal results if it upsets the functioning of a vital organ or deprives such an organ of certain of its adaptive properties.

A mutation may be beneficial if it results in qualities that enable an organism the better to adapt to its environment.

Finally, a mutation may be neutral in its effects, in other words, it may, for the time being, be neither good nor bad so far as the survival of the organism is concerned (a change in the shape of one's nose is an example).

We can now appreciate that every organism differs in random fashion from every other similar organism. Whenever a mutation occurs, nature takes a chance step, as it were, into the unknown. This step is then examined by life. If an organism whose development is affected by a particular chance mutation proves to be less stable and less well adapted to its environment, it dies sooner than others of the same species. This corresponds to a step taken in the wrong direction—a failure. But because it dies young, such an organism will not, as a rule, perpetuate the error in any descendants (since it dies too early to reproduce). On the other hand, if through mutation an organism happens to acquire new adaptive properties, it will live to reproduce and consolidate these properties in its descendants. This, as Charles Darwin discovered, is exactly how natural selection works.

Consequently, if mutation can be considered to produce random deviations from a certain average condition among the organisms representing a given moment in the evolutionary history of a species, then natural selection may be regarded as evaluating, as it were, the results of these deviations.

Natural selection takes place according to the following principle: the best adapted are those that reproduce and multiply. Mutations provide the raw material for the operation of this formula by producing organisms exhibiting greater and lesser degrees of adaptability. Clearly, if there were no mutations, we would not be able to observe in the living organism that 'cleverness' of construction, that remarkable adaptation to environment, that never ceases to fill us with wonder and delight.

Mutations represent, therefore, one of the greatest driving forces of evolution; and, insofar as the evolutionary process is a never-ending one, they remain as necessary as ever to the further development of life on earth.

This is one aspect of the phenomenon.

The other is that most new mutations are harmful or even fatal to an organism.

The reason for this is that each organism is the result of a prolonged process of evolution and is therefore thoroughly adapted to its environment down to the minutest details. Consequently, not every chance alteration in its structure is by any means to its advantage. Rather the opposite is true. In order to improve a highly organized organism, mutations of a special kind are called for, and, as one might expect, are of such rare occurrence that a considerable period of time may elapse before a desirable mutation appears. It may even happen that a species will die out while it is 'waiting' for the hoped-for mutation—not so much through the lack of this mutation, but through an overabundance of unnecessary and harmful mutations.

The result is that mutations may be just as harmful as they are necessary. A species that mutates too frequently—under the influence, say, of radioactivity—may vanish because too many individuals of the species have become weak and ephemeral as a result of unsuccessful mutations. On the other hand, a species that does not mutate often enough will not be able to survive once the ever changing conditions of life demand new adaptive abilities that will not be forthcoming in the species owing to the lack of sufficient variety amongst its members. This is evidently the reason why in comparatively recent times the mammoth became extinct: it was unable to adapt to the sudden cooling of its environment during the Ice Age.

Let us consider the following rare, but perfectly possible, situation. Suppose there is a species—animal or vegetable—living in complete harmony with its surroundings. It has almost no competition for food or shelter and almost no enemies. The individuals of this species are strong, healthy specimens of sound constitution. They multiply rapidly, but they are not in any present danger of overpopulation. Deterioration of the species due to mutations is as yet at an insignificant level and cannot dim the healthy optimism of this happy breed. Then, all of a sudden, their golden age comes to an end: there is a sudden and dramatic change in their external conditions, due, let us say, to the appearance of powerful competitors on the scene. Immediately, the cruel mechanism of natural selection goes into action: only a mutation that would enable them, if not to defeat their competitors, then at least to coexist with them could possibly save this species. And if the needed mutation came too late, the species would perish.

... AND ITS BLOCK DIAGRAM

In painting so joyless a picture the writer certainly had no intention horrifying the reader with the 'pitilessness' of the laws of nature. Most certainly not. The fact is that this example provides, I think, a good illustration of the relationship between mutations and natural selection. We can depict this relationship in the form of a block diagram showing how species and environment interact in the process of natural selection (Fig. 67). The diagram shows how the environment acts upon the species, making certain demands upon it. These demands are formulated, as it were, along channel *A*. The species develops a certain pattern of behaviour in the given environment and 'com-

municates' this behaviour along channel *B* for co
parison with the demands of the environment. T
results of this comparison stimulate the mechanis
of natural selection via channel *C*, the degree of st
mulation depending on the extent to which the sp
cies fails to conform to the demands of the enviro
ment. If the species fulfils all the demands of the e
vironment and its behaviour does not violate the env
ronment's rules, natural selection does not act; su
a state of affairs, however, is very rare. Natural s
lection acts upon the species via channel *D*. On t

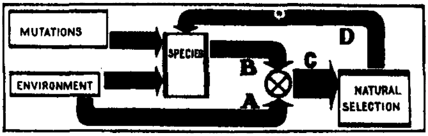


Figure 67

of all this the species is continuously subjected
random mutations.

The diagram works as follows. A change in exter
conditions—in the environment—results in either
appearance or the intensification of a contradicti
between the behaviour of the species and the dema
of the environment in which he lives. This cont
diction stimulates and intensifies the action of natu
selection, as a result of which only the best adap
individuals survive to produce offspring. Mutati
create a variety of deviations from the species aver
in different individuals. Owing to the random nat
of these deviations some of the individuals with n
traits—mutants, as they are called—may be bet
adapted to the demands of the environment th

others, and may form the basis of a new species. The rest will mostly perish as a result of the merciless intervention of natural selection.

THE HOMEOSTAT—A MODEL OF SELECTION

In 1951 an Englishman by the name of R. Ashby constructed a device that operated in much the same way as a species in process of adapting to its environment. He called this device the homeostat (from 'homeostasis'—the maintaining of the properties of a system within certain desired limits).

The homeostat is a dynamic system that may be in either a stable or an unstable condition, depending on the values assigned to its parameters. We use the term 'dynamic' to describe a system whose behaviour depends on its immediate past history. A stone, for instance, is a typical example of a dynamic system: the law of inertia guarantees its dependence on its recent past. If the stone is moving in a particular direction, that direction can only be changed by a definite force—the force of gravity, say—and the new direction will depend on the direction in which this force acts. Herein lies the dependence of the stone on its past history. If, on the other hand, the stone is lying motionless, then, in the absence of any force, it will continue to remain in the same spot.

We now need to distinguish the two states in which a dynamic system may exist: the stable (unchanging) state, and the unstable state (in which the motion changes). A stone flying through the air is an example of an unstable system; a stone lying on the road constitutes a stable system. Consider an ordinary clock: if it is working, it constitutes an unstable (self-excited) system; if it is broken, it constitutes a stable system.

Any single system may exist in a number of sta states. A television tower lying on its side is in a sta state just as much as one standing upright is. (T the first state is more stable than the second is anot matter and explains why a standing tower may ad a lying position during an earthquake, and why one has ever seen a recumbent tower suddenly er itself.)

But let us return to the homeostat. Like any ot dynamic system, the homeostat could exist in one two states—stable and unstable. In the stable st it was motionless and unchanging; but in the unsta state it 'rebelled', and its behaviour transgressed bounds of desirability.

We shall not describe precisely what the homeos did that was undesirable, because that is merely technical detail (the homeostat was only a mechani gadget, after all). The important thing is that one the states—the stable one—was desirable, the ot —the unstable state—undesirable.

The transition from the stable state to the unsta state took place under the action of uncontrolla factors: environmental effects; internal maladju ments within the device itself; and much more besid In other words, by virtue of the second law of th modynamics the homeostat tended towards instal ity.

Feedback was then introduced with a view to co trolling the homeostat. It worked as follows: as so as the homeostat became unstable, the several pa meters within it that determined its behaviour beg to change at random. In other words, it began a p cess of random search that continued until by cha it ceased its rebellious behaviour and returned to stable state. When this happened, it meant that homeostat had chanced to hit upon those values the control parameters that were necessary for stability. Thereupon the homeostat would end its random search and become dormant until a change in the environment or an internal maladjustment sent it on the rampage again.

Here the random search simulates the random mutations occurring within a species. The search continues until by chance the homeostat hits upon such values of its parameters as will ensure its stability. This event corresponds to the occurrence of an essential mutation. Thereupon the parameters of the homeostat stabilize and the mechanism for finding new 'mutations' switches itself off until for one reason or another the homeostat becomes unstable again—that is, until the external conditions change and thus necessitate the appearance of new mutations.

The block diagram of the homeostat is shown in Figure 68. As we have already seen, a dynamic system can have various parameters: values for these enter the system via channels *A*. These values are selected by a random numbers generator and transferred to the system along channel *B*. If the system is in an unstable state, the control block switches on the random numbers generator. The generator begins produc-

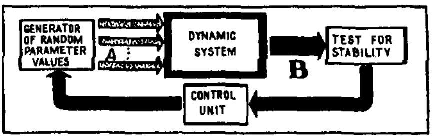


Figure 68

ing values for the parameters and feeding them into the system—trying them on for size, as it were. This process continues until the system becomes stable again.

When the control block receives the informatio that the system is stable, it switches the generato *off, so that the last values of the parameters—the one* that stabilized the system—are retained.

We see, then, that Ashby's homeostat represents *tolerable copy of the adaptive mechanism of livin* species and therefore serves as a natural selection simulator.

We have shown—and we emphasize it again—that species adapts to its environment entirely by chance. The chance element takes the form of mutations tha produce various random deviations from the hypothetical average among the individuals of the species. As a result of natural selection, individuals that suffe unsuccessful mutations die out; whereas organisms tha exhibit favourable changes will perpetuate the improvement in their progeny. In this way the specie reaches a stable equilibrium with respect to its environment. If the external conditions change, the mechanism of mutation plus natural selection begins a 'search' that continues until it brings the species back to a stable state.

The homeostat works in a similar way. It also searches for stability in a purely random fashion, and eventually finds and settles upon just those values of its parameters that correspond to a stable state. If an external circumstance upsets the stability of the homeostat, the mechanism for random selection of parameter values switches on and goes on working until stability is re-established, whereupon it switches off again.

The homeostat is like a sleeping cat. If you disturb a sleeping cat, it will wake up, select another comfortable spot, arrange itself to its liking, and drop off to sleep again. In exactly the same way, the homeostat 'wakes up', looks about in random fashion for such values of its parameters as will allow it to find a new stable condition, and, having found them, switches off its random search mechanism and 'falls asleep' again.

THE INTELLECT INTENSIFIER

The idea of random search that Ashby derived from his observations of nature is of enormous theoretical and practical significance. Ashby's study of the role of chance in nature gave him the remarkable idea of exploiting the boundless riches of chance. What, indeed, could be simpler than a random generator? 'Noise' constitutes an inexhaustible source of chance, a source that is easy to tap and costs virtually nothing. Consequently, we have the raw material available in abundance. But what can we make out of this material? The answer is: a great deal, if not everything.

A chance combination of letters may result not only in any known word, but also in words previously unknown, words that would possibly never have been thought of otherwise. A chance combination of words may form any sentence, that is, any finished thought that has already been expressed by men or that still awaits expression by our descendants. A chance combination of sentences may result in any work of art, or a description of any scientific investigation, or a report of any discovery made by man or still waiting to be made in the future. In general, chance conceals within itself limitless possibilities.

By combining letters, words and phrases at random we can extract new data, new results, and new thoughts. In short, we can create new information out of the raw material of chance.

We should note that this idea was first expressed—and ridiculed—nearly two centuries ago by Jonathan

Swift (who, by the way, is yet another pretender to the title of founder of cybernetics) in his well known novel *Gulliver's Travels*. When Gulliver arrived on the notorious island of Laputa, he saw how the Laputians created new scientific and artistic works by means of a machine that ran through all possible combinations of a set of one thousand letters. Bearing in mind that any scientific discovery can be incorporated in a summary containing one thousand letters, the Laputians hoped, not entirely without foundation, to assemble all possible scientific articles. Such a tempting prospect might easily beguile others besides the rather silly Laputians, for it implies that one could study the whole world from the comfort of one's armchair.

This paradoxical conclusion yields little in practice because although such a method will produce some true information, it will also produce an enormous mass of false information having all the outward appearances of being true; and the amount of false information will certainly far exceed the amount of truth.

Consequently, if we propose to make use of this idea, we must be able to weed out all the nonsense and all the falsehood. We can do this only by a process of selection.

It was by following this line of thought that Ashby hit upon the idea of a *selection intensifier*.

The selection intensifier works as follows. The source of chance ('noise') is tapped by a device that prints out a continuous stream of letters of the alphabet, each letter corresponding to a particular noise level. This stream of letters is tested against certain grammatical criteria and parts of it that can be regarded as words are set aside for further processing. A 'word' such as *strl* would be discarded because it has no vowel.

The next test picks out from all these words only those that form meaningful sentences.

A further test is needed to expunge any recognizably erroneous phrases, leaving only those that clearly do not contradict human experience.

After this, original ideas are separated out from the mass of trivia representing ideas that are already known or that could be readily derived from those that are known.

The final stage in the selection process is one that must be carried out at the highest level with the aid of the most refined criteria and can obviously be performed only by a human being. Its purpose is to decide which of the new ideas is to be subjected to experimental verification—which, as always, will have the final say.

In this way new information is obtained from a series of selective processes each of which is performed with the aid of various selection criteria.

The block diagram for the *abstract thought intensifier*, as Ashby called it, is shown in Figure 69. Here

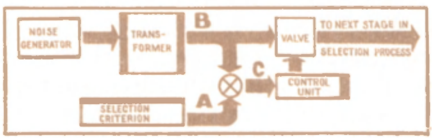


Figure 69

the transformer processes noise so as to produce at its output *B* a random stream of the material forming the subject of the selection process. The transformer 'rephrases', as it were, the information it receives at its input in the form of noise. As before, this information is meaningless. The transformer's product is compared with the appropriate criterion via channel *B*. If this product satisfies the criterion, the control block is informed accordingly, via channel *C*, and opens the valve to allow the selected information to proceed to the next, more advanced stage of the selection process.

In this way it is possible to generate information that was previously completely unknown. True, the process may absorb a great deal of time; but if the various selection stages can be carried out at extremely high speeds, the time required can be significantly reduced.

ARTIFICIAL SELECTION AS AN INTENSIFIER

It should be observed that Ashby's proposal to use multiple selection for the purpose of obtaining hitherto unknown information from chance interference is not a new one. Animal and plant breeders use exactly the same method in developing new strains and varieties—the method known as artificial selection.

Artificial selection is essentially very simple and has long been used by man in his day-to-day activities. Somebody notices that an organism has developed some useful peculiarity (as the result of a chance mutation) and decides to intensify it. At the first stage of the selection process that follows, the criterion is simply the presence—or even just the promise—of the given peculiarity. In other words, organisms are selected that either possess the feature or only give an indication of it. At the second stage—the selection of the offspring of the individuals selected at the first stage—a tougher criterion is applied and only those individuals that quite clearly possess the feature quality. And so on. Finally, a stage is reached at which the individuals possess the useful feature developed to an adequate degree. It only remains then to destroy the organisms not possessing the feature, and the new breed is ready.

The block diagram for artificial selection is shown in Figure 70. Here, mutations act upon the species

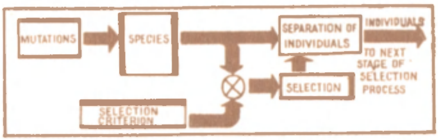


Figure 70

to produce individuals differing in random fashion from the norm for the species. Individuals exhibiting those differences that satisfy the selection criterion are selected by the breeder and proceed to the next stage of the selection process. Those not exhibiting the desired differences are destroyed.

It is obvious that the artificial selection diagram is closely similar to the block diagram for Ashby's selection intensifier. In fact, Ashby's intensifier is virtually a model of the artificial selection process.

As we conclude this chapter, we should stress that it was chance, and chance alone, that perfected the complex forms of adaptation of organisms to their environment, such as we may observe all around us. It is thanks only to chance that the enormous numbers of plant and animal species came into being. It is thanks only to chance that man himself appeared on earth. This striking fact of nature only became clear and comprehensible after Charles Darwin had explained the mechanism of formation of adaptive features. Until then, the only explanation considered possible relied on the idea, supported by religion, of the purposefulness and the rationality of nature. It is now clear that nature is entirely lacking in any sort of purposefulness or rational approach. If we are to speak of any rational principle in nature, then that principle can only be chance: for it is chance, acting in collaboration with selection, that constitutes nature's 'reason'.

7. SELF-ADJUSTMENT

ON CONNECTIONS

There is a well-known game that starts off by one player asking: 'What is the connection between...?' Two entirely different things are named and we are asked to find the connection between them. For example: how does the number of holes in a Swiss cheese affect the maximum speed of a Moskvich motor car: or, how does an eclipse of the moon affect the flavour of a shashlik. When we give up, the first player has to find the connection and explain it.

The most ridiculous thing about this is that such connections do in fact exist: the maximum speed of a Moskvich does depend on the number of holes in a Swiss cheese, and the flavour of a shashlik is somehow affected by an eclipse of the moon. But they are *weak* connections; and even if we could establish and investigate them, we should hardly entertain any hope of being able to use them. It is for just this reason that Moskvich owners trying to get the most out of their cars pay more attention to compression rations and octane ratings than to the finer points of cheese-making. And the gourmet ordering a shashlik is usually more interested in the qualifications of the chef than in the intricacies of the lunar calendar.

ENVIRONMENT AND OBJECT

When we select an object for study, out of the mass of material things and phenomena that surround us, we must preserve its connections with the outside world; otherwise it would cease to function normally and the study of it would not yield the desired results. By 'outside world' we mean the environment that the object is in most intimate contact with. Consequently, 'environment' includes everything that influences the object's behaviour without actually forming part of the object itself.

The interaction of object and environment can be represented by a block diagram such as that shown in Figure 71.

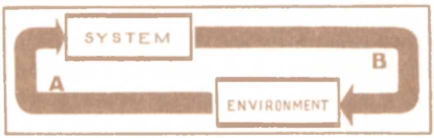


Figure 71

Here, the arrow *A* represents the action of the environment on the object we are studying; the arrow *B* represents the effect of the object on its environment. Using the convenient terminology of communication theory, we can call *A* the channel along which the environment acts upon the system under study; the system (object), then, exerts its influence on the environment along channel *B*.

Suppose, for example, we choose a thermometer as the object of our investigation. Heat is transmitted to the thermometer from the outside via channel *A*, and channel *B* informs us of the temperature in degrees. So channel *A* transmits heat, and channel *B* transmits information about the temperature of the surroundings (of which the observer, of course, also forms a part). There are other factors linking the thermometer with its environment—gravity, for example, but their connection with the thermometer is weak, and they are therefore ignored.

To take another example: let us consider an automatic lathe. Along channel *A* the lathe receives the blanks for processing, together with power and lubricant; along channel *B* it transmits finished components and vibrations back to the environment, or informs the outside world by means of crackling noises or silence that it is out of order. The fact that the lathe is illuminated by sunlight would constitute a weak connection between it and its environment, and can, of course, be safely ignored.

Now for an example from biology. A living organism always functions within a particular environment. This environment may be a forest, a desert, water, a chemical flask, and so on. Along channel *A* the organism receives its food and all manner of external stimuli; along channel *B* the organism acts upon its environment, changing its position within the environment, and so on.

Many more similar examples could be cited to illustrate the interaction of object and environment.

This is not a piece of idle theorizing; it contains a profoundly significant idea. It establishes the strict interrelationships that exist between the objects of the real world and essentially distinguishes and defines the fundamental causal connections that we seek to understand. Moreover, since every system has distinctively individual properties characterizing the connection between its input A and its output B, we can study any object simply by observing its A and B.

Different systems have different properties and different kinds of connections between their inputs and their outputs. Very often these individual properties can be expressed by a set of numbers that are usually called parameters.

The parameters of the thermometer, for example, are the quantity of mercury it contains, the diameter of the capillary and the distance between divisions on the scale. These three parameters together determine the connection between the thermometer's input and its output. If but one of them changes, the connection between the temperature of the mercury and the reading on the scale will also change, and the thermometer will begin to read false.

The automatic lathe takes in blanks at its input and turns out finished components at its output. The parameters of the lathe are its cutting cycle, the cutting speeds and rake angles of the various tools, the positioning and the rates of feed of the tools, the materials the tools are made of and so on. These parameters determine the dimensions of the finished product and its quality. Their values are by no means chosen at random. They depend on the material of the blank and the shape of the finished part, and their choice is influenced by such considerations as economy, the surface finish requirements of the component, the need to minimize tool wear, and many other factors. In one way or another the values of the lathe parameters are established in advance by taking into account all the requirements of both the lathe and the finished product.

For the time being we shall consider only one of these requirements: the quality of the finished product. Only when we have satisfied this basic requirement can we concern ourselves with such matters as tool wear, depreciation of the lathe, and so on.

So we shall demand from the lathe only one thing: that it turn out the best possible product, in other words, that it produce components that approach as closely as possible the ideal depicted on the designer's drawings.

It might be objected that there is no point in trying to produce components with highly accurate dimensions if tolerances are specified within which the component will be perfectly acceptable. On the other hand, it is well known that chance interference is a force to be reckoned with where manufacturing is concerned. Therefore we must constantly strive for the most accurate dimensions we can achieve, so that it will be more difficult for interference to send the dimensions beyond the stated tolerances and the percentage of defective parts may be significantly reduced.

CLOSENESS TO THE IDEAL

Naturally, if we are going to adjust an automatic lathe, we need to be able to determine the quality of its output. For this we require an *estimator* of the finished component's *closeness to perfection*. This estimator would measure the quality of the lathe's work.

It is wise to formulate an estimator like this in as precise a manner as possible—in other words, as a number. For a machine tool a suitable quality estimator is the sum of the differences between the dimensions of the finished component and the dimensions specified in the drawing. When the lathe is working at its best, the value of the quality estimator will be zero, meaning that the ideal has been attained (it is hardly worth remarking that this value of the estimator will never occur in practice owing to the impossibility of achieving absolute accuracy in the component's dimensions). If the sum total of the differences between the actual and the ideal dimensions is equal, say, to one millimetre, we say that the 'distance from the ideal' is one millimetre.

Other quality estimators may be used—the proportion of defective components, for example. Occasionally some additional characteristic may form part of the estimator. But whatever the case, when we decide on a particular estimator, we must ensure that there is only *one* and that its minimum value does correspond to the ideal we are aiming at. When people talk about an ideal that they would like to attain, they often use a series of superlatives—'the cheapest', 'the most accurate', 'the most beautiful' and so on—in an effort to cram into their ideal as many of the 'most bestest' qualities as they can. In choosing a bicycle, for example, the customer will insist that it be: (1) the most reliable; (2) the simplest to operate; (3) the cheapest; (4) the most attractive; and so on. After he has examined a few machines, however, he suddenly realizes that no one bicycle can satisfy all these requirements simultaneously. (All this presupposes, of course, that the customer has a choice; if he has none, he is unlikely to be nagged by doubts of this nature.)

This was the sort of 'unfortunate' situation, if you will remember, that Agafya Tikhonovna, the bride in Gogol's play *The Wedding*, found herself in. She had to choose one of four suitors, and instead of applying a single criterion to determine which would make the ideal husband, she tried several at once. And the poor girl finished up in an agony of doubt as a result.

'Dear oh dear, it's so hard to decide,' Agafya Tikhonovna laments. 'If only there were only one of them, or even two—but four! How am I ever to decide? Nikanor Ivanovich is handsome, but rather thin, of course,—Ivan Kuzmich is handsome, too. And to be quite truthful, so is Ivan Pavlovich—a little fat, perhaps, but a fine figure of a man all the same. What am I to do, I ask you? Baltazar Baltazarovich is also a man of estimable qualities. It's so hard to decide; I simply can't tell you how hard it is.'

Whereupon Agafya Tikhonovna expounds her conception of beauty. 'If you could take Nikanor Ivanonich's lips and put them with Ivan Kuzmich's nose, and—and add in something of Baltazar Baltazarovich's free and easy manner and perhaps a little of Ivan Pavlovich's portliness,—then I would say "Yes!" right away.'

You see what a difficult position our would-be bride found herself in as a result of her varied requirements.

But if it is possible, mentally at least, to combine the nose of one man with the lips of another, it is quite unthinkable even to try to combine minimum cost with maximum quality in a single article. These attributes are incompatible: they are mutually exclusive. Does this mean, however, that if we are after quality, we cannot take cost into account at all? Or again: if we try to get an article as cheaply as possible, do we have to ignore its quality? Well—if we did, we would simply spend our lives collecting rubbish.

Obviously, we have to take everything into account—but in different degrees. If we are chiefly interested in quality, we have to specify the maximum cost we are prepared to accept in solving a particular problem. On the other hand, if we are out to buy as cheaply as possible, we generally have a pretty good idea of the minimum acceptable quality below which we would not take the article even if it was going for nothing.

Therefore, when we are formulating estimators of closeness to our ideal, we have to consider the means that are permissible for the achievement of our ends. The proud and cruel aphorism 'The end justifies the means' is something of a paradox, because there are many ends in life; and any means that is intended to achieve even the most cherished goal must not conflict with the attainment of others. Our means are always limited for this very reason. There is no such thing as 'any means whatever'. The means for achieving the most important, the most lofty aims must not come into conflict with other aims and principles that are, perhaps, not so lofty, but are nonetheless important.

The above aphorism should be rewritten thus: 'the end justifies the permissible means'. After being edited in this way, it loses its grand flavour and becomes a scientifically demonstrable truth.

But let us get back to self-adjustment.

SELF-ADJUSTMENT AS A FORM OF CONTROL

Suppose our automatic lathe is properly adjusted and producing screwed nuts of excellent quality. This means that the parameters of both the lathe and the blanks must be maintained strictly constant. Suppose, now, a batch of blanks arrives at the lathe, differing slightly from the usual blanks either in shape or in hardness. This will naturally have some effect on the quality of the finished parts and is, in fact, the commonest form of interference. It is unlikely that the lathe will start working better as a result. We expect instead that the quality of the output will go down and that the operator will have to readjust the lathe—

in other words, he will have to find such settings of the controls as will ensure a product of the highest quality. This means that the operator will be trying to minimize the differences between the ideal indicated on the blue-print and the actual components. And this is nothing other than *control.*

Drawing a general conclusion from the above example, we can say that in the process of control the operator eliminates the consequences of unforeseen changes that happen to occur in the system and returns the system to its least probable condition, which corresponds to the minimum value for the closeness-to-perfection estimator (or quality estimator).

For the time being we shall not concern ourselves with the details of how this is done. The important thing for the moment is to emphasize the idea that the system moved away from the desired condition and then after a short time it was brought back to this condition again by the action of a second system (the operator, after all, is also a system). Once we have grasped this idea, it is but a short step to an understanding of self-adjusting systems.

If we regard the lathe together with the human operator as constituting a single, albeit more complex, system, we can, without a shadow of a doubt, call such a system self-adjusting. The operator then represents the adjusting element in the more complex system.

The reader is possibly wondering, in view of the above, whether any machine plus operator constitutes a self-adjusting system. The answer is: yes, if the operator acts within the system to improve any of its properties; no, if in the course of his interaction with the machine the operator does not pursue the specific purpose of improving its operation.

The professional chauffeur together with his car does form a self-adjusting system. The chauffeur drives the car, and from time to time he adjusts and maintains it. The amateur driver, on the other hand, does not, as a rule, act as an adjusting element; so the system 'car-driver' is not in this case self-adjusting.

It is only natural to wonder whether there is really anything behind all this talk of self-adjusting systems, or whether it is just an empty play upon words, a mere sophistry. After all, if the whole business hinges on the presence of an operator, then the resulting 'automatic' is not worth a bean.

We hasten to point out that the study of self-adjusting systems that incorporate a human operator is of profound significance because it leads to an understanding of the specific features of the human adjuster's behaviour. Once we achieve this understanding, we are in a position to try to build an automatic machine that replaces—and frees—the human operator.

If we are to build such a machine, we have to have a clear idea of what the machine needs to 'know', what 'skills' it should have, and what it needs to 'remember' while it is working. To find out these things we have to study in depth the functions and the procedures of a human operator under similar conditions. While we are studying him, we are also building up a 'replacement' for him—in other words, we are compiling an operating plan for the future automatic machine that will replace him.

Incidentally, the activities of a fitter-operator and his procedures for adjusting the complex system that an automatic lathe represents remain one of the most important and least studied problems in the whole field of control.

At this point we shall try to define what we mean by a self-adjusting system.

We shall call a *system self-adjusting if it tends,* *independently and without external intervention, towards its ideal state, that is, if the system's quality estimator is maintained at a minimum level by the system itself independently of any effects imposed upon it by its environment.*

The block diagram for the interaction of such a self-adjusting system with its environment is shown in Figure 72.

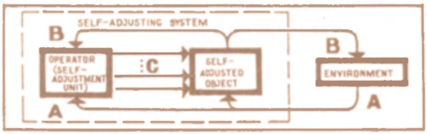


Figure 72

Here the self-adjustment block is a human operator who checks the quality of the lathe's output via channel B' and at the same time watches channel A' for any change in the properties of the blanks being fed to the lathe. If the blanks change their properties, the operator readjusts the lathe by determining new values for its control parameters. In order to be able to do this he has to know the correct procedure to apply in each particular case, that is, for any particular deviation of the properties of the blanks from standard. His actions must be strictly predetermined by the nature of the work being done on the machine. Either he has to know everything there is to know about the particular manufacturing process, if he is to decide for himself how the parameters are to be changed; or he has to have a complete set of instructions that tell him what to do for every imaginable deviation from standard that the blanks may exhibit.

Anyone who knows anything about automatic lathes knows that a set of instructions that took care of every contingency would tend to look like the Encyclopedia Britannica. On the other hand, a complete understanding of the manufacturing process would require of the operator the ability to deal with the most detailed problems of manufacture in general, as well as a profound knowledge of the process he is in charge of in particular. Naturally, such a state of affairs would hardly be to anybody's liking, and, more importantly, it would constitute an intolerable burden to the operator himself. After all, the task of trying to find one's way about a multi-volume instruction manual would indeed be a most unpleasant one. If we add to this the fact that components made by automatic lathes are usually produced in batch lots according to planned average requirements—in other words, lathes are assigned to produce a different component at roughly monthly intervals, it becomes pretty clear that for an operator who has to cope with a voluminous new instruction manual every month, and for the engineer that has to write such a manual every month, life would soon become utterly unbearable.

The difficulties experienced by operator and engineer alike in such a case would arise because the method that they chose for adjusting the lathe was not the best method, the so-called 'method of compensation'. With this method it is virtually impossible for the operator to make use of the feedback channel B' that would otherwise permit him to judge the quality of the product. A method such as this raises difficulties not only for the operator, but also for the engineer.

Indeed the sole purpose of channel B' is to check whether the lathe is turning out a good or a bad product. When a defective component appears, it is a signal to the operator to pay attention to the input end of the system, that is, to the blanks because, all things being equal, a deviation of the blanks from standard is the most likely reason for defective components to appear at the output.

But suppose 'all things' are not 'equal' at all. Suppose one of the parameters of the lathe changes for some internal reason at the same time as the blanks change—one of the tools becomes loose in its holder, say. Since the instruction manual was compiled to deal only with changes in the incoming blanks, it will not be of the slightest help in deciding what to do about this new eventuality, for all its great bulk.

The only thing the operator can do is to study closely the causes of the defects and try to eliminate them by readjusting the lathe. And to do this he has to keep his eyes glued to channel B' because this channel carries virtually all the information available as to the quality of the lathe's operation.

By watching the feedback channel B' to keep a check on the quality of the output an operator can regulate a machine even if he is completely new to the kind of work it is doing.

... AND HOW TO OVERCOME THEM

Suppose we have an operator who possesses the absolute minimum of knowledge required to operate a lathe, but is blessed with plenty of common sense. He needs a certain amount of knowledge if he is to get the lathe to produce the required type of component. For example, if nuts are required, he must know how to 'compel' the lathe to make nuts, and not bolts. (He has to have a minimum qualification to be able to do this.) He needs common sense if he is to adjust the lathe whenever it proves necessary in order to keep to a minimum the number of defective parts it produces.

Suppose, now, that the lathe is running out of adjustment and producing nuts having certain dimensions not agreeing with those specified in the drawing—enneagonal (nine-sided) nuts, say, instead of the desired hexagonal ones. The operator has no prior knowledge of exactly what he should do to correct the error, but he knows quite well which of the lathe's controls he can turn to change some of the dimensions of the nuts though he does not know which controls correspond to which dimensions. Obviously, the first thing he would do would be to try to establish how the dimensions of the outgoing nuts were affected by changing the parameters of the lathe. By turning one control (changing one parameter) and machining one nut the operator can discover what effect that particular control has on the finished product. In the course of this analysis he may encounter three types of controls.

THREE TYPES OF CONTROLS

Type one controls are those that affect only one dimension of the nut irrespective of the positions of the other controls. Turning a type one control in one direction, say to the right, results in an increase in the dimension in question; turning it in the opposite direction results in a decrease.

The operator will probably be overjoyed to discover controls of this type because they are so easy to use. For example, if a certain dimension suddenly increases, all he has to do is to turn the appropriate control in the appropriate direction by however much is necessary to eliminate the error.

An example of a control of this type is shown in Figure 73. Here the position of the tool in relation

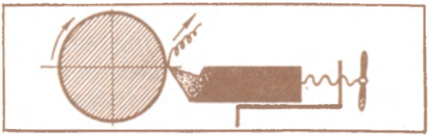


Figure 73

to the revolving work-piece and, consequently, the diameter of the finished part both depend on the position of the handwheel controlling the tool. When the handwheel is screwed in, the tool is forced towards the work-piece and the diameter of the part will be decreased; when it is screwed out, the diameter of the part will be increased. Suppose one complete turn of the wheel moves the tool through one millimetre. If, then, the operator discovers that the diameter controlled by this particular handwheel is 0.1 millimetre larger than it ought to be, he only has to turn the wheel through one twentieth of a full turn in a clockwise direction to decrease the diameter of the part by 0.1 millimetre and thus correct it.

It should be noted that the operator can only do this with confidence after he has carried out the appropriate experiments. These experiments consist of turning the wheel to right and left and observing the resulting changes in the dimensions of the parts issuing from the lathe. This done, the operator is then fully equipped to deal with any change that occurs in a dimension controlled by a type one control. After measuring a finished part he will immediately be able to set the controls so that these dimensions will be correct in the next part.

Let us draw one small conclusion at this point. For a dimension controlled by a type one control any deviation from the ideal contains complete information as to how far and in which direction the control has to be turned in order to reduce the deviation to zero. Controlling the dimensions of a finished part by means of type one controls presents no real difficulty and is called *deviational control.*

Type two controls will rather puzzle the operator to begin with. A single turn of one of these controls will immediately result in several dimensions of the nut changing simultaneously. An example of an arrangement containing a pair of type two controls (handwheels) that determine the position of a rectangular work-piece is shown in Figure 74. Here the work-

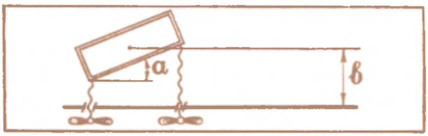


Figure 74

piece has to be placed in a definite position in relation to the lathe: the angle of inclination *a* and the centre-distance *b* have to have particular values. It is quite obvious that if either of these controls is moved, both these parameters will change. It is also clear that to adjust either of the two parameters *a* and *b* to their correct values both controls have to be moved simultaneously. To change the setting *b* both handwheels have to be turned in the same direction and by the same amount; the angle *a* will then remain constant while this is done. To change the angle *a* the handwheels have to be turned through the same angles, but in opposite directions; the centre-distance b will then remain unchanged in the process.

From this example we can appreciate that to alter one dimension only of the finished article we have to move several type two controls in a particular manner simultaneously.

It is easy to see that once again any deviation from the required dimensions of the product contains complete information as to how the settings of type two controls should be changed in order to reduce the deviation to zero. The relationship in this case is a little more complicated, to be sure: but there is nothing to prevent the operator from establishing it through experimentation. And once he has established it, he will have no difficulty in determining such settings of the type two controls as will guarantee correct dimensions on the finished component.

Obviously, type one and type two controls permit very effective control of the dimensions of the finished component. Any deviation from the required dimensions can be immediately corrected by a single measurement; that is, a single observation of the deviation is sufficient to determine exactly what has to be done to reduce the deviation to zero.

We can draw a further conclusion: type one and type two controls hold no fear for the operator. Once he has established the purpose of each of them, he can rest assured that whatever happens he will be able to correct any defect in the finished components as soon as it appears—provided, of course, that the defect involves a dimension controlled by these particular types of controls.

And finally, there is a group of controls that will leave the operator completely nonplussed when he first encounters them in his analysis of the lathe.

Type three controls (the ones we are talking about) are more complicated than either of the first two types. Each of them controls several dimensions of the finished part simultaneously; but the actual changes they produce in the dimensions are different for different positions of the control. In one position, for example, turning the control to the right will increase a given dimension, whereas in another position exactly the same movement of the control will result in a decrease in the same dimension. An example of a type three control is shown in Figure 75.

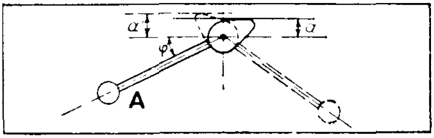


Figure 75

Here the dimension a depends on the angle φ at which the cam operated by the lever A is set. Obviously, when the lever is on the left and is moved counterclockwise, the dimension a increases; but a similar counterclockwise movement with the lever on the right will cause the dimension a to decrease.

In this example we can see that there is one position of the control lever where a slight movement produces no change in the dimension a; namely, when the cam angle φ is equal to 90°. This angle represents a critical setting of the control and, in this example, is independent of the settings of other controls. This, however, is by no means always the case. Very often the critical setting of a particular control depends on the settings of other similar controls.

Another typical example of a type three control is the tuning knob on a radio set. Suppose the station we are listening to 'drifts away' and we want to tune the set to bring the station back again. Which way should we turn the knob? We have no way of knowing until we have done some experimentation because the tuning knob is a type three control and therefore requires systematic investigation on every occasion. We may have to turn the knob in a clockwise direction on one occasion and in a counterclockwise direction on another, depending entirely upon which way the station 'drifted'.

Obviously, type three controls possess the treacherous property of changing the degree to which they influence the output of the controlled system. Whereas with type one and type two controls a single initial examination suffices to determine their effects on the system, type three controls have to be closely watched all the time. When we are adjusting a system, we have to be constantly on the alert for type three controls that may have reversed their effect on the system; and each time one of them does, we have to proceed with strict regard for its changed characteristics.

Consequently, whenever we are confronted with type three controls, we should observe the following rule: always turn the controls through a small angle so as not to miss the point at which any one of them passes through its critical setting.

Working with type three controls is reminiscent of fighting a war, the only difference being that we are not pitted against a human enemy, but against nature. Every battle is preceded by a reconnaissance the purpose of which is to collect information about the strength and disposition of the enemy forces. This information then forms the basis for the operational plan. At the end of the operation the situation changes: the enemy reorganizes his defences, so that further reconnaissance has to be carried out before the next operation; and so on.

A machine-tool operator 'wages war' with type three controls in exactly the same way. Before he readjusts the controls, he has to carry out a 'reconnaissance' in order to find out which way he should turn them to achieve the desired effect. When the lathe gets out of adjustment again, he has to carry out another 'reconnaissance' before he can readjust it; and so it goes on.

The general plan of 'reconnaissance-battle-reconnaissance' applies in this situation as well, but with the terms changed to 'experiment-adjustment-experiment'. 'Experiment' in this context means such manipulations of the object as are necessary to obtaining the information needed for controlling the object.

8. SEARCH (PATHS AND WANDERINGS)

So far, we have seen how an operator first analyses the operation of type three controls and then adjusts the system by means of them. Clearly, this kind of control has a dual nature, and because of this it is termed *search*.

Here the term 'search' refers to an active process of collecting information. It signifies more than mere observation of the object: it involves performing experiments with the object in order to find out how it will behave in the various circumstances that may arise during the process of adjustment. If the object does not change its behaviour, the investigation need only be carried out once and will be good for all time, as is the case with type one and type two controls.

We remind the reader that the operator will continue to experiment with these first two types of controls only until such time as he establishes their type.

Once he has done this, no further experiments are necessary because it will be quite clear which way and by how much any particular control has to be turned on any future occasion.

Type three controls are a different matter altogether because they change their characteristics all the time. If one of them increased the dimension of the part when you turned it in a particular direction yesterday, today it will decrease the dimension. And vice versa.

Type three controls necessitate a constant search because whenever the actual dimension of the component diverges from the ideal there is no way of telling which way they should be turned or how far. The search consists of making a number of trials to begin with—and only then does it become possible to adjust the system. So type three controls represent the chief difficulty that the operator controlling a lathe has to face. Consequently, the principal ingredient of his work is a process of search.

How is he to organize this search, then?

Having given the problem a little thought, an operator may choose one of the following methods.

METHOD No. 1

The operator makes a slight adjustment to the first of the type three controls, machines one nut with the control in this setting and then compares it with the previous one. For this purpose he applies a quality estimator such as the one we discussed earlier. If the value of the estimator is smaller for the new nut than it was for the previous one (meaning that the new nut is better), the operator continues to move the control in the same direction. On the other hand, if the quality estimator increases and the nut gets worse, the control must be moved in the opposite direction. So, by gradually changing the setting of this particular control, the operator works towards the best nut possible. This will still be a long way from the ideal; but it will be better than the first, quite useless, nut, and it will be the best nut he can obtain by adjusting only the first of the lathe's type three controls without touching any of the others.

The operator then moves on to the second control. He proceeds as before to find a setting for this control that produces the best nuts. Having found it, he goes on to the third control; then to the fourth; and so on.

When the operator has completed this process for all the lathe's controls, it may happen that the resulting nut, although greatly improved, is still not close enough to the desired dimensions. In that case he has to go back to the first control and find its best setting all over again, because when he changed the settings of the other controls the original 'best setting' of the first may no longer be the best at all (remember, we are talking about type three controls, which interact with one another). Then he does the same with the second control, then with the third, and so on. And he may have to go through the whole process several times before he can get the nut within the required dimensions.

Could it be that this is an endless process?

No; in fact it never is, the reason being that with every new setting of the controls the operator always improves the product and never lets it get worse. So there must come a point when the lathe starts producing nuts of the required dimensions.

All the same, this procedure for setting the controls on a machine (known as the *method of sequential parameter change* or the *Gauss-Seidel method*), though it yields the required result, is far too laborious. So the operator may choose another method for adjusting his lathe.

METHOD No. 2

The second method requires the operator to spend a certain amount of time analysing the effects of the controls before he proceeds to adjust—or perhaps we should say 'to tune'—the system.

He would be quite right to reason as follows. Each control affects the quality estimator to a different extent; so it would probably save time to take those controls that have a greater effect on the quality estimator and turn them through a greater angle, and those that have a lesser effect through a lesser angle. He concludes, therefore, that he should move the controls in proportion to their effect on the quality estimator. If one of them has twice as much effect on the estimator as another, it should be turned twice as far as the other. Let us consider a concrete example.

Suppose there are three type three controls that the operator has to adjust in order to minimize the quality estimator of the finished component. Before he can apply the above proportionality principle, he must find out the extent to which each of the controls affects the quality estimator. This he does as follows. He turns one of them in a particular direction, say clockwise, through a certain angle, say ten degrees. Then he machines one nut on the lathe and determines the value of its quality estimator. Knowing the value of the quality estimator for the previous nut, he can determine the difference in quality between the two nuts. Suppose these values were: 20 for the previous nut and 22 after the trial adjustment. The change in the value of the estimator is thus +2, which tells him the effect of the first control. He now knows that if he had turned the control through ten degrees in the opposite direction (starting from its initial setting) he would have obtained the value 18 for the estimator (a change of —2); in other words, he would have decreased the value of the quality estimator and improved the quality of the nut. But he should not actually carry out this adjustment until he has worked out the effects of the other two controls.

So he returns the first control to its original setting and carries out a similar trial with the second control, and then does the same thing with the third control, and thus discovers the effect of each. The results of these experiments are tabulated below.

In the table a plus sign signifies turning a control in a clockwise direction, a minus sign in a counter-clockwise direction. And this completes the analysis phase.

The operator now carries out a first adjustment. He changes the settings of the controls by amounts proportional to their effects on the quality estimator, that is, proportional to the changes he observed in the estimator during the experiments he has just been doing. The amount of each adjustment is equal to the change in the quality estimator multiplied by a constant coefficient (in our example the value —5 has been chosen for this coefficient).

The last column in the table shows the angles through which the controls are to be turned to effect the adjustment. As you can see, they are proportional to the results of the tests.

We observe that the choice of the proportionality constant in this case was arbitrary. However, it is usually chosen so as to reduce the value of the quality estimator by as much as possible. Naturally, if the constant is made too large or too small, the result

	Original settings	Analysis			First adjustment
		1st test	2nd test	3rd test	
1st control	0°	+10°	0°	0°	—10°
2nd control	0°	0°	+10°	0°	+10°
3rd control	0°	0°	0°	+10°	—5°
Value of quality estimator	20	22	18	21	15
Change in quality estimator	—	+2	—2	+1	—

in either case will be a poor adjustment. The constant has to be specially chosen on each specific occasion.

At this stage the operator has to determine the effect of each of the controls once again because they are now certain to have changed. This done, he makes a second adjustment of all three controls simultaneously in accordance with their new properties. And so he continues this procedure until he has reduced the quality estimator to the necessary minimum.

This method of adjustment is called the *method of proportional parameter change* or the '*gradient method*'. It is distinguished by particular accuracy, and in many cases it is superior to the Gauss-Seidel method.

It may happen, however, that even this method is not particularly suitable because of the large number of times the properties of the controls have to be determined. At each stage of the adjustment process the operator has to perform as many experiments as there are type three controls. What if there are a hundred of these controls? Or a thousand? What then? We must conclude that the gradient method is fine for adjusting simple systems in which the number of controls is small—two or three, say; but if there are a great many controls we shall have to look for another method of adjustment. And this is where chance comes to the rescue again.

METHOD No. 3
(RANDOM SEARCH)

The third method demands of the operator not only common sense, but also the courage of his convictions because he will be required to perform actions that will seem at first extremely odd and unreasonable. To make use of this method the operator has to turn all the controls at once through a small angle in random directions (we remind the reader once again that we are talking about type three controls). How does he ensure that he chooses random directions? He can do this by tossing a coin for each control and turning the control to the right, say, for a head and to the left for a tail. He also needs to make a note of all the changes.

In performing this unproductive, 'subversive' operation the operator relies on the remarkable property inherent in chance of encompassing every possibility. And among all the possibilities are those that will improve the product—and they will constitute a sizable proportion of the total. The procedure thus consists of making a series of random overall adjustments. If a particular set of adjustments results in a deterioration in the quality of the finished product, each control must be returned immediately to its former setting. Another set of random adjustments is then made again by turning each control a small amount in a new random direction.

At first sight it seems as though the operator's behaviour is utterly senseless. Whereas the gradient method (Method No. 2) was certain to improve the process, the present method not only does not guarantee any improvement, but may actually make the situation worse. And how can we be sure that the operator will not have to spend an excessive amount of time juggling the controls? Is there, in fact, any end to this process?

All these fears and doubts arise because the random method of adjusting a lathe is an unusual one. Yet upon closer examination its great advantages over non-random (regular) methods become readily apparent.

Indeed, a set of random adjustments may make the finished part either better or worse; and either result may be expected with equal probability. This means that on the average every second turn of a control will improve the product. As is shown both by theoretical calculations and by practical experiments, the time required for adjustment by the random method represents a considerable saving. For example, with the method of random search a system incorporating one hundred controls can be adjusted on the average in *one tenth* of the time needed for the gradient method. Each time the operator makes an unsuccessful random adjustment he immediately returns the system to its former condition and performs another random adjustment. His adjustments will not all be equally successful: there will be *bad* ones that lower the quality of the product (these he cancels out immediately); and there will be *good* ones. Among the latter there will always be those that are *just good*, which hardly improve the product at all, and those that are *very good* and result in an immediate, significant improvement in product quality. In such cases the chosen random directions in which the controls were moved happened by chance to have been the 'correct' directions for nearly all of them; which means that the optimum settings corresponding to the highest possible quality should be sought in just these directions. However, these 'very good' random adjustments will occur extremely rarely, and the superiority of the random search method does not depend on them. Its strength lies in the 'just good' adjustments because they occur much more frequently and are consequently easy to come upon by chance.

A game involving random search

The well-known children's game of 'hot or cold' provides a good illustration of the method of adjustment we have just been describing. No doubt the reader used to play this simple game as a child. Its rules are clear and straightforward: the person who is 'it' has to find some object that the others have hidden in the room. Whenever 'it' moves away from wherever the object is, the others cry out 'Cold!'; if he moves more or less towards the thing he is urged on by shouts of 'Warm!'; and when he heads straight towards the hiding-place, he does so to the accompaniment of shrieks of 'Hot!'.

Let us analyse the player's procedure. The first thing he does is to take a step in any direction at random. If he hears the negative 'cold', he tries taking a step in another random direction. He keeps on at this until he gets the affirmative 'warm'. Thereupon he concentrates his search in the successful direction. When he hears 'hot', he moves forward with confidence.

It is easy to see that the player behaves in exactly the same way as a self-adjusting system. The signals 'cold', 'warm' and 'hot' warn him of changes in his 'quality estimator'—his closeness to the concealed object. He chooses the method of random search for the obvious reason that he does not know of any other method of search. This lack of knowledge serves him well, in fact, because any other method of search would complicate his task and draw the game out interminably, so that it would simply become a bore.

Passions engendered by random search

Let us continue our discussion of self-adjustment by the random method.

Ten years ago, when this approach to the problems of self-adjustment was put forward by a group of people (this writer among them), it was met with *anything but indifference.* Everyone that came in contact with random search in any way gave vent to his feelings on the subject in a spirit of frenzied passion. Some—at first, in fact, most—openly ridiculed the idea and made random behaviour a vehicle for the extensive exercise of their wit. Others sprang just as openly to the defence of random search, seeing in it distinct possibilities for overcoming the 'dimensional curse' that plagues complex systems. (This curse threatens all who undertake to adjust any very complex system having a large number of controls. At present nobody knows the answer to problems of this order.)

The controversy over random search gradually lost its intensity, for it became clear that in certain complex situations involving large numbers of controls, random search was the only viable method for solving the problem. It was also granted that if the number of controls was small and the system straightforward, then it was better to use one of the regular methods of search (Methods No. 1 and No. 2).

It must be said, however, that even today there are people who cannot get used to the idea that in certain difficult cases the random method is both quicker and more reliable. On one occasion, after a stormy debate over random search at one of our regular scientific conferences, a friend of mine implored me to confess that it was all nonsense.

'Look,' he said, 'I know you need the subject for your thesis—and I'm sure that it will be accepted; but tell me, in all honesty: random or non-random search—doesn't it all come to much the same thing in the end? And mightn't the gradient method really turn out to be better after all? Come on, admit it!'

I admitted nothing of the sort.

On another occasion an eminent theoretician tried to talk me out of it, using the full weight of his authority: 'Young man, why do you waste your time with random search? I took it up when I was your age too, and was able to demonstrate that random behaviour is always inferior to regular behaviour, on balance, and that in particular, random search is inferior to regular search. I advise you to drop it.'

But drop it I did not.

Another time I chanced to hear a particularly zealous opponent of random search holding forth: 'I am engaged in the development of electronic devices of the greatest complexity. I frequently have occasion to use a search process when I am simulating the behaviour of a system on a high-speed computer. And now I find that my programmers are determining the optimum alternatives by using random search. Not that it matters to me what method they use—so long as they come up with the optimum mode for the machine—but surely this random method is against all logic? However much I tell them that random search is absolute nonsense, it never seems to get through to them. Once a programmer acquires a taste for the random search method—and one calculation is enough for that, I can tell you—wild horses couldn't stop him. But, for the life of me, I can't understand what they see in it.'

Random search and learning

If in applying the random search method the operator remembers the details of each step and instead of making each succeeding random choice completely random he takes into account the results of the preceding step, this method will yield still greater gains. The operator will be able to adjust and readjust his lathe in record time. Moreover, if he combines the random search method with self-instruction he will be in the best possible position for keeping his lathe in the required state of adjustment.

It is common knowledge that while a lathe is working both the lathe and its tools are subject to wear; and wear tends to raise the proportion of defective components. Consequently, the lathe is in constant need of adjustment to keep this proportion to a minimum. This is usually done as follows. Each time the operator adjusts the lathe he relies on the experience gained from the previous adjustments, because the lathe's loss of adjustment is evidently due to the same cause as before—tool wear, for example. Before long he is able to perform a complete adjustment in only one or two steps because he has taught himself how to correct for this particular kind of loss of adjustment, that is, he knows which controls to turn which way and how far in order to keep the lathe adjusted.

The process of self-instruction that takes place during the random search process is reminiscent of the techniques that are used in the training of animals. If the animal chances to do what its trainer wants it to do, he rewards it in the hope of reinforcing this particular chance act by associating it with a feeding reflex.

If the animal fails to do what is expected of it, it is punished with a view to encouraging it to behave otherwise in future; for amongst these other modes of behaviour will be the one that the trainer is waiting for.

We should bear in mind that punishment will only serve its intended purpose when the number of possible modes of behaviour is small. The animal will then have a chance of stumbling upon the desired action in a fairly short time in its efforts to avoid punishment. In a more complicated situation where the animal is presented with a large number of possibilities, punishment will prove more or less useless: it will hardly shorten the learning process at all, and will simply upset the animal. This is the theoretical basis for the advantages of rewarding an animal as opposed to punishing it.

Let us examine a simple learning experiment that uses laboratory rats and a T-shaped maze as shown in Figure 76. The rat is released into the maze and is faced with a choice of paths—one to the right and one to the left. The experimenter wishes to train the rat to turn to the right and encourages it to do so with a piece of bacon fat placed in the right-hand leg of the maze. At the same time he discourages it from turning to the left by giving it an electric shock whenever it does so. After a few trials the rat turns right and heads for the bacon fat without any hesitation. This means that it has learned and that its learning was assisted by both punishment and reward.

We could have employed punishment alone, by omitting the bacon fat in the right-hand leg of the

Figure 76

maze and simply giving the rat an electric shock if it turned left. The rat will still learn to turn right to avoid the punishment; but it will take longer to learn.

If we abolish the punishment and limit ourselves to reward, the rat will only learn to turn right after it has chanced to stray into the right-hand passage.

This example of teaching a rat in a maze shows quite plainly that it is possible to achieve a desired result by combining reward with punishment.

Self-instruction during random search proceeds in exactly the same way. Here, reward (tending to raise the probability of a successful adjustment) can be combined with punishment (tending to lower the probability of an unsuccessful adjustment) to achieve the desired effect of speeding up the search process.

So far we have confined our discussion to a fitter-operator, that is, to a human being who carries out the adjustments himself. Suppose, now, we wish to replace the human operator with an automatic adjustment system, in other words with a machine. Obviously, an operator using the random search method possesses one fundamental advantage: he is very easy to replace. This is because a computer programme for random search is extremely simple and can easily be incorporated in an automatic device.

The block diagram for such a device is shown in Figure 77. Here the system under adjustment has a definite number of control parameters (controls), which are set in motion by *random generators*. The output of the system passes to a converter that evaluates the quality estimator for the system and sends a signal corresponding to the value of the estimator along to the control block. (The signal will only reach zero if the system is in perfect adjustment.) The control block monitors this signal and operates the random generators, switching them on and off, according to the value of the estimator.

This arrangement operates in a very simple manner. The random generators change the settings of the con-

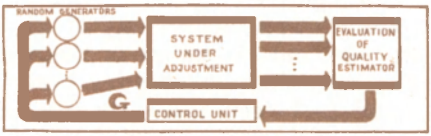


Figure 77

trols in random directions. If the system does not improve as a result of the last set of changes, that is, if the quality estimator does not decrease, the control block sends a command along channel *G* returning the controls to their previous settings. If, on the other hand, the estimator does decrease, the random generators transmit the next set of random adjustments to the system. And that is all there is to it.

To incorporate an element of learning into the process all we have to do is to send instructions to the random generators regarding their readjustment. This is done as follows. At the same time as the controls are returned to their previous settings following an unsuccessful adjustment, the control block sends another command along channel G; this command changes the characteristics of the random generators themselves so as to ensure that the situation that required the controls to be returned to their previous settings occurs as seldom as possible in future. This means that after the generators have been readjusted, this particular set of movements of the controls will occur much less frequently than before. The effect of this procedure is that the controls will be moved more frequently in those directions which lead to an improvement in the system.

We can describe this learning process in terms of the game 'hot or cold', as follows. The player remembers which of his moves resulted in cries of 'cold' and is therefore less likely to repeat any of them. In other words, he tries not to move in the 'cold' directions, which means that he will try other directions more often. And as he eliminates the various 'cold' directions one after another, he increases his chances of stumbling upon a 'warm', or even a 'hot', direction until he is bound to do so.

Obviously, the element of learning saves the player from senselessly repeating steps known to be 'cold' and at the same time guides him towards the 'warm' ones.

You may have noticed that the process we have just described is the same as 'learning from one's mistakes'. 'Punishment' is represented by the special instructions that reduce the random generators' probability of applying unwanted sets of random adjustments to the system's controls.

We can also employ a system of 'reward' by increasing the probabilities of those adjustments that improve the operation of the system, that is, that reduce the value of the quality estimator. In either case the system will succeed in learning how to adjust itself, so that it will take less time to correct its operation than it would without the element of learning.

As we have shown, a self-instructing system such as this constantly strives to improve the quality of its performance. At any moment something may happen to upset the smoothness of the system's operation and the system must then be ready to undertake a search for such new settings of its controls as will reduce the quality estimator to a minimum. But since the zero value of the estimator is, alas!, unattainable, the system does not 'know' how to account for its inability to achieve a zero value: it cannot tell whether it is simply impossible to get any closer to the ideal than it has, or whether chance interference is preventing it from operating perfectly. For this reason the system constantly endeavours to improve itself, ceaselessly testing various ways of changing its parameters, searching, searching, searching... One of the problems of random search is to determine the point at which the system can be considered sufficiently well adjusted and the search stopped.

Self-adjusting systems are currently finding a wide range of applications. After all, it is very convenient to have a system that adjusts itself and does not require the attention of a human operator. And yet convenience is not the most important consideration here. Systems like these are used in applications where a human operator would constitute a weak link in the system and would not be able to ensure the system's normal functioning because of his limited capacities. Sometimes it is absolutely essential to use a self-adjusting system, especially in a situation where conditions change too rapidly for a man to be able to keep track of them. Apart from this, adjusting a machine is not, for the most part, a particularly exciting occupation: the liberation of human beings from such tedious, monotonous work is a great and noble task.

In many cases, however, the problem of adjusting large systems, such as a complete production line for example, becomes very complex and requires the efforts of vast numbers of specialists to solve it. In such a case adjustment bears all the marks of a creative process, and in automating it we have to be able to simulate its creative aspects. And this we can only do if we understand what creativity is.

The problem of automating the adjustment of a system is thus connected with the problem of creativity. Its solution will therefore be a first step towards the automation of creative processes.

In concluding this chapter we should observe that the method we have described for automating the adjustment of a system has a definite limit. The system under adjustment tends towards a state of perfect operation (although in practice it almost never reaches it) and can therefore never become better than its ideal.

The last year or so has seen the appearance of a new type of system, the so-called self-organizing system, that does not have this limit of improvement and has much the same capacity as a living organism for improving its properties indefinitely.

But that is another story.

CONCLUSION

And so we come to the end of our journey around this thrice chancy world. And now, as we shake the accumulated dust from our feet, we can admit to ourselves that it was not the easiest going.

The first half of our journey was entirely devoted to overcoming the difficulties that chance puts in our way, hindering every kind of purposeful activity. This destructive tendency of chance is a manifestation of the second law of thermodynamics—the law that expresses the negative side to our world. We saw that the only reliable defence against the chaos resulting from chance is control; and cybernetics, the science that investigates the laws of control, is the science that deals with the struggle against chaos.

In the twenty-five years that cybernetics has been in existence it has developed effective methods for combatting chance, methods that are designed to crush and to annihilate chance obstacles on the road to knowledge. But this is not the only means at our disposal for dealing with chance. We have also developed methods for peaceful coexistence with chance, methods that enable us to work effectively despite the presence of chance interference.

During the second half of our journey around the world of chance, things took on a rosier hue. Here chance rose to the occasion in a new and, for chance, unusually positive role. We learned how man can use chance to advantage in his practical activities. We saw that the Monte Carlo method, having only the merest connection with the gaming houses of Monaco, is a powerful tool for solving a great many of our most important practical problems. We convinced ourselves that in play situations chance deserves a great deal of attention because it does not allow an opponent to proceed with certainty and thus reduces his chances of winning.

We became acquainted with the statistical hypothesis concerning the structure of the brain, a hypothesis that argues, boldly and strongly, for the idea that the structure of the nervous system is to a great extent a random one and that its rational behaviour is founded on the establishment of conditioned reflexes, these reflexes being formed through a process of instruction and self-instruction in which an element of chance once again plays an essential part.

We have analysed the operation of the perceptron—that remarkable machine that possesses the 'gift' of being able to recognize all kinds of visual shapes. The element of chance that is deliberately incorporated in its construction is responsible to a significant degree for this ability.

Chance is also of tremendous significance in living nature. The processes of evolution and improvement of living organisms in association with natural selection take place only because chance mutations produce in an organism accidental changes that are perpetuated in succeeding generations by heredity. We became acquainted with the first mechanical device to use the method of random search, namely the homeostat, and convinced ourselves that it simulated natural selection. The basic raw material for the abstract thought intensifier that we examined was interference noise; and the block diagram of the intensifier was virtually a copy of the process of artificial selection that man has long made use of.

Finally, we examined various ways of adjusting complex systems and saw that here, too, the method of random search possesses several advantages over regular methods.

The study of the remarkable world of chance is only just beginning. Science has as yet barely skimmed the surface of this world of strange happenings and limitless potential.

But the excavation of the priceless treasures of chance has begun, and there is no telling what riches it may yet uncover. One thing, however, is certain:

we shall have to get used to thinking of chance, not as an irritating obstacle, not as an 'inessential adjunct to phenomena' (as it is in the philosophical dictionary), but as a source of unlimited possibilities of which even the boldest imagination can have no prescience.

LEONARD RASTRIGIN graduated in aircraft design from the Moscow Aeronautical Institute and, in 1960, presented his Ph.D. thesis on mechanics. He then made a 179-degree turn and 'retreated' into cybernetics, where he studied random search — a new technique for finding optimum solutions to complex problems.

Cybernetics brings him both joy and sorrow. His work in this field has gained him his doctorate and a professorship, and he is now Director of the only random search laboratory in the world. Here his task is to vindicate the claims of random search and to demonstrate its advantages in practical applications.

Professor Rastrigin is a very busy man. Yet no sooner does he have a day off duty than he reaches for his pen. In the space of a few years he has written two monographs and over a hundred scientific articles. *This Chancy, Chancy, Chancy World* is his first book devoted to acquainting the general reader with his special field of study.