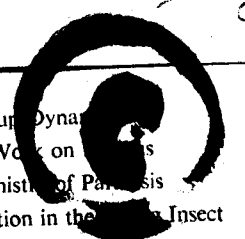
16



Group Dyna
Recent Wo on s
The Chemist of Pa sis
Fuel Consumption in the Insect
A Theory of Chess and Noughts and Crosses
Making Chemicals from Petroleum
On Looking at Old Instruments
Robots which Play Games
Human Colour Vision
Research Report

SCIENCE NEWS

EDITED BY J. L. CRAMMER

16

PENGUIN BOOKS

HARMONDSWORTH · MIDDLESEX

EDITORIAL

CULTURE is a word which has been tossed about a good deal in the last thirty years. For some it has called up visions of the long haired ineffective intellectual or the finicky don; chatter of Baudelaire among the teacups, talk of Goya over the port. For others it is part of the class war, the stolen heritage of the People. To thousands of careful scholarship students with their eyes on exams and the haven of the Good Job beyond, it is something one can get later, like a dinner jacket – if it is really necessary. And now, with the great expansion of scientific and technical education since 1945, come the anxious and bewildered voices crying 'Beware: Science is not a cultural subject. We are breeding a race of barbarians, what shall we do?'; and a few of them offer a study of alchemy or a short course in the history of Science as the panacea.

Pondering the matter we decided to penetrate boldly through the hubbub of approval and contempt to see if we could discover the real meaning of this battered and rather worn-out word. If it were really snobbish affectation, or even upper-class polish then it would be something extraneous and unworthy of effort to preserve it. But in fact, as any anthropologist or sociologist will tell you (see, for instance, page 101), culture is a word which summarises the whole pattern of social living. It is impossible to have none of it, just as it is impossible to do without breathing: one can merely have more or less. In our complex Western life, culture implies an awareness of all that is going on in the world and a self-awareness of oneself as a citizen. It is part knowledge and part humanity, that is to say a knowledge, understanding and feeling for one's fellows – which brings us to the reason why French and History are cultural studies, while Chemistry and Physics are not. It is quite simply that Science *as taught* deals only with things and abstract relations, while languages and

stories inevitably deal with people. The student who has been isolated from the education of real life and the unconscious cultural influence of fellow workers by seclusion first in school then in university requires something to be done for him if he is not to qualify from his laboratory as humanly innocent as a newborn babe – but unfortunately, twenty years older and fixed in his narrowness.

Culture is not just a matter of hobbies, or of preferring a picture gallery to a dog track (both have their points). To come down to brass tacks and hard cash, it is a matter of a man's personality and capabilities, of being able to write reasonable English and make a good speech, of knowing how to get along with superiors and colleagues and inferiors, of knowing his own limitations and prejudices.

For the improvement of scientific training the introduction of history is good if it is social history and shows the scientist his place in the community. The introduction of philosophy is good if it teaches the precise use of language, and a critical analysis of everyday assumptions. But there is also a third improvement possible: the teaching of Chemistry and Physics as applied science as well as pure science, human work as well as intellectual exercise, mass production and labour relations as well as thermodynamics. Chemistry would become a humane subject if alongside the courses on laboratory reactions went courses on chemical industry and its social significance.

The relevance of all this for *Science News* is to be seen in the following pages. Read for instance on page 124 how petroleum is taking part in a new technical revolution, with repercussions in employment and trade and politics. Or read the article on robots which shows how old is the idea of a chess playing machine. Or appreciate the beauty of mathematical method in *A Theory of Chess and Noughts and Crosses*. In this and following issues we are trying experiments in widening our scope, and we shall be grateful for your comment.

RECENT WORK ON MESONS

TERENCE PRICE

ONE of the characteristics of present-day nuclear physics is that the world no longer seems explicable, as it did to our grandfathers, in terms of a few fundamental units; if we are to make sense of it we have to believe in the existence of a much larger number of constituents of matter than had once been hoped. To call them all fundamental would be begging the question – it is better to say that, although the relationships between some of them are known, we do not yet know which of them are primary and which can possibly be products of reactions between the primary particles. Besides the protons and electrons that made up the classical atom, physics now recognises the positron (or positive electron), the neutron, possibly the neutrino, and a whole range of intermediate particles heavier than an electron but lighter than a proton, to which the general name meson is given.

Historically the discovery of mesons sprang partly from cosmic-ray research and partly from theory. An outline of their story up to 1947 is given in *Science News* 4. Since then, however, an intense effort in both research and theory has been made. In spite of new and at first confusing discoveries, the picture is beginning at last to seem clearer, though it still has its difficulties. We now know that there are at least four kinds of mesons; there are probably even more. We know something of their relationship to the structure of the atomic nucleus (which is the next, and perhaps the final hurdle for physics to take); we know a considerable amount about their life history; and we know something about how they are related to electrons and neutrons. Undoubtedly if we were to wait a little longer many of the difficulties now remaining would have been cleared up. But the work of these last few years is a fascinating story that is worth telling for its own sake.

It is told here from an experimentalist's viewpoint. A few landmarks in the work have been selected and dealt with in some detail, so that the reader can gain some insight into the way in which the information is hunted down by research workers the world over, all contributing to the task of making the jigsaw fit.

PHOTOGRAPHIC PLATE EXPERIMENTS

When a fast-moving charged particle passes through a photographic emulsion, ions are formed in the silver bromide grains. If the grain is to appear as a black spot after developing, a certain minimum number of ions must have been formed; a smaller number will show no effect in the finished plate. Now it happens that the kind of charged particles which make up the majority of the cosmic radiation have an ionising power which is only just sufficient to liberate the critical number of ions. If such a particle does not go straight through the middle of a silver grain, but only through the edge, no 'latent image' will be formed. The result is that cosmic ray particles – especially the fast-moving ones – leave a characteristic kind of dotted, grainy track behind them. The ionising power of a charged particle usually increases as it is slowed down by repeated collisions with the nuclei of the emulsion. This is shown up by the gradual crowding together of the exposed grains towards the end of the track.

When the speed of the particle is close to that of light, the ionising power has a minimum value, which is the same for all particles with equal charges. It is only recently that the sensitivity of photographic emulsions has been increased to a point at which very fast-moving singly charged particles can be recorded; these so-called 'electron sensitive' plates were not available when the experiments that we are going to consider were started. This did not greatly matter, however, since many of the important observations were made on particles which were not – by the standards of physics – travelling very fast.

When a cosmic-ray track in a photographic emulsion is being examined, two things are looked for: the departure from straightness (which is more marked for light and slow-moving particles) and the spacing of the exposed grains. From these two things taken together the particle which was responsible can usually be identified.

Many of the tracks are very short, and are visible only under a high-powered microscope; the experimental results can be collected only after many hours of tedious searching.

In the summer of 1947 a group of physicists at Bristol University – Lattes, Occhialini and Powell – were carrying out some experiments using photographic plates which had been exposed to the action of cosmic radiation at high altitudes. It had long been known that the cosmic ray intensity increased with height; the exposures for these experiments were therefore carried out at convenient mountain locations in the Andes, in the Pyrenees and in Switzerland. The method is quite simple: one takes a stack of photographic plates and leaves them in a light-tight box for a time which is decided by previous experience, then develops them and examines the tracks which have been left.

The sort of thing that was found is shown in Inset, Nos. 28 and 29, which are examples of the raw material of this kind of research. They are photomicrographs of the plates, and since the tracks extend over many microscope fields, they are built up in the form of a mosaic.

Inset, No. 29 shows a kind of event which is one of the important discoveries of the Bristol group. The track which comes in near the top must be from something going from left to right, since the bending becomes more obvious in this direction, showing that the particle was slowing down. It must be a meson, that is, with a mass somewhere around 300 times the electron mass, because of its characteristics (the scattering, etc.). This was to be expected, because it was already known from experiments described later in this article that a particle of about this weight existed.

The particle slows down, and then from the place where its track ends another particle appears. This has only a moderate energy and soon stops, showing as it does so the typical decrease of grain spacing. The range of these secondary particles in the

travelling through solid matter, they will be captured by nuclei before they have a chance to disintegrate into μ-mesons; instead, it is the nucleus that breaks up, forming a characteristic 'star' on a photographic plate. Negative μ-mesons, however, do not form stars when captured by nuclei.

(c) π-mesons are formed when cosmic-ray protons strike nuclei.

(d) Both π- and μ-mesons have masses about 200 – 300 times that of an electron, the π- being the heavier.

We have dealt with these experiments in considerable detail, first, because of their fundamental importance, and secondly, in order to become familiar with these newly-discovered pieces of matter. We can now glance at some of the other work which was going on at the same time.

THE ARTIFICIAL PRODUCTION OF MESONS

The nature of the forces that prevent the nucleus flying apart, under the combined repulsions of the positively charged and very closely packed protons within it, has been the central problem round which physics has been mounting its most recent attacks. It is scarcely possible to discover anything until you are able to affect the thing in which you are interested. When you can make it do something at will, then you are making a start on discovering exactly what it really is.

It was hoped that the behaviour of mesons would offer a clue, but the difficulty with meson research was, until recently, that the only source was cosmic radiation. This gave so few mesons that in order to collect sufficient results experiments often took weeks. These experiments were usually difficult to do well, and contradictory results in which no one had faith were commonplace. One of the important things about the photographic plate method was that the tracks were there for all the world to see and argue about. The other thing about it was that instead of having to draw tentative conclusions from a large number of rather inaccurate results, it was possible to sort out individual happenings; we could get down to detail.

In the early part of 1947 the monopoly of cosmic rays as the

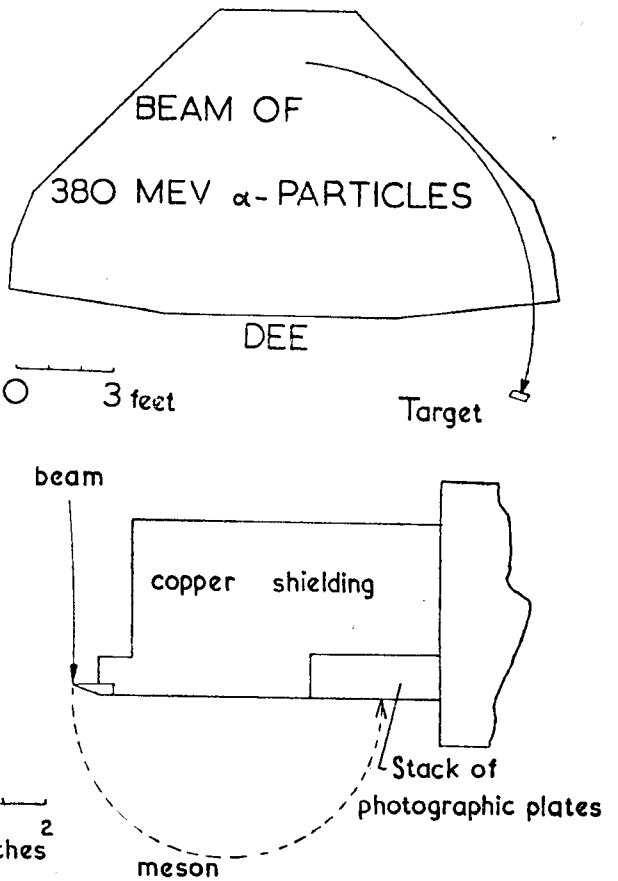


Fig. 1. Diagram of the experimental arrangement for studying the production of mesons in the Berkeley cyclotron

Top: Plan view of cyclotron showing target and accelerating electrode ('dee')

Bottom: Detail of target

Redrawn by courtesy of Eugene Gardner & C. M. G. Lattes, from *Science*, Vol. 107, p. 270 (12.3.48)

only source of mesons was lost, when the large synchro-cyclotron at Berkeley, California, was completed. The building of this instrument was itself a major technical feat. The pole pieces of the magnetic field are 184″ in diameter, and an 18 Kilowatt radio frequency drive is needed to provide the 15,000 volts peak which accelerates the alpha-particles at each revolution (alpha particles are fast-moving nuclei of helium).

Scrutiny of the photographic tracks of cosmic rays had already given an idea of the energy necessary to break up a nucleus. It was certain that with the new cyclotron it would be possible to produce nuclear disintegrations, and that there was a good chance that it might also be possible to produce mesons artificially. If so, an immensely more powerful source of them would be available; one would be able to collect as many 'tracks' in a few minutes as in months spent at high altitudes using cosmic rays.

Lattes went on from Bristol to join the Berkeley team, which was trying to see if a supply of mesons could really be turned on and off at will. The first successful results were announced in March 1948, about a year after the publication of the photographic plate experiments of the Bristol group.

Fig. 1 gives an idea of the arrangement that was used. The alpha-particles in the cyclotron start at the centre of the magnetic field, and as they are accelerated spiral outwards, in nearly circular paths of ever-increasing size. A thin slab of graphite is put in their track at the point where they reach their maximum energy.

Of course, the nuclei of atoms are very small, about 10^{-12} cm. radius, but the number of alpha-particles travelling round in the beam is large enough to provide a very substantial number of nuclear disintegrations rather similar to that shown in Inset, No. 28, in the graphite. It is from these that π-mesons are born. At Berkeley as they escape from the graphite plate they are moving under the influence of the magnetic field, and their path is therefore curved. Negatively charged mesons will, of course, be bent in the opposite direction from the positively charged alpha-particles; that is, away from the centre of the cyclotron. Being on the outside of the machine they are more easily accessible

and were studied first simply by placing photographic plates in their path.

These experiments have an important advantage over those originally performed at Bristol: it is possible to obtain a fairly accurate measurement of how heavy the mesons are. The two things that have to be known are the radius of the circular path the mesons follow in the magnetic field, and the distance they can penetrate the photographic plates. Both of these things depend on the mass and the speed of the meson, and, by knowing both, the mass can be found. The best result was that the mass of the negative meson was 286 $\pm$ 6 times the mass of an electron (m_e). This meson had all the properties of the negative π-mesons found in the cosmic ray experiments. The search for positive mesons was a little more difficult, because the apparatus had to be arranged in such a way that those alpha-particles which were not yet up to full speed were not intercepted. Positive π-mesons were found, with about the same mass as for the corresponding negative particles. There was also a group of lighter mesons – μ-mesons – with masses of about 200 m_e, into which the π-mesons had changed (or, as the physicist says, 'decayed') before striking the plates.

Thus there was no doubt that in the Berkeley cyclotron the world had an instrument capable of producing π- and μ-mesons for research, in numbers about one hundred million times as great as those available from cosmic rays. Subsequently, mesons have also been produced at Berkeley using the large electron synchrotron.

RADIOACTIVITY OF THE μ-MESON

Soon after the discovery of the meson, in the late nineteen thirties, a most interesting thing was noticed: passing mesons through matter reduces their numbers, and since the effect is easy to measure, it was one of the first things to be studied. The rather surprising result was found that different absorbers, which should have given the same result – according to predictions – in fact did not do so. If, for instance, one compared the effect of the air in a few thousand feet depth of the atmosphere with

that of an equivalent solid, the air seemed to be better at stopping mesons than it ought to be. There were a few attempts at explaining this effect away by assuming that the density of the material did have a direct effect, before the real explanation was hit upon. As is often the case with correct explanations it was delightfully simple. It was this: the meson (it was the μ-meson) was unstable; that is to say, if left alone it would spontaneously change into something else, and cease to exist as a meson. Now, during its passage downwards from high altitudes, it can be captured by the nuclei in the atmosphere, or it can decay spontaneously. In the laboratory experiments, however, the absorbers are very small; the mesons do not take very long to pass through them, and so there is not time for many disintegrations to occur. The result is that the atmosphere seems to be behaving as a better absorber than it should, and from the difference between how it actually behaves and how it ought to have behaved, we can find out how long the meson 'lives'.*

It was soon recognised that to explain this absorption anomaly the average lifetime of a μ-meson must be about a microsecond. A number of experimenters set out to obtain a more accurate measurement, one of the best results being that of Rossi and Nereson in 1943. The way in which these measurements were carried out is interesting because they are typical of the technique that has to be used – a technique that is worth a short digression.

The first basic tool is the Geiger Counter, which gives a pulse of electricity when a single charged particle like a meson passes through it. The other tool is the whole subject of electronics. Without electronics modern physics could not exist.

With a *coincidence unit* we can say whether the pulses from two Geiger counters occurred at the same time. More precisely, 'at the same time' usually means within about a millionth of a second, although with much more difficulty it can be made to mean within a ten-millionth of a second.

* We speak of the life-time of unstable particles without, of course, meaning that they are in any way connected with life. The lifetime is the average time they exist before breaking up. In breaking up they follow the exponential law common to all radioactive materials.

Pulses coming as close as this are recognised by the unit as occurring 'instantaneously'. Because they occur so close together they are very likely to be connected with each other. There is, of course, the possibility that two unconnected things happen by chance during this 'resolving time', but by making it as short as possible this danger can be made remote.

This type of unit makes possible a 'cosmic ray telescope' or stack of counters. We can arrange that only mesons going simultaneously in a particular direction – that is, through all the counters – are recorded. The rest are disregarded, and in this way we can select the particles in which we are most interested.

With a *delayed coincidence unit* we can find out whether a particle passing through a counter produces an effect elsewhere, not at the same time, but some few microseconds later.

The wartime development of radar gave an enormous impetus to electronics technique, and has resulted in a number of important experiments which would have been impossible without its help.

The experiment of Rossi and Nereson used the technique of delayed coincidences. What happened was roughly this: a Geiger counter telescope was placed on top of an absorber, so that when the counters were discharged the coincidence unit 'knew' that a meson had passed through, and was wandering around inside the absorber. Then, a few microseconds later, the meson disintegrated into an electron and something else. The electron shot out of the absorber and triggered off another Geiger counter placed at the side. The delayed coincidence unit could then say how long had elapsed between the two Geiger pulses. The average time that had passed is the average lifetime of the μ-meson. Rossi's value was 2.15 ± 0.07 microseconds.

We now know that the behaviour of μ-mesons varies according to whether they are positive or negative. The positive ones will have the 2.15 microsecond mean life, but the negative ones, although similarly radioactive, will often be absorbed by nuclei *before they have a chance to decay;* in this case there is no electron produced when the meson disappears. Rossi made no attempt to distinguish between the two sorts of meson, but was in fact

measuring what happened to the positive ones. At the time he carried out his experiment the existence of the π-meson was not suspected, and the μ-meson was thought to be the particle which Yukawa had predicted as responsible for nuclear forces. The first proof that this was not so arose out of an experiment by a group of physicists in Rome: Conversi, Pancini and Piccioni. They too, of course, did not have the benefit of the discovery of the π-meson.

The argument which lay behind their experiment went like this: if μ-mesons were really responsible for nuclear forces, then it could be shown on general grounds that, when negatively charged, they should be easily captured by nuclei. There would, in fact, be overwhelming odds against any negative meson in solid matter living long enough to decay and give rise to a decay electron. If one looked for these electrons, one ought not to find any. The positive mesons would be in a different category, because their electric charge would hold them away from nuclei, and so give them time to decay.

Conversi, Pancini and Piccioni studied the decay electrons from mesons absorbed in solids. They found to their surprise an appreciable number of electrons when they absorbed negative mesons in graphite, though there were none when the absorber was iron. The conclusion was that low atomic number nuclei do not have a very great tendency to capture (negative μ) mesons; the probability for this happening is so small that quite a number live long enough to show the alternative method of disappearances by radioactive decay. For higher atomic number nuclei like iron the increased electric attraction makes capture much more probable. However, the results for the case of the graphite absorber could mean only one thing – that these mesons were not the particles Yukawa had postulated.

Soon after this upsetting experiment had been announced, Lattes, Occhialini and Powell produced the π-meson, and theoreticians once more had something around which to spin their webs.

We said just now that a μ-meson disintegrates into an electron and something else. The electron is fairly certain: the third track

in Inset, No. 29, which we ignored earlier, is an example. This is actually the first photograph Powell obtained in which the chain of events $\pi \rightarrow \mu \rightarrow$ electron was clearly shown. To be quite fair, this photograph does not actually prove that the μ-meson gives an electron when it decays. From the characteristics of the track we can only be sure that the decay particle is not more than ten times as heavy as the electron. Hincks and Pontecorvo have whittled down the uncertainty by performing an experiment in which they investigate the gamma rays (Bremsstrahlung) produced when these fast charged particles are slowed down; they are fairly sure that the decay particle is less than twice as heavy as an electron, and almost certain that it is less than 3 m_e.

The electron carries away with it an appreciable quantity of momentum, even when the meson producing it is travelling very slowly. It therefore seems as though at least one other particle must be emitted, otherwise the law of conservation of momentum would be violated. Since all the charges are accounted for, the particle or particles must be uncharged.

Another law which has to be satisfied is the law of conservation of energy. This says, in effect, that similar transitions release the same amount of energy. If a constant amount is to be divided between two particles, only one method of division will satisfy these laws. The disintegration product will always come away with an energy which is typical of the disintegration. This happens in the $\pi - \mu$ decay. The π-meson disintegrates when at rest, and the μ-meson always starts with an energy of 4 million electron volts. From this we conclude that only one other disintegration product is involved.

In the 2.2 microsecond μ-meson decay, however, the decay electron has an energy which can vary over wide limits. This implies that there must be at least three disintegration products, since the conservation laws would then no longer demand a unique distribution of energy: provided the total remains constant, the share of energy taken by each of the products can vary.

These extra decay particles are believed to be neutrinos – one in the $\pi - \mu$ decay and two in the μ-meson – electron transition. The reasons for this assumption are that the particles are cer-

tainly uncharged, and that the zero rest mass, which the neutrino is assumed to possess, explains satisfactorily the observed decay energies. The existence of a neutrino is a convenient hypothesis, originally put forward to explain a continuous distribution in the energies of electrons given off by radioactive nuclei. Its properties are such that at the moment we cannot prove its existence directly, but the mass of indirect evidence is sufficient to make belief in its reality almost universal.

MASS OF THE μ-MESON

We have seen how a fairly good idea of meson masses can be obtained from the grain-spacing in photographic plate experiments; and how the mesons produced in the Berkeley cyclotron revealed their mass by the manner in which they reacted to the magnetic field.

Many measurements on the cosmic ray mesons have also been made using a similar kind of magnetic bending. The main difference between such experiments and the ones on artificial mesons is that the cosmic ray particles are moving faster, and are therefore much more difficult to deflect from their path. The best value obtained from the combined measurements of Fretter, Retallack and Brode for the mass of the μ-meson – which is the one commonly occurring at sea-level – is 215 ± 5 electron masses.

OTHER KINDS OF MESONS

It is a natural question whether these π- and μ-mesons are the only kinds of 'intermediate particles'. The first reply is purely philosophical and was given to the writer when he asked someone the same question. It is that particles are discovered when the physics techniques necessary for their discovery are ready. We have seen how, when the new tool provided by the sensitive photographic plate was used, the π-meson was found. The two sorts of mesons we know happen to have just those properties that make them detectable with the present tools of physics. We can argue that this is accidental, but a more likely train of thought is that when improved methods are available we shall find still more particles.

In fact, there is already a possibility that another meson – the τ-meson – of about 1,000 times the mass of the electron exists. Unfortunately, every manifestation of its presence seems to endow it with different habits; so that it is not certain whether this is a case of being misled by experimental errors or not. The French physicist Leprince-Ringuet has jokingly named it the 'Fregoli' meson after a famous Italian clown who was adept at impersonations.

During the last year the existence of a very unstable neutral π-meson seems to have been proved, but little further progress has been made with the reported discovery of light mesons.

MESONS AND NUCLEAR THEORY

At the beginning of the article it was stated that the approach would be from the point of view of the experimenter, and that theoretical matters would be touched on only lightly. There is in this an implied criticism of the present state of physics theory, which, according to Professor Pryce, is now sadly in need of a second Einstein to clear up the mess.

The history of the meson is peculiar. It began as an entirely theoretical idea of the Japanese physicist Yukawa, who, in 1935, was attempting to solve some difficulties that had arisen in the theory of nuclear structure. The nucleus, according to his ideas, should be held together by forces in the creation of which particles of mass intermediate between protons and electrons took part. When the μ-meson was discovered in 1937, it was thought that his theory had been brilliantly verified. This view persisted until 1946, when the crucial experiment of Conversi, Pancini and Piccioni showed that whatever the μ-meson was, it was certainly not the particle that Yukawa had predicted. While physics was pondering the difficulties, and preparing alternative theories, the π-meson was discovered. Since it reacts strongly with nuclei, it is in this respect similar to the Yukawa particle. This is not to say that all the difficulties have disappeared; but at least they are not as severe as they appeared to be in the early part of 1947.

No mention of meson theory, however brief, can be made

without some reference to the work of the theoreticians, who attempt the interpretation of the enormous quantity of confusing and sometimes misleading experimental results. Their names rarely reach the ears of the lay public, for whom their work is something of a mystery. At its lowest valuation it is very necessary correlation; at its best it is a pointer showing the experimentalist where next to break the ground.

The importance of the enormous quantity of work on mesons that is now going on does not lie in the direction of practical returns. No one can foresee that it will make any contribution to atomic energy techniques. It is a search for knowledge for its own sake. This is the traditional method of science, and the only one possible. It is the sphere of human activity where the touchstone of immediate utility is least valuable and most misleading.

APPENDIX

Here in a form convenient for quick reference is a summary of the principal facts known about mesons at the present time (December 1949).

π-mesons

(a) Produced when fast moving protons (constituents of the primary cosmic radiation entering the atmosphere from outside) strike nuclei.

(b) It bears strong, but not complete, resemblances to the particle predicted by Yukawa.

(c) The mass of the π-meson is 286 $\pm$ 6 electron masses.

(d) π decays, with a mean life-time of 0.9×10^{-8} seconds, with the emission of a μ-meson and a neutrino; but in the presence of an absorber this process is only observed for positive π's.

(e) Negative π's are captured by nuclei with the production of stars (see page 13).

(f) Both π^+ and π^- can be produced artificially in the Berkeley cyclotron.

(g) Very unstable neutral π-mesons, with a lifetime less than 10^{-11} of a second, are now thought to exist. (Neutral mesons are sometimes called 'neutrettos'.)

μ-mesons

(a) Are the principal constituents of the penetrating component of cosmic radiation at sea-level.

(b) Decay product of π-meson. Mass 215 electron masses. Itself unstable, disintegrating with mean life-time of 2.15 microsecond, probably forming an electron and two neutrinos.

(c) Negative μ's are captured by all except the lightest nuclei, and subsequently neutrons are emitted from the resultant compound nucleus.

Other mesons

Heavy (τ) mesons of about 700-1,000 electron masses have been reported by a number of experimenters, though many of their claims have since been withdrawn. The existence of such mesons is still very doubtful.

FUEL CONSUMPTION IN THE FLYING INSECT

DR V. B. WIGGLESWORTH, F.R.S.

THE flight of insects presents many puzzles in physiology and physics. Leaving aside the intricate problems of aerodynamics, the rate of movement of the wings is itself truly remarkable. In the most efficient fliers the wings are moved by 'indirect muscles', the two main groups of which run respectively longitudinally and obliquely across the thorax. These groups of muscles contract alternately, and in so doing they deform the wall of the thorax and thus raise and depress the wings that are hinged to its two sides. By the use of the stroboscopic flash lamp (Chadwick and Gilmour, 1940), by listening to the musical note that is given out by the beating wings (Sotavalta, 1947), and by other means it has been shown that in the bee there are about 200 complete beats per second; in the mosquito 500-600; and in the tiny midge *Forcipomyia* about 1,000.

In the muscles of most animals each act of contraction and relaxation is brought about in response to nerve impulses. But no neuromuscular mechanism is known which can respond with a rapidity approaching that seen in the indirect flight muscles of insects. Moreover these flight muscles are very poorly supplied with nerves.

The answer to this problem has recently been found by J. W. S. Pringle (1949) of the Department of Zoology in Cambridge. When the longitudinal set of muscles contracts, the oblique muscles are slightly stretched, and vice versa. It appears that the chief function of the nerves is to maintain the muscles in a reactive state. In this state the one set of muscles is stimulated to contract instantly and as a whole in response to the sudden increase in tension set up within its fibres by the contraction of the antagonistic group. In this way a very rapid alternate oscillation is set up, the frequency of which is determined largely by the physical load and inertia offered by the beating wings.

These high rates of muscular activity naturally demand a high rate of energy consumption. In a man going from rest into the fastest sprint the consumption of oxygen is increased about 10 or 12 times. When the bee goes over from running to flying its oxygen uptake rises 50 times or more. In the fruit fly *Drosophila* it is increased about 100 fold.

Calculations of this kind at once raise the question of the nature of the fuel that is burned, the amount that can be carried, and so the duration and the range of flight. The honey bee seems to be dependent on the store of sugar in its stomach. Sugar is present in the blood in the form of glucose at the surprisingly high concentration of 2 per cent. The bee weighing 100 mg. uses this sugar at the rate of about 10 mg. per hour and has a flying time of only about 15 minutes with a flight range of about 4 miles. Some insects can utilize fat. The fat content of the beet leafhopper *Eutettix* may be used as a measure of the distance over which it has flown from its breeding places. Locusts lay down large stores of fat before making their migratory flights; and it has recently been shown by the late Professor A. Krogh and Dr T. Weis-Fogh, working in Denmark on behalf of the Anti-locust Research Committee of the British Colonial Office, that during sustained flight in the desert locust *Schistocerca*, the respiratory quotient (the ratio of carbon dioxide given out to the oxygen consumed) equals 0.74–0.75, indicating that fat is being burned.

Curiously enough it is in the little fruit fly *Drosophila* that we know most about these matters – in spite of the fact that the commonest species *D. melanogaster* is only about 2 mm. long and weighs little more than half a milligram. Dr L. E. Chadwick and his colleagues, working at Harvard and at the U.S. Army Chemical Centre, Maryland, related the rate of oxygen consumption in *Drosophila* during flight with the rate of wing beat as measured with a stroboscopic lamp. They found that the respiratory quotient was 1.0 and so concluded that glycogen, a carbohydrate (animal starch), is the only reserve substance used

by the flying insect (Chadwick, 1947; Chadwick and Gilmour, 1940).

It was interesting to obtain an anatomical picture of the location of the reserves of fuel in the fruit fly and of the changes that take place in these reserves during exertion (Wigglesworth, 1949). There are simple methods of staining by means of which fat and glycogen are easily demonstrated. The fly, suitably fixed, is simply slit in half with a fragment of a razor blade and the two halves are examined with the microscope after the appropriate staining. Fig. 2 shows the distribution of glycogen.

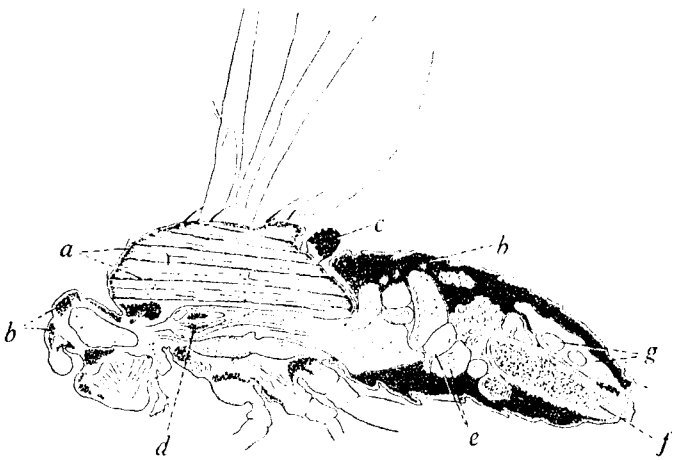


Fig. 2. A section through the fruit fly *Drosophila* showing, in black, the distribution of the reserves of glycogen. Head on the left, abdomen on the right, thorax with wings above and legs below in the middle. (*a*) glycogen in the flight muscles; (*b*) in the fat body; (*c*) in the knob of the balancer or haltere; (*d*) in the proventriculus; (*e*) in the wall of the intestine; (*f*) in a developing egg; (*g*) rectal glands containing no glycogen

There are deposits along the flight muscle (*a*), massive deposits in the so-called 'fat body' (*b*) in the abdomen and head, etc.; there is a dense deposit in the knob of the haltere or balancer

(*c*), and smaller amounts in the proventriculus (*d*), the wall of the gut (*e*) and in the yolk of the developing eggs (*f*). Fat is present in relatively small amounts and is confined to the 'fat body'. When the fly is newly hatched from the pupa these reserves are small in amount; but the stores of fat and glycogen are rapidly built up and reach a peak within about 4 days (Fig. 3, A).

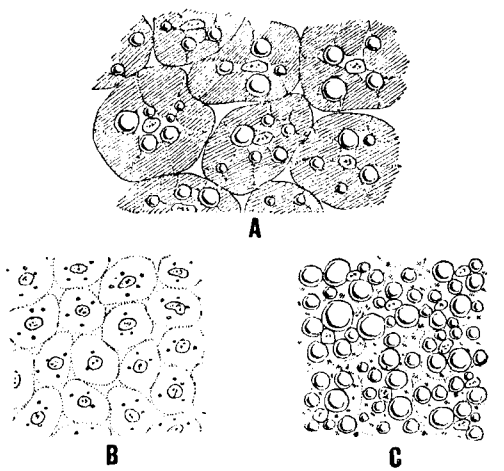


Fig. 3. Groups of fat body cells in *Drosophila*, drawn to the same highly magnified scale. The grey shading indicates glycogen, the refractile droplets are fat

(A) the resting fly; (B) after starvation until moribund: all the glycogen had disappeared from the shrunken cells; there are a few minute droplets of fat; (C) in fly which has flown to exhaustion: glycogen is reduced to small granules between the unchanged deposits of fat. The shrinkage of the cells gives the impression of an actual increase in the quantity of fat

If the fly is starved, being given water only, it runs actively about and dies within 48-72 hours. In the moribund fly both fat and glycogen reserves in the fat body are completely used up (Fig. 3, B). Clearly for this form of activity the fruit fly can utilize both fat and carbohydrate.

It is otherwise during flight. In order to study the flying insect

it is secured with wax to the head of a very small entomological pin which is inserted into the cork of a small vial (Inset, Nos. 17 and 18). A slight tap, or the sudden withdrawal of the cork, will set the wings going, and the insect will continue for long periods at a time in 'stationary flight'. If the wings stop the insect is at once made to fly again. A well-nourished fruit fly will keep going in this way for as long as 5 hours. Towards the end it requires re-starting at more frequent intervals, until finally the fly is exhausted: when stimulated, it spreads its wings, but they fail to vibrate.

The changes in the reserves of fuel in the exhausted fly are very different from those in the fly that is starved to the point of death (Fig. 3, C). Glycogen only has been used; the droplets of fat appear to be unchanged. And not all the glycogen has gone. It seems that 'exhaustion' supervenes when the reserves of glycogen are so reduced that they cannot be mobilized rapidly enough to supply the high demands of the muscles. It is not only the reserves in the fat body that are used up, but all the other stores of glycogen, including the deposits in the halteres. The halteres of flies represent modified hind wings. It is of interest to note that among their other functions they carry a reserve of fuel.

If the 'exhausted' fly is allowed to rest for a few minutes, it is able to fly once more – but only for a very short time; and the longer the rest, the longer is the duration of flight before it is once more exhausted (Fig. 4): a rest of one minute may be succeeded by a flight of only 10 seconds, a rest of half an hour may be followed by 2 minutes flight. While it is resting the fly evidently transfers some of its reserves of glycogen from the fat body to the muscles.

It was of interest to see how soon after a meal the power of sustained flight was restored in the exhausted fly. Flies which had been brought to the point when they could fly for only a few seconds after a rest of one minute were allowed to suck up a droplet of syrup (10 per cent glucose) during 30 seconds. At the end of this time flight was re-started and was continuous for at least half an hour. Sugar is made available for flight with extraordinary rapidity.

By feeding different sugars and other substances to these exhausted flies it is possible to compare their efficiency as sources of energy. Some, such as lactose, sorbose, rhamnose and ethyl alcohol, have no effect at all; they cannot be utilized. Some, such as glycerol and mannitol, lead to a small increase in the duration of flight after a standard period of rest. Others, such as galactose, xylose, etc., will allow repeated brief flights but will not support

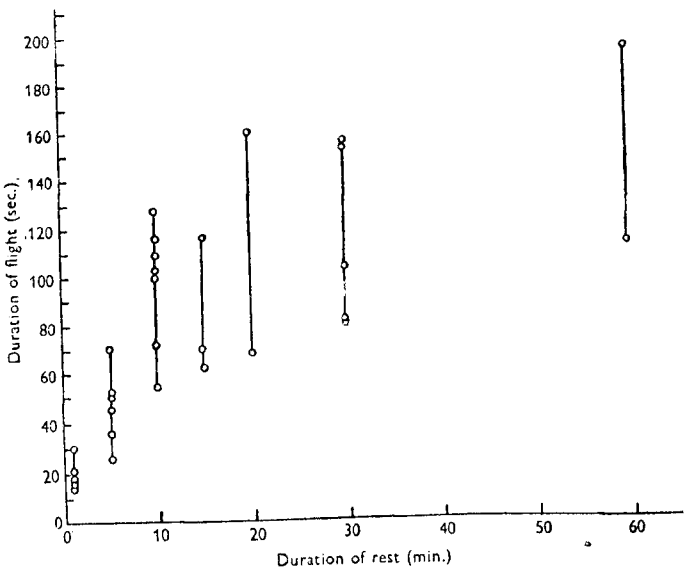


Fig. 4. Total duration of flight in exhausted *Drosophila* after varying periods of rest. Data from five flies of mixed age and sex

uninterrupted flight. Whereas fructose, maltose, sucrose, etc., will restore the capacity for continuous flight – although not quite so rapidly as glucose.

The exhausted fly can also be used to test directly the rate at which sugar is consumed during flight. A micropipette is made by drawing out a very fine capillary, coating it inside and out with paraffin wax and mounting it inside another tube (Wigglesworth,

1937). It may be graduated with the aid of a measured bead of mercury. It is then easy to offer to the suspended fly measured volumes of sugars of known concentration and to see how long flight will continue before the insect is once more exhausted. For example, 0.217 cubic mm. of 20 per cent glucose maintained flight for 290 minutes; and 0.108 cubic mm. of 5 per cent glucose maintained flight for 27 minutes. From the average of a fairly large number of experiments it was found that 1μg. (one millionth of a gram) of glucose will last the fly 6.3 minutes – during which time, as we know from the observations of Chadwick, it will make about 75,000 wing beats.

We have seen that the glycogen carried in the knobs of the halteres is also consumed during prolonged flight. It was of some interest to form an estimate of the value of these reserves. Making various assumptions it may be concluded, as a very rough approximation, that the two halteres carry about 1.4μg. of glycogen – enough to provide for about 9 minutes flight, out of a total flying time of about 4½ hours. This is not a very important reserve. The halteres, as Pringle (1948) has shown, act as gyroscopic sense organs which enable the fly to perceive deviations in its course through space. Perhaps the glycogen deposited in the knob may serve to improve this function slightly by increasing the specific gravity of the knob and so adding to its inertia.

REFERENCES

Chadwick, L. E. 1947. *Biol. Bull. Woods Hole*, **93**, 229.
Chadwick, L. E., and Gilmour, D. 1940. *Physiol. Zool.*, **13**, 398.
Pringle, J. W. S. 1948. *Philos. Trans. Roy. Soc.* B, **233**, 347.
Pringle, J. W. S. 1949. *J. Physiol.*, **108**, 229.
Sotavalta, O. 1947. *Acta Entom. Fenn.*, **4**, 1.
Wigglesworth, V. B. 1937. *Biochem. J.*, **31**, 1719.
Wigglesworth, V. B. 1939. *The Principles of Insect Physiology* [Methuen & Co. Ltd, London.]
Wigglesworth, V. B. 1949. *J. Exp. Biol.*, **26**, 150.

ON LOOKING AT OLD INSTRUMENTS

A. R. HALL

INEVITABLY the growth of professionalism in science during the last century had its drawbacks as well as its great achievements. Some of them, especially the political ones which are of such great interest to us to-day, are too obvious to require comment. But the changing attitude is noticeable in minor ways too, producing, at least for a period, some rather unhappy results, as for instance in the design of scientific instruments. There have of course always been a few professional scientists or quasi-scientists, men who lived by the practice of astrology or alchemy and in more recent times by teaching navigation, architecture or fortification. And there have always been instrument-makers – a trade which became in a sense universal after the introduction of spectacles about the beginning of the fourteenth century. Until recent times, however, the names which live in the history of science are those of amateurs, with the exception of a few medical men whose work was connected with their researches in anatomy or physiology, and of a very few university teachers, like Newton between 1669 and 1696. Even from these research was not expected as it is in a modern university department; indeed innovations were looked upon with suspicion. The remainder were mostly learned clerics in the Middle Ages, men of all professions whose leisure enabled them to indulge in scientific investigations.

As a result of this, science fitted much more naturally into the whole intellectual, even habitual, background than it does to-day – or at least than it has done until the last years, for there are signs that as science and the scientific method dominate more of our lives, we are on the verge of attaining a new scientific synthesis comparable to the Aristotelian-Christian synthesis

which was destroyed by the seventeenth-century revolution. As yet habits, tastes, and forms of speech survive to remind us of the pre-scientific age. Although there has been built up over the last century and a half a great store of knowledge about genetics, biochemistry, psychology, we continue to marry, eat, and think much as our ancestors have always done. Because we are so well aware of the existence of a dual standard, the commonplace and the scientific, we have debased the meaning of this latter adjective till it implies no more than 'exact' or 'efficient'.

When we consider in the history of science the various relationships which have existed between scientists or natural philosophers and the civilizations in which they have lived, we should expect to find corresponding changes in the making of the instruments they used. There are of course many ways in which collections of old instruments and apparatus may be studied. One can begin by seeing how capable a particular instrument is of fulfilling the purpose for which it was designed. One can see how the increasing usefulness and power of the microscope or telescope have followed upon the progress of glass-grinding and optics, and have in turn permitted a new depth of observation in biology or astronomy. The functional study of a type of instrument is important, but relatively straightforward.

Such collections can also be divided otherwise than into evolutionary series. One can, for instance, distinguish between instruments intended for teaching or routine use, and those made for research. As we go backwards in time the former become proportionately more common; for contemporary science we have great numbers of individual pieces of apparatus used in the laboratory, very few of which have distant cousins familiar to the non-professional. These two classes must ultimately return to the same common stem – an expensive modern machine for taking photomicrographs with every possible refinement is as much descended from the primitive compound microscope of the seventeenth century as are the toy microscopes given to children. If we should choose the clock as our example, we find

that the modern electronic time-keeper is not merely completely distinct in principle from a lady's jewelled wristwatch (of which the mainspring and train of gears were already known in the fourteenth century, the balance wheel in the seventeenth), but that the lady's watch, since it is no longer a scientific instrument, though it may keep time more accurately than the astronomical clocks of three centuries ago, preserves as an article of personal adornment those characteristics of beauty and elegance which the astronomical clock then had, and ours no longer possesses.

Another chronological distinction is possible, more closely related to design and style. Here I would tentatively suggest four main periods. The first, the traditional, extends up to about 1600. Of this we have few relics – astrolabes*, sundials, clocks, surgical instruments and quadrants being the most important. There were very few novel instruments; design was stereotyped because men went on making after an established pattern, and it was greatly influenced by aesthetic considerations (See Inset, Nos. 19–22). In the next period, roughly 1600–1800, many new instruments were invented, but again design was largely influenced by contemporary taste and the functional approach developed slowly. In the third period, covering the nineteenth century and a little beyond, the tradition of design and fine craftsmanship disappeared, the elegance of the eighteenth century was lost, and functionalism was dominant. Greater flexibility and accuracy were obtained, but by means of complicated and intricate mechanisms at the expense of neat and simple designing. Finally in the modern period we see instruments in whose design the maximum utility is combined with a pleasing appearance given by clean lines, balance, and effective use of colour contrasts. These are features reflected in all modern designing, since rich detail has given way to a mood almost of severity, though it may be asked whether this change in taste is not itself due partly to the touch of science directly in providing new materials, such as plastics, and indirectly by stimulating our desire for efficiency.

* Or planisphere – a flat projection of the circles of the celestial sphere, with the principal stars, used for determining latitude, time, etc.

Artificial aids to the natural powers of observation are prehistoric; Stonehenge is an observatory of sorts. A pole or stone set upright in the ground for making solar observations is the most elementary of scientific instruments, and the sundial is the same thing in a more accurate, calibrated, and of course attractive form. (See Inset, No. 19). The astrolabe was the universal mediaeval instrument for astronomical observation, for navigation, for casting horoscopes and chronometry, but in the works of the Danish astronomer Tycho Brahe (1546–1601) the idea of an observatory equipped with the necessary quadrants, both fixed and portable, time-keepers, celestial globes, and so on, is already clearly delineated. Similarly there was an established design for chemical apparatus, particularly that required for distillation; the apothecaries' balance had reached a stabilised form from which to-day's chemical balance is derived; instruments were in daily use before the beginnings of modern science for levelling, surveying, and navigation, on well-defined principles. The skill of the spectacle-maker, developed during the late Middle Ages, was the source of the rapid improvement of the telescope and the microscope from the opening years of the seventeenth century, which saw the invention of many other new instruments – the barometer, thermometer, air-pump, slide-rule, and pendulum clock.

If the number of inventions of instruments were plotted against time throughout history the curve would probably rise very gradually to about the end of the sixteenth century, with minor upward trends representing the peaks of Greek and mediaeval civilization; then it would bend ever more steeply upwards to the present day. It would resemble but not exactly correspond to the similar curve for technological inventions. Now it is obvious that the closer the point at which a new instrument emerges is to the present, the more rapid its evolution seems to be. The design of an astrolabe was little changed during nearly 400 years, whereas radar underwent a complete transformation in ten. At first sight it seems as though the growing faculty for invention, whatever that may be, operates as well through the restless modification of existing instruments, as in the discovery of

entirely new ones. But we may observe that this is not universally true, that even in the modern period a decline in the rate of improvement of established apparatus may be consistent with a fertile power of invention. The structure and optical system of microscopes and portable telescopes changed much more notably between 1740 and 1840 than between 1840 and 1940, although the theodolite, perhaps owing to its greater mechanical complexity, was improved more rapidly in the latter than the former century. Perhaps, therefore, before assuming from the stability in the design of instruments represented in our collections over a considerable period of time a certain lack of inventiveness, we ought to consider whether that established design was not practically perfect for the work required of it. Thus in a dynamic period interest would shift from the increasing ingenuity of the instrument-maker to the change in the frame of reference, which is often far wider than that of the research worker.

Towards the end of the seventeenth century we begin to perceive a new trend. In the first place men began to appreciate that natural philosophy – natural science – could be learnt more easily with the aid of visual demonstrations. And as in the Middle Ages it would have been unthinkable that a man with any pretensions to learning should not know Aristotle's physics and John of Holywood's astronomy, so in the eighteenth century when the new science of microscopes and telescopes, pulleys, levers, and magnets had swept away mediaeval philosophy, the popular lectures of Hauksbee and Desaguliers (the forerunners of the Royal Institution) were illustrated by experiments, and collections of apparatus even began to find their way into the universities. While, however, many new instruments (microscope, sextant, opera-glass) were quickly accepted outside the field of the research scientist, we may also note, first of all in observatories, the development of exclusively laboratory equipment, though this still provided only a small part of the instrument-maker's trade.

The unique pieces of research apparatus which we sometimes hear about in the pre-scientific age have rarely survived; early instruments have been preserved because they were made in large

numbers, like the astrolabe, or because, like the magnificent astronomical clocks from Vienna shown in London in 1949, they were made as treasures for wealthy patrons. Right up to the beginning of last century instruments were made to please as well as to carry out precisely a particular function. They were the products of a craftsmanship whose power to display rank and taste was fully appreciated. Exactly the same attitude of mind is of course displayed in other crafts which we should regard as purely utilitarian, such as those of the armourer and the gunsmith. It was gratified without loss of accuracy; a gilded and engraved astrolabe made for a man of rank was more useful, because more complete, than that made for a mariner, and more skill was devoted to perfecting it.

In the same way comparatively few of our instruments deriving from the period stretching roughly from the middle of the seventeenth century to the early years of the nineteenth are laboratory pieces in the modern sense. Luxuries, such as the orrery* by Graham now at Oxford, or the massive silver microscope made by Adams for George III, become relatively rarer – and what fantastic vestiges of a decayed tradition these are! But instrument-making did not yet exist solely for the benefit of scientists and technicians, who would not have been interested in the change from oak to mahogany (in for example, the framework of a sextant) that accompanied the similar alteration in the style of furniture. The size of the trade is evidence of this in itself. Dr Johnson (who used to recreate himself from time to time with chemical experiments), in one of his walks round London, without going very far could have called in the shops of Adams, Dollond, Nairne, Short, Cuff and Martin if he had wanted to consult only those instrument-makers who were most famous for skill and inventiveness. They served not a profession but a public which was charmed and amused by the wonders of science, and to which they offered 'diverting observations', 'pleasing and profitable spectacles'. And a society which placed a microscope and telescope, a pair of globes and perhaps an

* A mechanical model of the solar system, showing the revolutions of the planets and their satellites.

orrery just as naturally in the domestic background as a library (where, moreover, the *Ladies' Diary* might be found as a magazine of mathematical puzzles) could not fail to stamp its own character upon its scientific instruments.

In the nineteenth century, as science became more professional, interest declined. The drawing-room telescope and pocket sundial were forgotten in cupboards and were replaced by real toys – the kaleidoscope, the stereoscope, the magic lantern. Investigation was replaced by a mania for collection, resulting in cabinets of butterflies and books of dried flowers. This was, perhaps, only a temporary trend, for now again, though in a different way, we are all becoming customers of the instrument-maker. The cathode-ray tube is only the latest of many devices to escape from the laboratory into general use. A considerable industry has grown up during the last half century to furnish instruments which, like the astrolabe or the sundial, are intended to perform routine functions in office or factory, on board ship or in the home, and as before good appearance is a desirable quality; indeed poor or crude design from a technical point of view is often revealed to the unpractised eye in a clumsy or unbalanced instrument. The tools of science are once more being made to conform to social pattern – an ideal of simplicity most aptly suited to the particular problems of their design.

A THEORY OF CHESS AND NOUGHTS AND CROSSES

D. W. DAVIES

EVERYONE who has spent the tedious hours of the mathematics lessons in playing noughts and crosses will be aware of certain features of the game. In Noughts and Crosses you need never lose if you play well, whether you or your opponent takes the first move. Since the same applies to your opponent, it follows that you cannot expect to win if he is smart enough. The game is therefore rather trivial, because if both players are good the result will always be a draw. It is easy to imagine games in which one of the players can always win, though there seems little point in playing such a game. All these games are trivial in the sense that the outcome of the game, win, lose or draw, is to all intents and purposes a foregone conclusion.

At first sight, a game like Chess is an entirely different matter. If black could always win then nobody would play white, and if it could always be drawn, the game would consist of waiting for your opponent to make a mistake. Chess does not seem to be like this at all, but perhaps this is merely because we do not understand it sufficiently. Is it possible that the only difference between Chess and the trivial games in which the outcome can be foreseen is in complexity? We propose to show that this is so, and that there is a wide class of games in which the outcome is just as much a foregone conclusion as in Noughts and Crosses.

There seem to be different opinions on this question – some people feel that it is obvious and some, particularly chess players, rebel against the idea. Fortunately, we can decide in favour of the 'foregone conclusion' theory by mathematical methods, and that is what we propose to do. The mathematics is such that it can be developed from first principles and does not require the reader to have any special knowledge; in fact many readers will

probably think that very little mathematics is used at all. The same subject matter could have been written up in mathematically rigorous form, to please the mathematician, but for the average person a 'common sense' treatment such as we have given is easier and just as convincing.

Not all games have this feature of a foregone conclusion. There are four factors which can prevent it, which will now be described.

1. *More than Two Players*

The subject of games with three or more players is a difficult one, and quite different from the theory of games with two players. The 'foregone conclusion' result is a characteristic of two-person games, and may be thought of as arising from the simple opposition of interests. In games with more than two players the possibility of coalitions must be considered, and there is no longer a simple opposition of interests.

2. *The Lack of a Stop Rule*

A well-known phenomenon in draughts is the dodging about of two opposing pieces in a corner. If this is allowed to go on for ever, the game has no outcome at all. However, in most games a rule is introduced to bring the play to an end. This may be called a stop rule, and to be completely effective it should ensure that the game ends before a certain finite number of moves have been made. The game is then called a finite game. While Noughts and Crosses is obviously finite, having a maximum of nine moves, Chess, with such stop rules as are usually applied, can have an astronomical number of moves. Nevertheless, the number of moves should be restricted by the stop rule to be finite, or the stop rule would not be serving its purpose, which is to force a conclusion whatever the players do.

3. *Chance Moves*

Most card games begin with shuffling and dealing, a process which is governed only by chance (or should be). This prevents any 'foregone conclusion' except in a probability sense. The

theory can be extended to cover this case, and to show that the mathematical expectation of the outcome of the game is a foregone conclusion, but we shall not go further into this.

4. *Lack of Perfect Information*

By this, we mean that a player does not know to the full what moves his opponent has made. A good example is the game called 'battleships' in which the first moves, the drawing of 'cruisers' and 'destroyers' on the paper, are unknown to one's opponent. The result of foregone conclusion can be proved for such games only in a probability sense, and we shall not treat these games here.

We therefore state that our theory will apply only to 'two-person games which are finite, have no chance moves and have perfect information'. For shortness, we shall simply refer to *games* in future, meaning this special class of games.

While, as we mentioned above, there is a theorem of 'foregone conclusion' for all two-person finite games including chance moves and imperfect information, this theorem is quite different from the one we shall prove. For our particular *games* the theorem is a stronger one, and has more appeal to intuition.

PUZZLES

In order to introduce the ideas of the theory in their simplest possible form, we shall first treat a much simpler but allied problem, that of the 'one-person game', better described as a puzzle.

Suppose that we have some sort of *apparatus*, for example a draughts board and draughtsmen, and we begin by setting it up in some *initial state*. We then have rules which tell us what moves are allowed in any given *state of the apparatus*. When a move has been taken, the state of the apparatus has been altered, and the rules then tell us what new moves are possible. We have other rules also which tell us when the puzzle is finished. Let us suppose that the puzzle is *finite* in the same sense as the games we have been describing, so that after a certain number of moves the rules must always tell us that the puzzle is ended, and also

whether it has been solved or remained unsolved. We shall also contemplate a more general sort of puzzle in which, after the rules tell us the puzzle is ended, we can count up the number of points we have won, and the object is to win as many points as possible.

The former type of puzzle can be treated as a special case of the puzzle in which points are awarded. We have only to allocate 0 points to an unsolved puzzle and 1 point to a solved one, then the object of the game is once more to win as many points as possible. We can cover all cases by using the following convention:

The outcome of a puzzle is one of a certain set of integers 0, 1, 2, 3 . . . (There may be only 0 and 1 in the simplest case) which can be interpreted thus

0 – The puzzle has not been solved
1 – The puzzle has been solved
2 – The puzzle has been solved in a better way than 1
3 – A better way than 2
etc.

The object of the puzzle is to increase the outcome number. The outcome 0 corresponds to a complete failure.

We have now described the sort of puzzle which we are considering. Such puzzles usually have little mathematical interest. To obtain the best solution one would merely have to consider all the possible alternatives and find the set of moves which leads to the greatest outcome. This is simply a maximising problem, and, while it may be interesting in certain isolated cases, the general form of the problem is easily understood – unlike the case of games. In spite of this we shall develop a mathematical model for the general puzzle because this will serve as a simple introduction to many of the ideas to be used later with games.

The Tree

The mathematical model to be used for puzzles and games is called a 'tree' and is illustrated in Fig. 5 which also has a key to the nomenclature.

The elements of the tree will be called spots, because of their appearance, and certain relations between these elements are shown by lines joining them. The chief characteristic of the tree

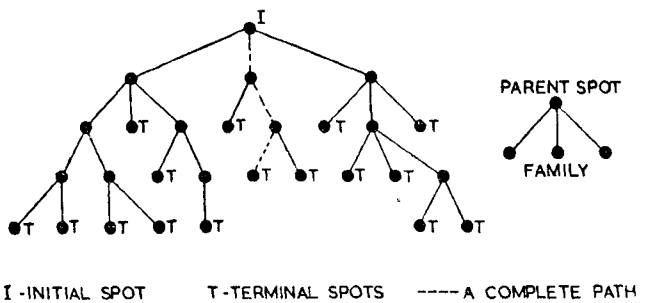


Fig. 5. A typical tree

is that its spots are naturally arranged in levels so that each line goes between adjacent levels. In general a spot is related (joined by a line) to only one spot in the next higher level, and this spot is called its *parent*. A sole exception to this is the spot in the topmost level, from which the whole tree springs and which has no parent.

Spots may be related to several spots in the next level below. These spots are called its *family*. Since the tree is finite, all the branches must end somewhere, and these ends are called *terminal spots*. They are the only spots without a family.

One further idea associated with the tree is that of a *complete path*. An example is shown in Fig. 5. It is a set of the lines reaching from the initial spot down the tree, level by level, to a terminal spot. The spots on the complete path define it, but, if necessary, a complete path can be defined merely by specifying its terminal spot. From this spot, the path may be traced upwards to the parent, and from each spot reached to its parent until the initial spot is reached. There is an exact correspondence between terminal spots and the complete paths which they determine.

Each of the terminal spots determine a complete path, and the complete paths of all the terminal spots together make up the whole tree. If some of the terminal spots are selected, their

complete paths make up a tree which is only part of the whole tree. This will be called a *subtree*, and its distinguishing features are that its initial spot is the initial spot of the main tree and all its terminal spots are terminal spots of the main tree. A subtree is illustrated in Fig. 6 by dotted lines. This particular subtree includes all the complete paths corresponding to the terminal spots marked S.

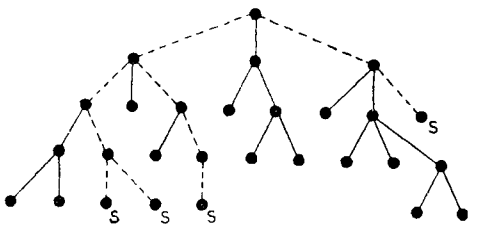


Fig. 6. A subtree

The Tree as a Mathematical Model of a Puzzle

The general scheme of using a tree as a model for a puzzle should be fairly obvious. The start of the puzzle is represented by the initial spot. All the possible first moves are represented by the lines leading down to the second level, at which the new states of the apparatus (after the moves) are represented by spots. For each of these states, new moves allowed by the rules are represented by lines down to the third level of spots at which the new states reached are represented, and so on. Thus the tree diagram represents those rules of the game which determine, for every state of the apparatus, what new moves are now possible, and the choices are represented by the branching of the tree.

It often happens that some state of the apparatus can be reached in more than one way. On the tree diagram, this would be represented by a 'confluence' like that of Fig. 7. This would spoil the tree structure, and it would be awkward if, for example, some state were obtainable by two alternative routes involving different number of moves. This difficulty is completely avoided and a genuine tree-structure obtained if a slightly different scheme is employed, which will now be described. It is still true

that the branchings represent the available choices, but the spots no longer correspond uniquely to states of the apparatus.

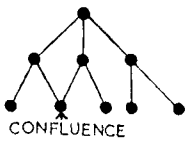


Fig. 7. A confluence

The new scheme is this. A complete set of consecutive moves which, starting from the initial state, reaches one of the end states of the puzzle is called a *complete play*. Each such complete play is represented by a terminal spot on the tree. A set of consecutive moves starting from the initial state but not reaching the end is called a *partial play*.* Each such partial play is represented by a spot of the tree (excepting the terminal spots which are already accounted for). In the previous scheme any move allowed by the rules altered the state to a new one. In this new scheme any new move added on to the partial play converts it into a new partial play. The new partial plays obtainable from a given one by adding one new move constitute its *family*. Correspondingly, if the last move in a partial play is removed, we have its *parent*.

When the spots of the tree represent partial plays instead of states, confluences like that in Fig. 7 are avoided. A state obtainable by two different sequences of moves, like that in the third level of Fig. 7, would be represented by two spots, one for each

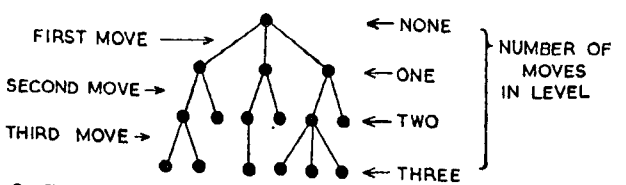


Fig. 8. Correspondence of spots in a tree to partial plays in a game

* The use of partial plays is quite common in describing Chess games. A typical partial play on the sixth level of the tree is *P-Q4, P-QB4, Kt-KB3, Kt-B3, B-Kt5*.

of the partial plays leading to it. The levels of the tree have a simple significance, for they classify the spots according to the number of moves in their partial play. This is illustrated in Fig. 8. The virtue of partial plays is that they have a unique parent, as the definition shows at once. To summarise the relation between 'tree' on the one hand and 'puzzle' on the other:

TREE	PUZZLE
spot	partial play
line	move
terminal spot	complete play
initial spot	a 'partial play' with no moves at all
parent spot	partial play obtained by deleting the last move of a given partial play
family of spots	partial plays obtained by adding some possible next move to a given partial play

All that remains is to represent the outcome of the puzzle, and this is simply done. Possible outcomes are represented by numbers, 0, 1, 2... and one of these outcomes is associated with each complete play. To incorporate this on the tree, the outcome numbers are placed alongside the terminal spots.

The system of a tree with an outcome associated with each terminal spot will be called a 'tree with outcomes'. We assert that it represents all the essential features of a puzzle.

Solution of the Puzzle

Although finding the best solution of such a puzzle is mathematically a trivial problem (however difficult it may be in practice), we shall go into the question in some detail to form an introduction to the same problem for games.

Solving a puzzle when the 'tree with outcomes' is given consists of examining the terminal spots to find which has the greatest outcome. There may be several terminal spots all having this greatest outcome. These represent complete plays all equally meriting the title of 'best solution' of the puzzle. This case is interesting because the set of 'best terminal spots' defines a subtree, which may be called the 'best subtree'.

An example is shown in Fig. 9. This is not taken from an actual puzzle, because they tend to be too complicated to show easily in one diagram.

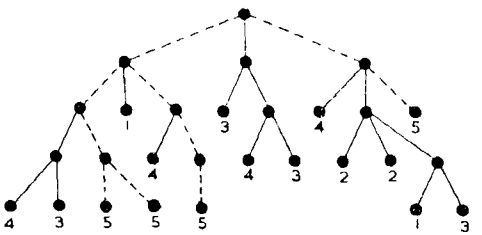


Fig. 9. A puzzle (with best subtree shown dotted)

The best outcome obtainable is 5, and the subtree contains all complete paths leading to terminal spots of outcome 5. This subtree is a guide to the player for obtaining the best outcome. It never lets the player down because at each spot reached by the subtree at least one choice is available for the next move – in other words it has no terminal spots except at the terminal spot of the original tree. Such a 'system' showing the player what to do (perhaps a number of alternatives) in all cases he can come across, is called a *strategy*. Our result is that any subtree gives a strategy, and the best subtree gives the best strategy.

Partial Puzzles and Values

Before leaving the Puzzle for the more interesting field of the game, we wish to introduce a new idea, that of the *partial puzzle*. Suppose that the puzzle is played up to a certain point by one player, then left to be completed by another. To the second player it is a new puzzle and the tree diagram is that part of the original tree leading downwards from the partial play at which the puzzle was abandoned by the first player. This new puzzle is called a partial puzzle. Two cases arise. If the first player played well, the play will still be in the best subtree, and the second player need only keep to this subtree to achieve as good a result as if he had started from the beginning. If the first player played badly, a new best subtree applies in the new puzzle,

but the best outcome obtainable will not be as good as formerly.

For each spot in the tree, the partial puzzle starting here can be analysed and the best outcome available can be determined. This outcome will be called the *value* of the spot in question. If the spot is in the best subtree, the value will be the best outcome of all, and this value will also apply to the initial spot and will be called the *value of the puzzle*. For the example in Fig. 9, the values have been written in by the side of the spots in Fig. 10.

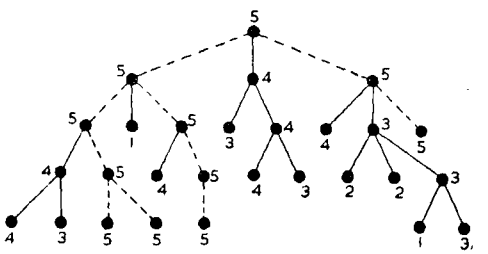


Fig. 10. The values for the puzzle of Fig. 9

It will be seen that 5 is the value of the puzzle, and all the spots of the best subtree have this value. All other spots have less value. The advantages of the best subtree can therefore be expressed in two statements.

(i) by keeping to it, the best outcome can be obtained.
(ii) Any straying from it involves sacrificing this best outcome for something less.

By examining the values in Fig. 10 it may be seen that simple rules permit the values to be entered on the tree without the necessity of considering all partial games. The rules are these:

A. The value of a spot is the maximum of the value of its family spots.

B. The value of a terminal spot is its outcome.

Obviously the rules A and B define all the values in the tree in terms of the outcomes because when the outcomes are known, values may be entered against the spots, working up the tree to the top.

All these ideas will be of particular importance for games, where the approach by means of values is the only one possible. In order to help in preparing for the more difficult case of games, some more remarks about puzzles will be made, which can be verified for Fig. 10, and in any case are intuitively obvious.

(1) In any choice of moves, the value stays constant for at least one possible choice, possibly more. For all choices where the value is not constant, it decreases. If only one move is possible, the value stays constant. A typical case is shown in the family of choices of Fig. 11.

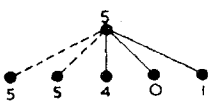


Fig. 11. The values in a family (puzzle)

(2) In a complete path, the sequence of values of its spots starting at the initial spot and going downward to its terminal spot is a decreasing or constant one. A typical sequence would be 5, 5, 5, 4, 4, 4, 3, 3.

(3) If complete paths are classified into two kinds, constant ones and ones which decrease somewhere, the former will be found to constitute the best subtree, the latter will be found to lie partly outside it.

When all these ideas are fully grasped, the problem of games is merely a simple extension of that of puzzles; as in all problems, the difficulty is to find the right approach.

DESCRIPTION OF A GAME

A game differs from a puzzle in that two players are involved, and their interests are opposite (if their interests were the same, it would be the same problem as a puzzle). The two things nevertheless have a great deal in common, in fact the tree structure applies just as well to games as to puzzles; but whereas in a puzzle all the choices are made by one player, in a game they are divided between two players. The branchings in the tree which correspond to these choices are therefore of two types,

which must be distinguished. Another way in which games differ is that the players, having opposing interests, look at the outcome in a different light. We shall consider these points in this order, and introduce new definitions to cover them.

The first part – the tree – is the same as for puzzles, since the state of a game can quite well be expressed by a partial play, as indeed it usually is for Chess. As with puzzles, we do not make much use of the state of the game as recorded on the apparatus, except to know what moves are allowed and when the game is ended. On the tree we represent partial plays.

A *partial play* is a set of consecutive moves starting from the initial state but not reaching an end state of the game (the moves must be allowed by the rules).

A *complete play* is the same thing, but reaching a state in which the rules tell the players the game has ended.

The *parent* of a partial play is the different partial play obtained when its last move is deleted. It is unique.

The partial plays are represented by spots on a tree, with the 'parent' relation represented by lines joining them. Complete plays correspond to terminal spots.

All this is exactly the same as for puzzles. The choice of moves in a game is given by rules depending on the apparatus, and on who moved last. We shall not assume that the moves are alternate. We call the players A and B, and the choices belong sometimes to one, sometimes to the other. This is shown on the tree by affixing to each spot the letter A or B according to who, in the partial play represented, has the next move. Thus all the lines downward from an A spot represent the various moves open to player A. The new system in which all spots (except the terminal ones) have A or B affixed will be called a 'tree with affixes'. The complete specification of a game is a 'tree with affixes and outcomes', but we first have to decide on a standard notation for outcomes.

As an example, when the game ends A and B may have to count up the points they have gained. The player with the greatest number of points will win, but the object of the game, besides winning, will be to gain as many points as possible.

For the outcome number in such a game, a suitable value is

A's points minus B's points.

When this outcome is positive A wins, when it is negative B wins, while zero means a draw. The object of A is to obtain the highest outcome number, while that of B is to obtain the most negative.

Whatever set of outcomes the game has, a similar outcome number can be devised so that the larger (more positive) the number, the better for A and the smaller (or larger negatively) the better for B. Since only the order of preference among outcomes matters (compare with the case of puzzles) the outcome numbers can always be the positive and negative integers:

$$\ldots -2, -1, 0, 1, 2, 3, 4, 5, \ldots$$

$$\longleftarrow \qquad \longrightarrow$$

B's interest A's interest

All this consideration of many outcomes is an extravagance from one point of view, since the object of all games is to win, or if this is not possible, to draw. Any game can be looked at from this point of view, so the outcomes:

A wins draw B wins

have a special significance, and will be assigned the numbers:

$+1 \qquad 0 \qquad -1$

Even when many outcomes arise, they are divisible into these three cases, and the numbers —2, —1, 0, 1, 2 . . . can always be arranged so that 0 corresponds to a draw and positive numbers to A winning, negative numbers to B winning.

Assuming that the object of the game has been decided, outcomes can be assigned in accordance with the convention given above. The essential part of the convention is the rules

A's object is to attain maximum outcome
B's object is to attain minimum outcome.

An example of a game specified by a 'tree with outcomes and affixes' is given in Fig. 12 and will be used in the remainder of the article as the illustration of the theory. We could not use an actual game as example because the tree diagram would have

been too large. We therefore give in Fig. 12 an arbitrarily concocted tree with outcomes and affixes chosen at random. The reader is advised to concoct a similar tree and carry out the processes described in the next three sections for this tree in order to understand them better.

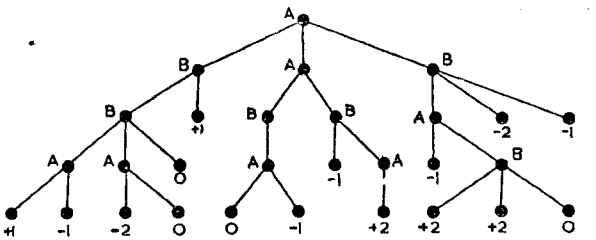


Fig. 12. Tree diagram of a hypothetical game

THE STRATEGIC ASPECT OF THE GAME

Imagine the two players A and B playing the game of Fig. 12 having been given nothing but this Figure to work with, that is, not knowing the original apparatus and rules of the game.

The play which ensues is represented by a complete path. The spots along this path are the partial plays, and the terminal spot at the lower end is the complete play. The outcome of the game for this play is marked by the terminal spot.

Neither player has complete control over the path which the game takes. A controls the choices at the A spots and B the choices at the B spots. A way of mechanising the game would be this – the lines would be channels, sloping down from initial to terminal states. A ball would roll down these channels for each play of the game. Switches at the A spots controlled by A and at the B spots controlled by B would guide its path. The terminal state at which the ball arrived would give the outcome of the game. This synthetic game of controlling the switches is in every way equivalent to the original game from which the tree might have been derived.

In the case of the puzzle, the one player had complete control over the path in the tree. He could decide to restrict his play to the 'best subtree', and since this gave him a complete set of rules

for how to play next we called it a *strategy*. In the puzzle, any subtree gave a strategy, but this is not so in a game because one's opponent must be taken into account.

A set of rules (expressed by means of a subtree) which tell Player A what to do under any circumstances will be called an *A strategy*. Similarly a set of rules for B will be called a *B strategy*. Consider what must be the form of an A strategy. For any A spot included in the subtree, it must give certain 'preferred moves'. There may be several, but there must be at least one so that if the play arrives at this spot, player A will know what to do. Hence the rule:

(i) For every A spot in the strategy, there is at least one spot of its family included.

This rule is actually covered by the statement that the strategy is a subtree. The distinguishing feature of a *strategy* subtree appears at the opponent's spots. Thus, at a B spot which is included in this A strategy, the choice is out of the hands of player A, for whom the strategy is intended. To allow for all eventualities, all the family of the spot must be included. Thus:

(ii) For every B spot in the strategy, all the spots of its family are included.

Rules (i) and (ii) suffice for an A strategy. Similar rules, *mutatis mutandis*, apply to a B strategy. While there is no essential difference between the two players, we shall always for convenience adopt A's point of view.

Returning to the mechanised game mentioned above, we see that player A need not set all the switches when deciding on a strategy, he need only set those corresponding to an A strategy subtree, the other switches will never be used. Exactly similar considerations apply to B.

It is possible to look at a game from the strategy viewpoint alone, forgetting the individual moves. Instead of associating a given outcome with a sequence of moves, it is associated with a certain A strategy and a certain B strategy. This strategic approach to a game, in which A and B play by chosing strategies rather than moves is an invaluable one for games of imperfect information, but for games such as we are considering it merely

complicates the issue, and even obscures certain points. While we shall not use the strategic approach to games, we shall need the idea of strategy subtrees later, and this is the reason for its inclusion here.

We may summarise our conclusions thus: A's control over the game is that of being able to confine the play to any A strategy subtree.

THE CONCEPT OF VALUE IN A GAME

In a puzzle we defined the value of any partial play as the best (greatest) outcome obtainable, starting at that point in the game. For a game in which one's opponent's moves are incalculable, such a definition is impossible, but one rather like it will be used. Suppose you have a choice of several different moves, each leading to a spot of which the value is known, presumably you will go to the spot with the best value. That is, A will go to the spot with greatest value, and B to the spot with least value, since value will be measured like outcomes. This will be used to determine the value of the spot from which you started. The value of a spot will be defined as the value of the spot in its family to which a 'rational' move would be made.

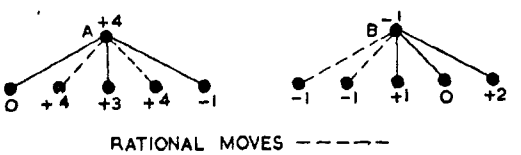


Fig. 13. Rational moves in a game

Examples for the two kinds of spot are shown in Fig. 13. From the A spot, a rational move is one to the greatest value, +4, in the family. The value of the parent is therefore determined as +4. For the B spot, the rational moves lead to −1, the least value, therefore the best for player B. The value of the parent is therefore −1.

In exactly the same way as in a puzzle, these rules determine values for all the spots on the tree in terms of the outcomes. The value of the terminal states are, of course, the outcomes

themselves, and all the values can be written in on the tree by working upwards, since the value of each spot is determined by those of its family. This process has been completed for the

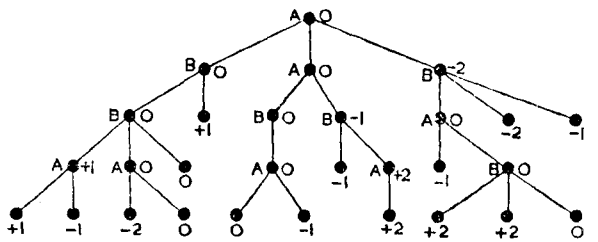


Fig. 14. Values for the game of Fig. 12

illustrative game of Fig. 12 and is shown in Fig. 14. In this game the possible outcomes are

$$-2, -1, 0, +1, +2$$

and, naturally, all the values of spots belong to this set of numbers. In a win/lose/draw game the values would all be $+1$, 0 or -1.

The concept of value, although it was defined rigorously enough, was obtained by considerations involving 'rational moves'. As far as your own moves are concerned, you may well decide to make them all 'rational' in this sense, but you cannot ensure that your opponent will do the same. If you say that to take a non-rational move makes him worse off any way, then the assumption that your opponent will play rationally may be justified, but this argument is only intended to show the correctness of your own 'rational moves', and hence is circular. We shall later justify the making of 'rational moves' and show that the value of a partial play is the 'foregone conclusion' which we are seeking, but these interpretations are not intuitively obvious, as they were for puzzles. If they were obvious, then the same reasoning would apply to three-person games, whereas in fact there is no foregone conclusion for such games.

In spite of this, the term 'rational move' will be used as a technical term with this meaning:

A *rational move* is one to a spot of the same value as the parent. A rational move is always possible at every choice.

The ordinary meaning of the word must (officially) be forgotten for the present, but actually the term 'rational move' is no more tendentious than the term 'value', the implication of which also remains to be justified.

JUSTIFICATION OF RATIONAL MOVES

By referring to the cases of Fig. 13, certain facts can be verified, which follow from the definition of value given in the previous section. They express the difference between rational and non-rational moves.

1. A non-rational move by player A decreases the value of the partial play; this follows because the value of the parent A spot is greater than or equal to the values in the family. The rational moves lead to the spots in the family with value equal to that of the parent; hence the others lead to spots with less value.

2. A non-rational move by player B increases the value of the partial play. Finally:

3. A rational move by either player maintains the value constant.

If both players use rational moves entirely, the value will remain constant throughout the game. The outcome, which is the final value, must equal the value of the initial state. This value will be called the *value of the game*, because of its paramount importance.

Suppose that A uses rational moves, and B at least once makes an irrational one, the value remains constant at the value of the game until B makes his irrational move: then it increases. It still remains constant for A's moves, and for B's when they are rational, but if B makes another irrational move it increases again. The values of the partial plays in such a game form a sequence which stays constant or increases, such as

$$0, 0, 0, 1, 1, 1, 1, 2, 4, 4, 4, 4.$$

As a result of this, the outcome, which is the last value in the

sequence, is greater than the value of the game, which is the first value in the sequence.

Again, suppose that B uses rational moves and A does not always do so. The value stays constant except at A's irrational moves, when it decreases. In this case the outcome is less than the value of the game.

In the case when neither player keeps to rational moves, no statement can be made, the outcome can equal the value of the game, or be less, or greater.

The results are summarised below, where 'rational play' means making only rational moves, and irrational play means making at least one non-rational move.

TABLE GIVING OUTCOME FOR VARIOUS SORTS OF PLAY

	A plays rationally	*A* plays irrationally
B plays rationally	Equals the value of the game	Less than the value of the game
B plays irrationally	Greater than the value of the game	Unknown

The contention put forward at the start that these games have a 'foregone conclusion' has now been proved. The foregone conclusion, for good players, is that the outcome will equal the value of the game. This 'value of the game' is to be determined by the methods given above, and it may or may not be a draw. In the case of Noughts and Crosses it is a draw, and the table reads:

	A rational	*A* irrational
B rational	Draw	*B* wins
B irrational	*A* wins	Unknown

In the case of Chess, we do not know what the value of the game is. If it happens to be a draw, then the game consists, to a superman who knows the game perfectly, of waiting for the opponent to make an irrational move. It is, however, equally likely that the value is a win for white, for example, and white can always win by using only rational moves. In this case all the moves that black can make when white is playing correctly are rational because they all keep the value constant, although this value is a defeat for black.

The advantages that prove the superiority of 'rational play' will now be summarised. It must be remembered that you should take account of all that your opponent might do, you cannot assume that he is employing any ideas of 'rational play'.

(i) Using rational play you can achieve a result equal to, or better than, the value of the game.

(ii) Not using rational play, your opponent can always force a result worse for you than the value of the game. (He can do this by employing rational play.)

(iii) If your opponent has made an irrational move, and you have not, the value of the partial play has improved (for you). The considerations of (i) and (ii) now apply to the remainder of the game with the new value.

Thus the rational play makes the best of any situation that can arise in a game. One is therefore justified in saying that two players knowing the whole situation would always employ rational play, and always arrive at the same outcome – the value of the game.

The considerations of this section have not so far employed the tree diagram, but they can quite easily be stated in the terms of it. Consider, for example, the result of both players using rational play. The value remains constant throughout the play; hence the complete path in the tree is one with constant value. The collection of all such paths of constant value constitutes a a subtree of constant value, which will be called the 'rational play subtree'. It is the largest subtree all the spots of which have the same value. For the illustrative game, this subtree is shown in Fig. 15. In this case it is the biggest subtree with all values zero.

Now consider the result of A's using rational play. The paths in the tree corresponding to this are ones whose values stay constant or increase downwards, and they include the paths of

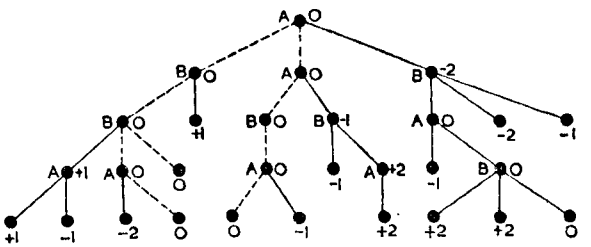


Fig. 15. Rational play subtree (shown dotted)

the rational play subtree. They include other paths in general, for at each B spot, all choices by B must be taken into account. All these choices correspond to constant or increasing values. All the paths whose values are constant or increasing form a subtree which will be called the *A subtree*. For the illustrative example it is given in Fig. 16.

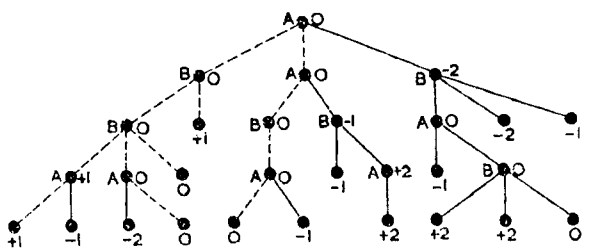


Fig. 16. The A subtree for the hypothetical game

This subtree is a *strategy* for A because at all the B spots, all choices are included. Player A can therefore keep the play on this subtree if he wishes, and the outcomes on the subtree are greater than or equal to the value of the game. It is the greatest subtree with all paths increasing or remaining constant in value.

A similar subtree of constant or decreasing value is called the *B subtree* and is the B strategy corresponding to rational play. It is shown in Fig. 17.

The part common to both the A and B subtrees is the subtree of paths with constant value, that is the rational play subtree.

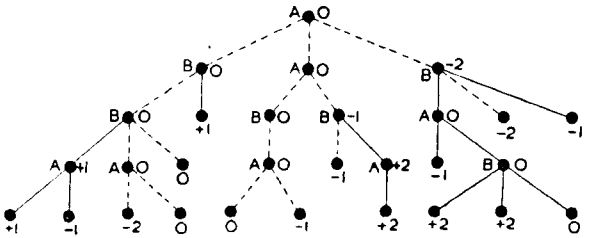


Fig. 17. The B subtree for the hypothetical game

The previous results can be restated in terms of these subtrees as follows: Player A can keep the play in the A subtree and hence ensure an outcome not less than the value of the game.

Player B can keep the play in the B subtree and hence ensure an outcome not greater than the value of the game.

If both players do so, the play is kept in the rational play subtree, and the outcome equals the value of the game.

If player A keeps the play in the A subtree, but B lets it go outside the B subtree, it goes outside the rational play subtree and the value is greater than the value of the game.

The parts of the tree outside these three fundamental subtrees do not matter to the game. The paths leading to terminal spots outside the fundamental subtrees have values which oscillate in some way, i.e. increase at some point and decrease at another. All paths which do not oscillate in value are included in the subtrees.

THE PRACTICAL PROBLEM OF GAMES

In the above theory a process for determining the fundamental subtrees of a game was given. Knowing these subtrees the game can be played in the best possible way. The process of finding them has three stages.

(i) Draw the tree of the game and add the affixes and outcomes. This process starts at the top and works downwards.

(ii) Calculate the values of all partial plays. This process starts

at the bottom and works upwards to the initial spot, where the value of the game appears.

(iii) Outline the fundamental subtrees. This process starts at the top and works downwards once more.

Incidentally, there is an alternative method in which all three processes are carried out simultaneously, working by complete paths and describing the tree path by path.

Although the processes are easily described, with actual games they are generally too complicated to carry out. By taking account of symmetry in the apparatus of the game and confluences (see Fig. 7), and by leaving out parts of the tree which can be shown to have no effect on the game, the work may be simplified, and small games like Noughts and Crosses can be analysed.

In a typical non-trivial game, there are, say, four possibilities for each move on the average, and 64 moves. The number of branches on the tree is thus 4^{64}, an astronomical number with about 40 decimal digits. The case of Chess and Draughts is incomparably worse, for the possible number of moves is much larger than 64. There seems to be no possibility, with electronic machines as at present contemplated, of analysing any game of reasonable size completely. To programme one of the electronic machines for the analysis of Chess would not be difficult. The programmer need not be a good Chess player, he need only know the rules. Unfortunately, the time of computation would be measured in an astronomical number of years.

THE THEORY OF GAMES IN GENERAL

The theory of games given above is a very good illustration of the method of Applied Mathematics. It may be objected that the treatment was not really mathematical, and this objection may be raised for one of two reasons. Firstly, because the treatment does not involve arithmetic or algebra or geometry. Actually this is not unusual in Mathematics (other than school mathematics), and many modern branches of mathematics are like the one described here. Secondly, because the treatment is not rigorous, a point mentioned in the introduction. If a rigorous treatment had been given it would have seemed more mathematical because of the symbolism and (in the most rigorous treatment) mathematical logic which would have been needed. Mathematical readers will have no difficulty in making the theory as rigorous as they wish.

The mathematical structure which was the basis of the theory was the tree. A rigorous treatment would have defined a tree in terms of 'elements' (spots) and a relation (parent) between them. Certain properties of this relation would have been postulated, and everything else would be proved therefrom. By a 'tree' we do not actually mean the collection of spots on paper, but a set of abstract elements with an abstract relation (having certain properties) among them. This 'tree' concept is valuable not only in the theory of games, but also in the theory of information and in certain electrical circuits met with in calculating machines. The same concept occurs, also, in alphabetic or decimal indexes and classifications. The abstract tree is an *extract* of the essential feature of all these systems.

Similarly, the 'tree with outcomes and affixes' is an abstract system which extracts the essential features of a game. If two different games corresponded to the same tree, they could be said to be mathematically equivalent.

The use of an abstract system as a model of a physical system has the advantage that the abstract system can be made very precise, whereas the physical one rarely is. Before applying the model, it must be verified that it fits the physical system. For example, before applying the theory of games one must test the postulates of the theory for the game in question. If they fit, then the results of the theory apply also.

The theory of games illustrates a feature of mathematics that becomes more and more common. The greater part of the theory consists of setting up definitions for new concepts (complete path, value, strategy, subtree) which are then proved to have certain simple properties. The fundamental theorems are dependent on several of these concepts, and the finding of the best definitions and the most useful concepts is the main part of a mathematical investigation.

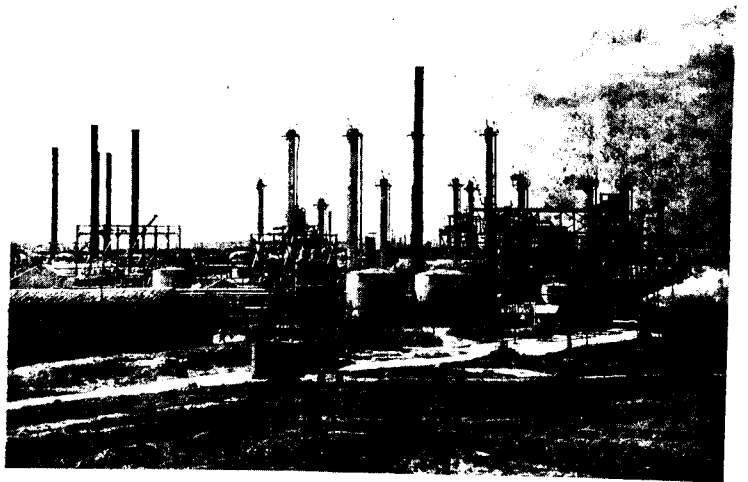
The treatment of games given here was based on that given in *Theory of Games and Economic Behaviour* by Von Neumann and Morgenstern. In this book games are treated mostly by the strategic method, because this method generalises easily to games of imperfect information. For games of perfect information without chance moves this strategic method obscures some points, and the method used here is preferable.

In generalising the result of 'Foregone conclusion' to the case of chance moves and imperfect information a different notion of outcome is required, which is quantitative. Since chance is introduced, a slight chance of large loss might be weighed against a large chance of small loss. For the case of imperfect information, the concept of 'strategy' is widened to include chance in the moves made by the players themselves. A simplified form of poker is treated, in which this new feature appears as the 'bluffing' with which we are familiar in the actual game. The bluffing is shown to be calculable, and not a psychological matter at all.

By these means, the foregone conclusion result is generalised to all two-person games, but it cannot be pushed further, for in games of more than two persons it is shown that only by consideration of coalitions is a solution obtained. There are in general many solutions of such a game, each taking the form of a 'standard of behaviour' for the players. While social implications are perhaps the first that come to mind, the authors of the book make most of their applications to Economics. In this field they show that it is more applicable than any branch of mathematics so far employed.

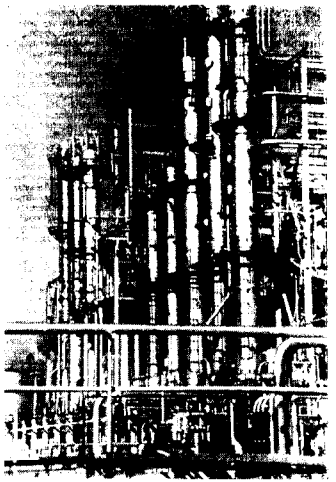
REFERENCE

von Neumann, John, and Morgenstern, Oskar. *Theory of Games and Economic Behaviour* (Princeton University Press, Second edition, 1947).

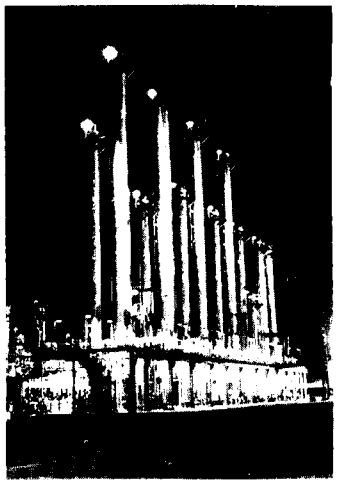


1. A general view of the new Shell Chemical Plant, Stanlow, near Chester, England. The Acid Concentration Unit is to the right and behind are the Fractionating towers of the Distillation Unit.

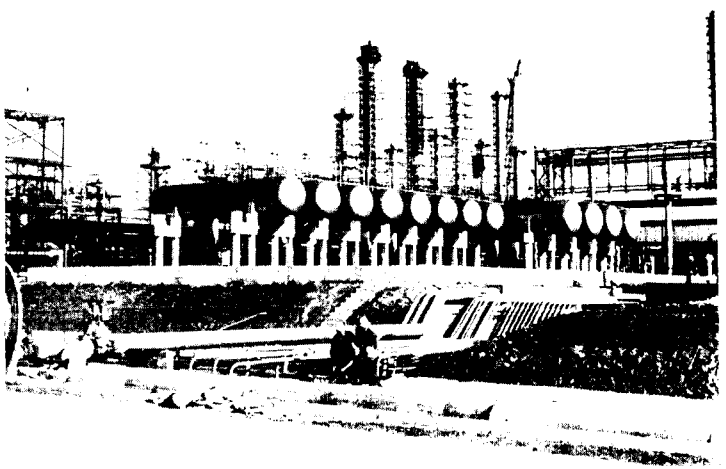
2. A view of the Hydrocarbon Fractionation Unit and on the right the Cracking Plant.



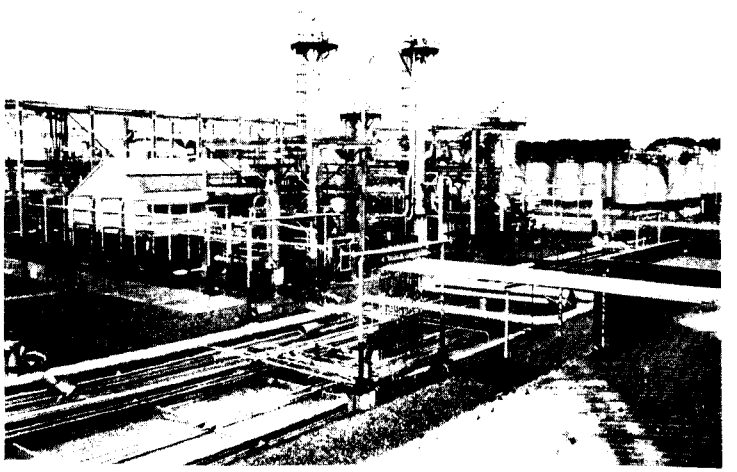
3. A closer view of the Hydrocarbon Fractionation Unit



5. The Distillation Unit working at night



4. Pressure storage tanks with the columns of the Distillation Unit being erected in the background.

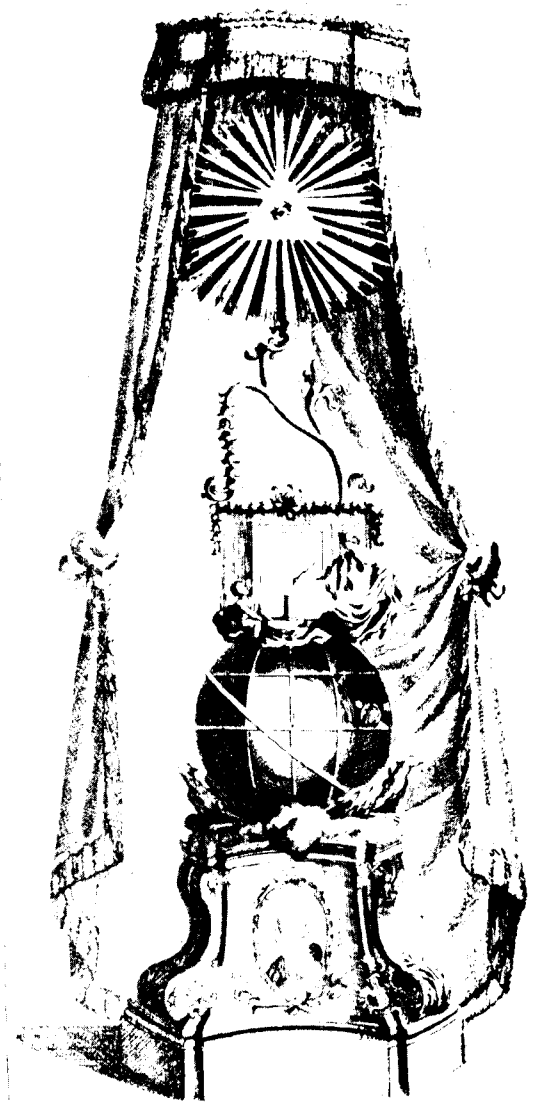


6. A view of the Diacetone-Alcohol/Mesityl Oxide Methyl Isobutyl Ketone Converters with part of the Tank Farm in the background

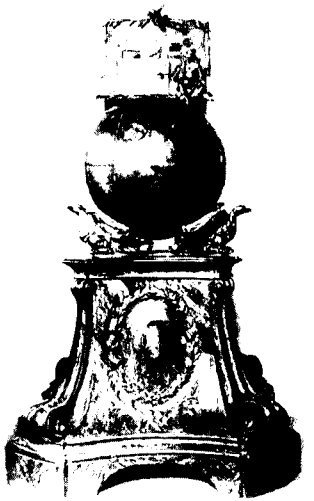
(By courtesy of Shell Photographic Unit)

7. Vaucanson's Automata in 1738

8. The Jacquet-Droz Writer. (*By courtesy of Cornish Brothers Ltd, from 'Musical Boxes' by John E. T. Clark*)

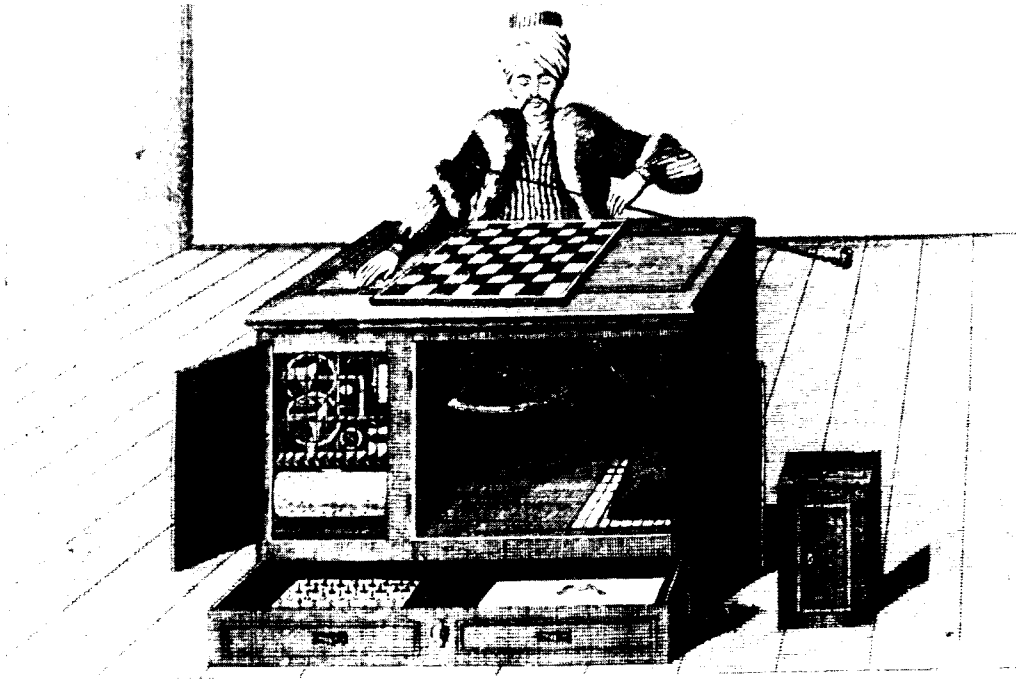


9. The writing machine of von Knaus as it appeared in 1760



10. Present-day appearance of the von Knaus Writing Machine

11. Part of the mechanism of the Writing Machine. (*Both illustrations by courtesy of the Director of the Technisches Museum für Industrie und Gewerbe Vienna*)

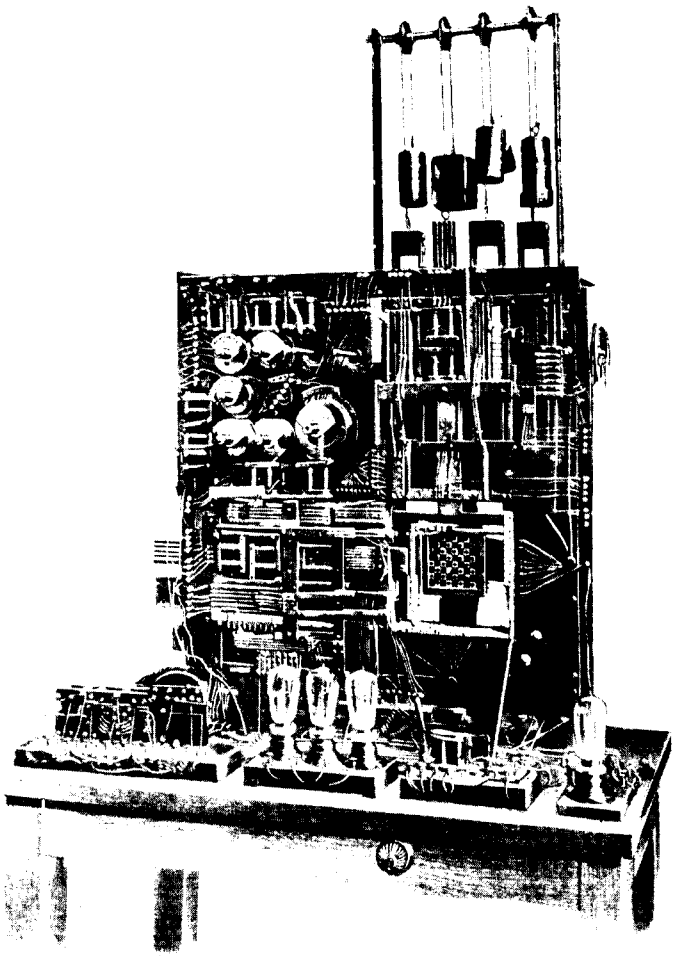


12. Baron von Kempelen's Automaton Chess Player

13. Front view

14. Back view

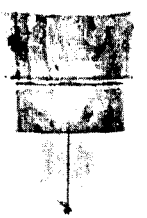
J. N. Maskelyne's Automaton Whist Player 'Psycho'
(By courtesy of the London Museum)



15. Chess Playing Machine of Sr. Torres
(Reproduced from 'La Nature', 1914)

16 Noughts and Crosses Playing Machine

FUEL CONSUMPTION IN THE FLYING INSECT



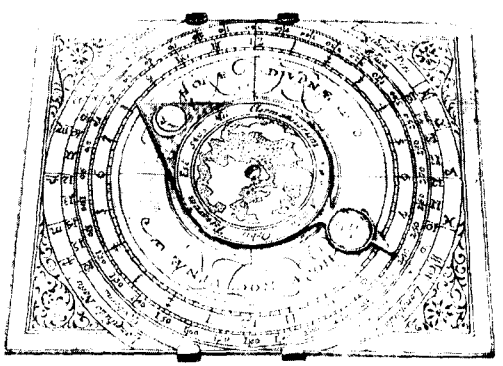
17

18

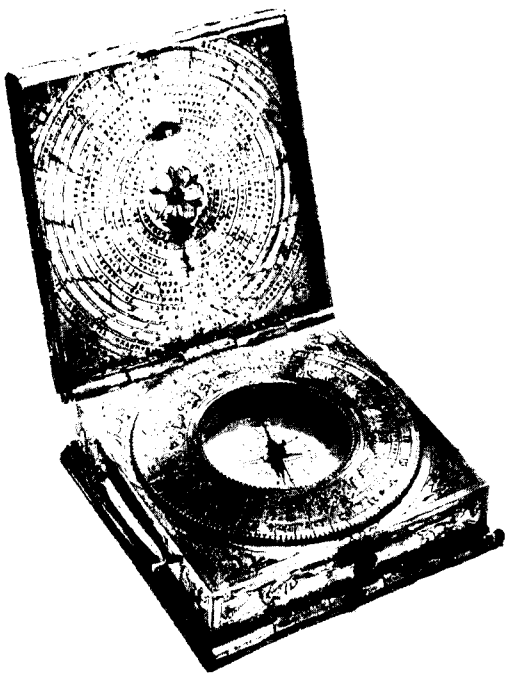
17. Glass vial 1″ 2″ showing fruit fly suspended from the cork.

18. Detail showing attachment of fruit fly by wax to the head of a fine entomological pin.

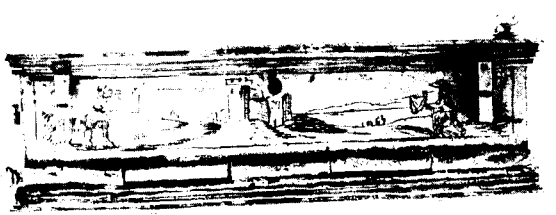
19. Universal Equinoctial Sundial by Johann Martin of Augsburg (*ca.* 1720), made of gilt and engraved brass, with the two hour-circles of silver, the whole folding flat to fit into a leather case. Note the rich decoration of thumbscrews and base.



20. Horological instrument by Erasmus Habermel (1580), made of gilt brass or copper, probably used in calculating the time of rising of the constellations. The centre piece rotates, with two pointers carrying the moon and sun, and has engraved on it a map of the world with its north pole under the sun. Again, note the decoration.

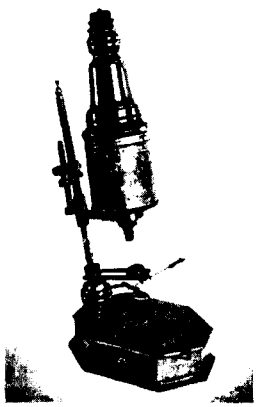


21. Another gilt portable instrument, this time including compass, sundials, quadrant and calendar, made by Tobias Volckmer (1645). There is another lid on the bottom of the box, which thus has all six outside surfaces engraved (see below).

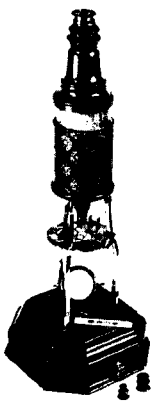


22. A detail of one of the sides, showing men surveying with quadrants.

Illustrations 23 to 27 show very briefly some stages in the evolution of the design of the microscope. Some of the changes were optical: note for example how more light is focussed on the specimen by the development of better mirrors, and of a substage condenser lens. Others were mechanical: the production of finer adjustment for focussing or for the quick change of lenses, the appearance of a mechanical stage for moving the specimen delicately under the high-power magnification, and the recurrent problem of stability – a heavy base to prevent the instrument being knocked over. But a third distinct type of change is also visible: in materials and design. The earlier microscopes have decorated bodies, and are made principally of wood, the late eighteenth century instruments have something of the geometrical grace and simple elegance of the furniture of the period, while the nineteenth century sees the appearance of solid, heavy, utilitarian microscopes, getting progressively more massive and complicated and even clumsy as the century advances.



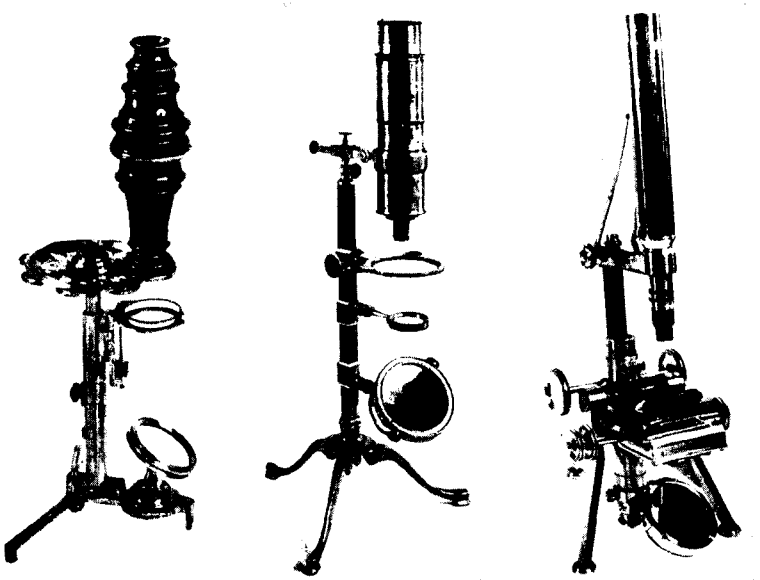
23



24

23. John Marshall's 'New Invented Double Microscope for viewing the Circulation of the Blood', *ca.* 1700. The body is made of wood and cardboard covered with gold-stamped vellum. Note the somewhat crude fine-adjustment screw, a device invented by the Danzig astronomer Hevelius, and the way the whole instrument swivels on its base through the universal joint.

24. A type of microscope named after Culpeper and probably made by Matthew Loft about 1740. The body is covered with tooled red leather, and mounted on three equally-spaced brass pillars. The tube slides up and down for focussing, and a disc of prepared objects rotates like a wheel on the specimen stage.



25 26 27

25. George Adams' 'New Universal Double Microscope', of the same period as No. 24, but made entirely of brass with the exception of the black ivory body and a steel screw worked by the nut visible at the centre of the base, as a fine adjustment. The wheel carries eight objective lenses of different magnifying powers to slide in turn under the lower end of the body tube. This is a portable instrument: the body tube unscrews and the base folds up.

26. Microscope by Powell and Leland (1846): one of the first in the modern style, mounted in trunnions, with a fine-adjustment operating on the nose-piece, and a mechanical stage.

27. 'Microscope Universel' by Dellebarre (Delft and Paris), dated 1785, and very similar to contemporary English types. All the optical parts are mounted on a bar hinged to the base.

(Illustration 23 to 27 are of instruments in the Collection formed by Mr R. S. Whipple, and are published by permission of the University of Cambridge. Photographs by Ramsey and Muspratt)

28 29

28. *Top.* Emission of a π meson during the explosive disintegration of a nucleus by a fast charged primary particle. (page 12)

Bottom. The π-meson is captured by a nucleus and causes a 'star'.

29. First observation of the successive decay $\pi \rightarrow \mu \rightarrow$ electron. (page 11)

(By courtesy of Professor C. F. Powell, F.R.S.)

ROBOTS WHICH PLAY GAMES

STANLEY BYARD

THE automaton fashioned into the likeness of man and capable of movement has for centuries excited the popular fancy. The savants of the Middle Ages, along with their supposed activities in magic and alchemy, were often credited with the construction of automata, the motivity of which was assumed to be partly mechanical and partly occult. Albertus Magnus, the Dominican systematist who lived in the thirteenth century, was said to have spent thirty years constructing a man of brass which served him as a domestic, and his contemporary, Roger Bacon, seven years in making a talking head of the same material which he hoped would advise him as to the defence of England. This interest in the automaton goes back to classical times. Then it was enough to simulate, for, as Ernst Mach has pointed out, it was hoped by imitating life outwardly to apprehend it from its inward side. The Promethean spirit is still within us and our interest is renewed to-day because computing machines are being made which, using means analogous to those which may exist in the central nervous system, perform some of the simpler logical processes of the human mind.

When in the fourteenth century striking mechanisms were introduced telling the hour and replacing the alarums of the earlier clocks, they were often made in the guise of automata. This was a very natural thing to do, for the mechanism did in fact take the place of a bellringer in relieving him of his more routine duties. The town of Montpellier in France in 1403 appointed a keeper of the belfry at Notre Dame, but his high wages and the unsatisfactory service he gave in return caused the authorities in 1410 to replace him with a large clock fitted with a wooden man which would strike the hours. The famous clock at Dijon boasts a family of such jacks, although originally it had but one.

It was not until the middle of the seventeenth century that Jacomart after nearly three hundred years of bachelordom took a wife to look after the half-hours. They were blessed with two children, the younger of whom did not arrive until 1855, and both are usefully employed in striking the quarters.

The seventeenth century saw in science the triumph of the mechanical world which was to be elaborated with such success in the following centuries. The Italian mathematician Borelli published in 1680 an essay on the motions of animals in which he analysed their movements on a mechanical basis. The duality of body and mind with man as an automaton directed by a soul, Professor Gilbert Ryle's 'ghost in the machine', became a fashionable philosophical concept which was frequently the subject of the metaphysical poetry of the age. Descartes was one of its most notable scientific exponents. In accordance with Church doctrine it was supposed that man alone possessed a soul and that in consequence animals were mere automata. Leibniz developed the mechanical concept still further and assumed a pre-established harmony between body and soul, comparing the dynamic elements or *monads* which made up his world to clocks which maintained perfect synchronism for all time.

The automaton in fact epitomised the philosophical outlook of the age and was symbolic of the new mechanical world. It is therefore not surprising to find mechanicians in the eighteenth century directing their efforts to the construction of elaborate automata. The best known constructor of such mechanisms was Jacques de Vaucanson, who built them to recoup himself, and who later in life contributed several important improvements to the silk weaving industry. His famous automaton flute player which took as its model Coysvoix' marble faun in the Tuileries was completed in 1736 after four years' work, Vaucanson then being twenty-seven. The figure was of wood, about five and a half feet high, seated on a pedestal which contained the greater part of the mechanism. The figure played twelve airs on the German flute, correctly stopping the pipe with its fingers, and moving its lips in order to voice the notes. It was also provided

with a tongue which was used to interrupt the wind to the flute. Vaucanson published a detailed account of the mechanism in a prospectus which was on sale during the exhibition of the automaton. The wind was supplied by three sets of clockwork-driven bellows differently weighted so that three values of air pressure were available for voicing the flute. The fingers, lips, and the tongue were moved by chains which were attached to keys or levers mounted in a frame within the pedestal. These levers were operated in the correct sequence and for the requisite length of time by a pegged barrel which was rotated by clockwork. Three additional levers were used to control valves which enabled the correct air pressures for the various notes to be selected.

The flute player was first exhibited in Paris in 1738 and its constructor received the praise of the Académie des Sciences. At this exhibition Vaucanson's mechanical duck, and pipe and tabor player, were also on show. The general appearance of these automata may be gathered from the contemporary engraving reproduced in Inset, No. 7. The pipe and tabor player took the form of a shepherd playing a three-holed flageolet which he held in one hand while he accompanied himself on the tabor with the other. Vaucanson says that he found the construction of this automaton extremely difficult, because in order to produce some of the notes on the pipe a partial stopping was required, while in addition the wind pressure had to be controlled over a very large range. In its perfect articulation of rapid passages it was said to excel the human player of the instrument.

Vaucanson's duck, which was life-size, excited considerable admiration on account of the anatomical correctness of its wing structure and because of its exact representation of some of the characteristic movements of the living bird. It would stretch out its neck to pick up corn placed before it, which it would swallow with avidity and later excrete. It could preen its wings and quack most naturally. These three automata by Vaucanson no longer exist and their fate is uncertain.

Friedrich von Knaus, who was born in 1724, and who in 1757 became mechanician to the Court of Francis I of Austria,

appears to have been the first to make machines which could write. He constructed no less than four writing automata, and the most elaborate of them has survived and is preserved in working order in the Technisches Museum für Industrie und Gewerbe in Vienna. Its original and its present-day appearance may be compared in Inset, Nos. 9 and 10. It is in the form of a small goddess seated pen in hand in front of a writing tablet and upon a sphere about three feet in diameter containing the mechanism. When presented to Francis I in 1760 it proceeded to write its own testimonial of some sixty words in the space of a quarter of an hour. The writing arm of the figure is under the control of three levers. In the sphere there are a number of triple cams each corresponding to a letter and each communicating movement to a set of three keys. The three levers guiding the arm of the figure may be engaged with any one of the cam mechanisms in turn. This engagement is effected by a pinned drum upon which the programme giving the required sequence of letters may be set up. The surface of the drum is drilled so that the pins may be inserted in the appropriate positions. The tablet is shifted horizontally after each letter has been written. Inset, No. 11 shows part of the mechanism and Fig. 18 an example of its writing to-day.

Fig. 18. Handwriting from the von Knaus writing machine. (*Courtesy of the Director of the Technisches Museum für Industrie und Gewerbe, Vienna*)

In 1774 three enchanting mechanical children made their appearance in Switzerland. They were the work of Pierre Jacquet-Droz, his son Henri and their assistant J. F. Leschot, and comprised two small boys, one of whom could write and the other draw, and a young girl who played a harpsichord. They are now in the Musée Historique de Neuchâtel. The young writer, Inset, No. 8, will when set in motion dip his quill in the ink, shake it and write a prescribed sentence in the fine careful hand of an

eighteenth-century notary, following his work attentively with his head and eyes. The whole of the controlling mechanism is contained within the body of the figure, occupying a space but 10 inches high, 5 inches wide, and 4½ inches deep. The forearm of the writer can turn about a vertical spindle at the elbow, thus giving a left to right or side to side movement to the hand carrying the pen. The hand is attached to a bar which can move in and out of the forearm to give an 'up and down' movement to the pen. This bar slides between four rollers carried in a cradle mounted within the forearm; the cradle itself is pivoted so that it may be tilted, thus raising or lowering the hand off the paper. By suitably combining the side to side and the up and down motions of the hand the figure may be made to inscribe any desired character, while a simultaneous control of the pivoted cradle will give the required variation in the pressure of the pen. These three motions are derived by lever connections from three sets of cams mounted upon a common vertical shaft. There are 120 cams in all, enabling 40 different actions to be made. By raising or lowering the whole bank of cams the levers communicating the motions to the hand and forearm can engage with the correct set of cams to produce the required character. The controlling mechanism which shifts the column of cams and starts and stops their rotation may be programmed to write a complete sentence, or alternatively it may be operated letter by letter. Fig. 19 gives

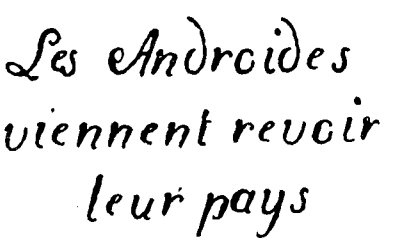


Fig. 19. Example of handwriting done by the Jacquet-Droz writer in 1906

an example of the writing done in 1906 on the return of the automata to Switzerland. The writing hand does not move bodily from

left to right as the work proceeds, but the paper is moved instead, a space at a time after each letter has been formed. The left hand which rests on the paper follows this movement, making it appear a natural one. The young artist is similarly constructed, except that the paper does not move. He has a repertoire of four drawings.

The automaton, true to its nature according to seventeenth-century thought, simulated in parrot fashion the appearance of reason. But in 1769 a new form of automaton appeared, a robot which could play chess. This was the work of Baron von Kempelen, Aulic Councillor to his Imperial Majesty Francis I of Austria and Director General of the salt mines in Hungary. It was of course a trick-automaton, but it entertained and intrigued Europe and America for more than eighty years. Frederick the Great and Napoleon are both said to have tried their skill against this automaton and to have lost their games. But the number of legends associated with this chess player is legion. Its career came to an end in 1854 when it perished in a fire in Philadelphia.

It was generally recognised that the performance was not a purely mechanical effort, but the modus operandi was so well hidden and the showmanship so perfect that the spectators found it difficult to reject the hypothesis that the machine was a pure automaton. Its appearance may be gathered from the contemporary engraving reproduced in Inset, No. 12. It was in the form of a Turk somewhat larger than life seated behind a cabinet on which the chessboard was fixed. At the beginning of the game the pipe was removed from his left hand and his arm supported on a cushion ready for play. A great display was made of throwing the interior of the cabinet open for inspection; doors were opened back and front exposing the mechanism, and a lighted candle held behind to prove that there was no obstruction which might screen a confederate. Notwithstanding this display the cabinet did contain a concealed player who changed his position as the various doors were opened and closed. Signor Blitz, a mid-nineteenth century American conjuror who met the chess player during its tour through the United States under the direction of Maelzel, referred to it as

the perfect counterpart of a magician's trick table. But the secret was well kept and the circumstantial solution which was published in 1821 and which has been attributed to Professor Willis of Cambridge, while probably correct, is no more than an intelligent guess.

Another famous rational automaton was J. N. Maskelyne's whist player Psycho, which was first exhibited in 1875. There was never any suggestion here that it was a true automaton and the interest lay wholly in the secret of the control. Psycho was a mechanical figure considerably less than life-size, dressed in oriental costume and seated crosslegged on a box or pedestal, Inset, Nos. 13 and 14. Doors gave access to his interior, which was packed with mechanism, and he was subjected to the closest inspection. He was then placed upon a large transparent glass cylinder set on end which apparently isolated him from all external control. His cards were placed in a rack in front of him and when his turn came to play he would select the best one available to him, taking hold of it between the finger and thumb of his right hand, holding it up for all to see before putting it down in front of him. The movements were imparted to the automaton by a confederate off-stage using a pneumatic tube control of which the glass cylinder in full view of the spectators formed a link. It was a consummate piece of magical misdirection on the part of Maskelyne.

There appear to have been two versions of Psycho. In the earlier one the automaton's thirteen cards were arranged in a quadrant so that as the arm turned about a vertical axis through the shoulder the right hand traversed an arc immediately above the cards. A solution giving details of a clockwork mechanism which would enable the above movements to be made was first published in 1877, and Fig. 20 is taken from this. When the confederate varies the air pressure a bellows inside the pedestal on which the figure is placed responds and the clockwork which moves the head and arm is set in motion. In the later version, which is the one illustrated in Inset, Nos. 13 and 14, the cards are arranged in three rows and the arm movement is correspondingly more complicated. No clockwork is used, the

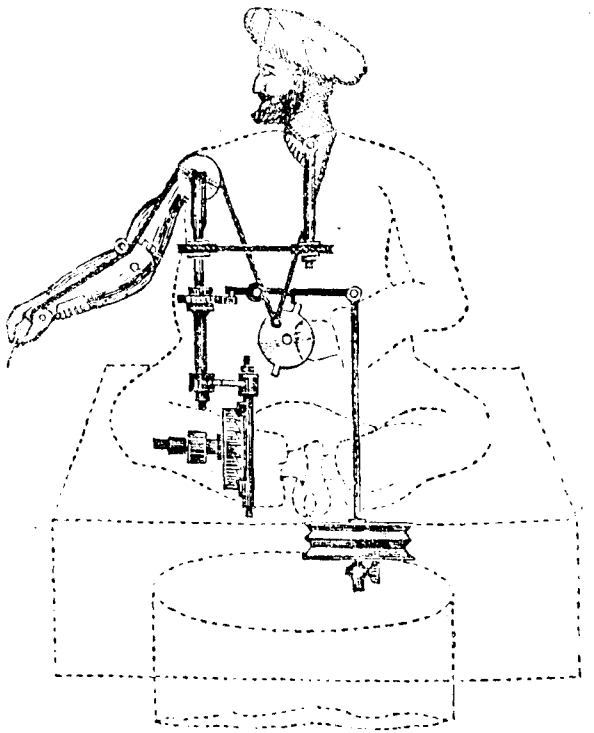


Fig. 20. A mechanism for Psycho. Extension of the bellows under an increase in air pressure operates a lever which (*a*) arrests the mechanism responsible for swinging the arm to and fro over the quadrant of cards, and (*b*) releases the mechanism which closes the thumb on the card and raises the arm. The clockwork drives for these mechanisms are not shown. (Taken from *Ancient and Modern Magic*, by Arprey Vere, George Routledge and Sons).

motion of the bellows themselves effecting the necessary and diverse movements of the hand and arm. The earlier Psycho could perform arithmetical calculations, displaying the results by disclosing with the left hand numbers contained within a small box. The later Psycho smoked a cigarette instead. Psycho

proved immensely popular and had a number of imitators in England and abroad. After a very long retirement, Psycho gave a special performance at St George's Hall in 1910. In 1934 the Maskelyne family presented him to the London Museum.

Charles Babbage, who invented the difference engine, a machine which can calculate and print mathematical tables, and who spent the greater part of his life attempting to construct an analytical engine which would be capable of solving mathematical problems, appears to have been the first to consider that machines could be devised which would take part in such games of skill as chess and draughts. In his self-revealing and entertaining autobiography *Passages from the Life of a Philosopher*, which was published in 1864, Babbage recounts how he sketched out the mechanism which would play the schoolroom game of Noughts and Crosses or Tit-Tat-To, and how he toyed with the idea of exploiting it as a sideshow with the object of obtaining further funds for his more serious work. In 1874 a paper was read before the United States National Academy by Dr Rogers, who also indicated how an automaton Tit-Tat-To player might be constructed. Neither Babbage nor Rogers ever got as far as constructing a machine and it is doubtful if either of them worked out a complete design.

Quite recently Mr D. W. Davies (the author of *A Theory of Games and Noughts and Crosses*, page 40) of the National Physical Laboratory has built an electro-mechanical Noughts and Crosses playing machine. This machine, which was first shown at a Royal Society soirée in May, 1949, uses standard automatic telephone equipment and indicates its play by an illuminated cross on the appropriate square. The machine will take advantage of a bad move on the part of its adversary; otherwise the usual draw will result. In this simple game with its very limited number of positions, the correct reply to any situation may be contained in a few definite instructions. Even so the mechanism required to implement these instructions is quite elaborate. The machine is illustrated in Inset, No. 16.

In 1914 Sr Torres y Quevedo, Director of the Laboratorio de Automática in Madrid, exhibited in Paris a chess-playing machine

which could play a simple end game checkmating an adversary's King with its own King and Rook. Its operation was made possible by the fact that in the particular end game chosen any one of its opponent's moves could be countered by one of five basic moves on the part of the machine. An additional provision enabled the machine to verify that its adversary had made a legitimate move, and to cease play if this were not the

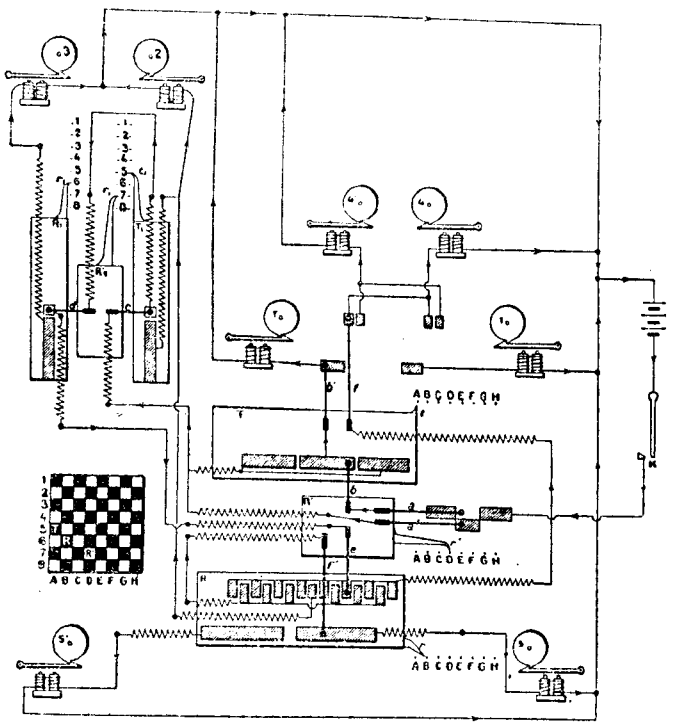


Fig. 21. Circuit schema of the chess-playing machine of Sr Torres. The two pairs of slides linked with the machine's King and Rook are lettered R, R_1 and T, T_1 respectively. Its opponent's King is coupled to the pair of slides R' and R'_1. (*Reproduced from La Nature*, 1914)

case. Inset, No. 15 shows the machine, which was electro-mechanical in operation. Each piece was linked to a pair of slides which recorded its horizontal and vertical position. These slides carried electrical contacts which determined the move to be made by the machine. The circuit scheme is given in Fig. 21, and is practically self-explanatory. It does not, however, show the ancillary circuit arrangements. For those who are interested the following table giving the basic moves is appended. Two zones are considered on the board (Fig. 21), one on the left formed by the three columns A, B, and C, and the other on the right comprising the columns F, G, and H. The numbered discs in Fig. 21 represent the motors which when released move the men the required number of squares.

The black king					
is in the same zone as the rook	is not in the same zone as the rook and the vertical distance between the black king and the rook is				
	greater than one square	equal to one square. The vertical distance between the two kings being			
		greater than two squares	equal to two squares and the number of squares which separates them horizontally is		
			odd	even	zero
The rook moves horizontally the extent of the board	The rook descends a square	The king descends a square	The rook moves horizontally one square	The white king moves one square towards the black king	The rook descends one square
1, 1'	2	3	5, 6	4, 4'	2

(The numbers refer to the numbered motor discs in Fig. 21)

Professor Norbert Wiener of the Massachusetts Institute of Technology has recently suggested how an automaton chess player might be designed which would be capable of playing a complete if somewhat indifferent game. The machine would be arranged to play, privately, all its own admissible moves together with those of its opponent for two or three moves ahead, and to assess each sequence of moves according to some agreed conventional valuation of basic situations. The machine would then

select and play that move which promised it the most advantageous sequence. The more extended the examination the better will be its play, but the time taken for a machine to explore a game for more than a few moves deep would be prohibitive.*

Sr Torres and Mr Davies have been able to arrange for their machines to recognise all possible situations and to provide them with instructions as to the correct behaviour. Where the complexity of the game is such that this is impossible, the method suggested by Wiener will have to be used. The first method is similar to that used by a man who is working a routine in which all the eventualities have been experienced before and the correct course of action found and learnt by heart so that it has become second nature, reflex in fact. The second approach is the trial and error method often used in playing a game and in solving a problem.

In common with Wiener a number of people see in modern computing machines – those electronic and electromechanical realisations of Babbage's analytical engine – an approach to a model of the human brain. These machines which can perform such prodigious feats in arithmetic have elements and functions analogous to those found in a nervous system. They can be programmed to play a game of skill by the analytical method referred to above, and they possess a memory mechanism which enables them to retain in a usable form data required during the course of a computation.

Although the computing machine can surpass the human brain in the mechanical accuracy of its logical processes and in the speed with which it can carry them out, it is a poor instrument by other standards. Lacking in initiative it does only what it is told and that by a rigidly prescribed method. While a machine could be constructed to show in some degree the phenomenon of the conditioned reflex, and thus capable of 'learning', the use

* Since the above was written C. E. Shannon of the Bell Telephone Laboratories has published a paper in the *Philosophical Magazine* for March, 1950, 'On Programming a Computer for Playing Chess', in which he indicates how by modifying the strategy the time of play might be substantially reduced.

it could make of this faculty would depend entirely upon the provisions made by its constructor. It is difficult to see how it could be capable of original thought.

READING LIST

The classic book on the automaton, *Le Monde des Automates*. (1927) by Alfred Chapuis and Édouard Gélis, has long been out of print, but Chapuis in collaboration with Edmond Droz has recently published *Les Automates, figures artificielles d'hommes et d' animaux* (Neuchâtel, 1949). Adolphe Blind has written a small book on the trick automaton, *Les Automates Truqués* (Geneva, 1927). Professor Wiener's *Cybernetics* (New York, 1948), is very relevant when the scientific aspect is being discussed (see page 124 of *Science News* 13).

RESEARCH REPORT

A. W. HASLETT

PENICILLIN AND BACTERIA

THE problem of killing bacteria in a living human body is to find substances which are more poisonous to bacterial cells than to the cells of an animal. There is no difficulty about killing either by itself. Concentrated sulphuric acid, for example, will do an excellent job. But differential killing is much more difficult; because, on the whole, the resemblances between one kind of cell and another are more striking than the differences. For example, several of the *B* group of vitamins were first recognised as vitamins because they were needed as 'growth factors' by bacteria. Then, there is the difficulty of killing cancer cells without killing normal cells, although in their uncontrolled multiplication they behave very differently. An illustration of another kind is the fact that there should seem to a biochemist to be sufficient resemblances between the chemistry and organisation of cells of different kinds to make it possible for him to choose *The Cell* – in general – as the title for an address. This was done by Professor R. A. Peters at the 1949 meeting of the British Association.

With this background, there is a real and serious interest in discovering how it is that substances such as the sulpha drugs and penicillin do their work. In the case of the sulpha drugs, there is a clear-cut answer. They are effective because they resemble in chemical structure and shape another substance, *para*-aminobenzoic acid, which is needed by bacteria. Crudely expressed, the bacteria are 'fooled' into accepting something which is useless to them in place of something they need; much as house construction could be effectively wrecked if builders could be induced to use flour instead of cement in mixing their mortar. There is no such simple picture in the case of penicillin – or none, at least, which is satisfactorily established. In fact, there has been doubt until lately as to whether penicillin acts

at the cell wall of the bacterium, by blocking the entry of some other substance or substances, or whether it penetrates the bacterial cell and does its work inside. What follows is an account of the first conclusive proof that its effect is inside the cell. It is of interest for its own sake as illustrating the point that two different methods, when used in conjunction, may give a decisive answer to a question which neither, separately, would be able to settle. It resulted from a pooling of techniques between different laboratories.

Dr P. D. Cooper and Dr D. Rowley of St. Mary's Hospital, London, had been working earlier with penicillin 'labelled' with radioactive sulphur which had been prepared by Dr E. Lester Smith of the Glaxo Laboratories. Penicillin has not yet been synthesised on any appreciable scale, and it is necessary, therefore, to use the *Penicillium* mould to prepare the radioactive form of penicillin. The procedure is to culture the mould under conditions such that the supply of sulphur available to it is a limiting factor in its growth. Then if extra sulphur is added which is, in part, radioactive, the penicillin which the mould produces will be radioactive too. It is then possible to follow the later movements of this penicillin by the radioactivity which it shows, and this method had been used without further addition at St Mary's Hospital. This is the 'tracer' method. Meantime, at the National Institute for Medical Research, also in London, Dr I. M. Dawson and Dr E. J. Elford had developed a method for obtaining relatively pure suspensions of bacterial cell walls, separated from the internal substance of the bacteria. This they did by mechanical shaking with very small glass beads. As a method of cell wall preparation this was a new development. Dr Dawson is an electron microscopist and he had been using the method, with Dr Elford, for a study of the fine structure of the cell walls of bacteria under the highest magnifications.

The two methods fit well together. The use of radioactive penicillin enables its presence to be identified, and the amount present to be calculated. The mechanical separation of bacterial cell walls makes it possible to distinguish between penicillin which has adhered to the cell wall and that which has penetrated

the cell – which, it will be remembered, was the point to be settled. One would not expect that each individual bacterium would be separately caught and broken by the glass beads, so that the division into cell walls and internal substance was literally complete. But this does not matter, since electron micrographs can be used to show the proportion of bacteria which have been disrupted in a particular experiment. The procedure, then, was as follows. The bacteria were exposed to radioactive penicillin. They were then washed twice to remove any surplus penicillin which had not either penetrated the bacterial walls or was adhering to them. They were then shaken with the glass beads, and separated mechanically into two portions. In one were the cell walls of those bacteria which had been broken and complete bacteria which had not been broken. In the other was the internal substance of those broken. The radioactivity ('tracer') method was then used to show the division of penicillin between these two portions; and direct counts were made from sample electron micrographs to give the proportion of bacteria which had been broken. On the supposition that all the penicillin had been inside the bacterial cells and none on the cell walls, it would be expected that the proportion of bacteria which had been broken would be the same as the proportion of radioactivity found in their outpoured substance. In three experiments, the proportion broken was 92 per cent, 60 per cent, and 92 per cent, with an uncertainty of 2 per cent in the count in each case. In the same three experiments, the proportion of radioactivity in the outpoured substance was 93 per cent, 59 per cent and 92 per cent, with an uncertainty of 5 per cent in each case. It would be possible that a relatively small amount of penicillin could have been on, or within, the cell walls. But there is no other explanation possible, from these figures, than that its main activity is inside the cell, and not on it.

Cooper, Rowley and Dawson. *Nature*, 1949, **164**, 842.

ANIMAL VIRUSES

The viruses which cause diseases in animals and human beings are thought to be more complex, both chemically and in organisation,

than those of plants. Some, but not all, plant viruses have been prepared in crystalline form; and, if there were a satisfactory test for the purity of a protein, it might be possible to show definitely that the individual plant virus *is* a protein. At present this remains a probability which cannot be proved. In the case of animal viruses, there is evidence from electron micrographs that the particles of vaccinia virus show some form of structure – that they are not the same throughout. Two of the pictures which have been obtained were reproduced as Inset, Nos. 30 and 31 in *Science News* 9. Their appearance suggested that this particular virus has a more dense centre, an outer region which is removed by pepsin (the protein-digester of the stomach), and some form of 'membrane' which surrounds the whole. But although an animal virus may have structure to this extent – and has in the case quoted – it is not self-supporting in the sense that an organism is self-supporting: there is no instance in which any kind of virus, animal or plant, has ever been reared or cultured except in the tissues of its host. It is supposed therefore that virus particles are only reproduced by interaction between those which already exist and the cells which they invade; in other words, that the multiplication of the virus and the damage which it causes to the cell are directly connected. If so, it is obviously important to study the process by which the virus multiplies – either directly, or through the sequence of damage which it causes in the invaded cell. The purpose of the present note is to record two recent examples of these two approaches.

The first arose out of work on the influenza virus, jointly carried out by Dr C. M. Chu of the World Influenza Centre, lately established at the National Institute for Medical Research in London, and by Dr I. M. Dawson and Dr W. J. Elford of the Institute's normal research staff. The significance of Dr Chu's contribution is that the great majority of work on the influenza virus – of which there are two main strains – has been on strains which have been maintained under laboratory conditions by repeated inoculations from one fertilised hen's egg to another. In other words, the strains of virus which are used are some long way removed from natural epidemic conditions. Both of these

laboratory strains appear typically in electron micrographs as small round bodies, of the order of one ten-thousandth of a millimetre in diameter, the particles of the 'B' strain being about 10 per cent larger than those of the 'A' strain. In 1946 V. M. Mosley and R. W. G. Wyckoff, of the United States National Institutes of Health, noticed in micrographs of both strains that, as well as the usual round particles, there were sometimes but not always a few of rodlike shape. But because their appearance was unpredictable, their identity as virus particles remained open to doubt. What has now been found is that in virus from an epidemic, the rod-like form is about as common as the round particles. The discovery has resulted from the combination of the 'collecting facilities' provided by the World Influenza Centre with a new technique of examination developed at the National Institute. Virus has been obtained from influenza cases in Italy, France, Holland, Czechoslovakia and even from Ocean Island in the Pacific; and it is through electron micrographs of the Paris and Dutch strains of 1949 that the discovery was made. As to the method of examination, G. K. Hirst of the Rockefeller Institute had found earlier that red blood cells were readily agglutinated or stuck together by certain viruses, including those of influenza, fowl plague and mumps. It therefore occurred to I. M. Dawson and W. J. Elford at the National Institute that red cells might offer a convenient means of collecting virus particles for examination, provided that the cells themselves could be made sufficiently transparent for electron microscopy – for a normal, intact cell is too thick and dense for this purpose. They have therefore used saponin – a form of natural soap – to dissolve away most of the cell's internal substance, leaving only its nucleus and a collapsed membrane, to which the virus particles will still adhere. The result was to enable larger numbers of virus particles to be 'photographed', and the reality of the rod-like forms, noted earlier by Mosley and Wyckoff, was quickly confirmed. The first of the 'epidemic strains' to be examined came from Holland, and it was at once seen that rod-like forms were very much more numerous. Later, virus from this same strain was passed through the tightest-meshed of graded

membranes likely to allow the passage of the ordinary round particles; and this material, when inoculated into hen's eggs, yielded the rodlike form as well as the round, and continued to do so even after several successive inoculations. The implication seems to be that either form can give rise to the other, and that the rod-like form is connected with the multiplication of the virus – which, under epidemic conditions, would be unusually rapid.

The alternative approach, through the changes produced by the virus in animal cells, has been followed by Dr G. M. Findlay and Mr M. Grigg, using normal optical microscopy. The virus was that of poliomyelitis or infantile paralysis, and some of the pictures which they have obtained were shown at a recent meeting of the Royal Microscopical Society in London. As with influenza, there is more than one type of this virus. With poliomyelitis four main types are recognised, all of which are transmissible through monkeys, but only one through mice, which were the experimental animals used in the present research. In addition there are two further groups of viruses which produce similar paralysis in a human being, but are not otherwise related. The experimental work was carried out by Dr Findlay in the research laboratories of the Pharmaceutical Society in London, the microscopy by Mr Grigg, who is an amateur microscopist.

The virus being invisible in the optical microscope, there is an interval between inoculation and visible changes, and only a few cells are at first affected. This interval is not more than 12 hours, and microscopic changes come earlier than symptoms. The sequence of change is from the centre of the cell outwards. The first change is in the nucleolus, which can be described as the 'nucleus of the nucleus'. It is a clearly defined region, roughly spherical in shape, and in a nerve cell occupies an appreciable volume in the larger nucleus of which it is part. It is surrounded by a layer which can be distinguished by staining with a suitable dye, and consists of chromatin, a complex of protein and nucleic acid. At 12 hours from infection, possibly less, the nucleolus begins to appear larger and less dense. Later,

the surrounding chromatin layer is seen to be disturbed and bits of it break off. One particularly fine micrograph showed what seemed to be a pear-shaped droplet, as if in the act of detachment. Smaller 'bodies' appear to move outwards from the nucleolus across the surrounding nucleus. One, it was suggested, had been caught passing through the 'membrane' which surrounds the nucleus. Further changes take place in the outer part of the cell, and finally the whole cell disintegrates.

There is a double interest in these pictures: first, in relation to the interaction between virus particles and cell, which is the main theme of this note; and, second, in relation to the natural course of the disease. The interest in the case of the first lies in the connection which is suggested with the theory of Professor Caspersson of Stockholm that the building up of proteins within the cell nucleus is by successive stages which can be localised; and that virus infections act by interference with some stage in this process. Professor Caspersson places the first stage in the building up of proteins at the boundary of the nucleolus – which is where these microscope pictures show the first physical disturbance. This could be coincidence, but at least is suggestive.

There are also two points which may be noticed in connection with the natural disease. The first is that, by exercising the mice, it has been confirmed under controlled conditions that exercise after infection hastens the progress of the illness. This agrees with the conclusion reached in 1947 by Dr W. Ritchie Russell, neurologist to the Radcliffe Infirmary, Oxford, from the analysis of 42 cases for which full histories were available, and since extended by Dr Russell to 100 cases. The second connection depends on relatively late changes in the outer part of the cell – the existence of which has so far been mentioned only in general terms. Here, outside the nucleus, there is a progressive stratification of certain angular-shaped bodies which are seen in the normal cell and which can also be reduced in numbers merely by repeated stimulation of the nerve cell. This is what happens during exercise, and may help to explain the accelerating effect which exercise produces: for the cell would then have to work

harder to replace these bodies, and would be more vulnerable to interference by the virus.

Chu, Dawson and Elford. *Lancet*, 1949 i, 602.
Dawson and Elford, *Nature*, 1949, **163**, 63.

THUNDERCLOUD GROWTH

Preliminary results of a research programme on the structure of thunderclouds, carried out over a period of years by the U.S. Weather Bureau, in conjunction with the U.S. Air Force, Navy, and Committee for Aeronautics, have lately been summarized by Dr H. R. Byers, of the University of Chicago and the Weather Bureau, who directed the project. Earlier electrical measurements, also made in the United States, and obtained during a series of aircraft flights through cloud, were mentioned in *Science News* 10. This has been both a more substantial project, and more fundamental in its approach; its aim being rather to establish the 'life history' of a thundercloud in terms of air movements. To some extent, it has been anticipated in publication by a smaller British investigation in which radar observations from a station at East Hill near Dunstable were combined with observers' reports from co-operating aircraft supplied by the Royal Aircraft Establishment, Farnborough. The coincidence is natural enough: for once it had been discovered, as it was during the war, that tropical rainstorms showed a substantial echo on centimetre-wave radar sets, it was inevitable that the same method should be applied to thunderstorms. What follows is based mainly on the American investigation, as being the more complete.

Measurements were made in the air during 1,363 flights. They were primarily of turbulence, up and down draughts, and temperature. The standard procedure involved the simultaneous flight of five aircraft through the same storm at heights of 5, 10, 15, 20 and 25 thousand feet. On the ground there were extensive networks of recording stations. The first, established in the summer of 1946 in Florida, consisted of some 50 stations, spaced one mile apart. In the following summer an even more extended network was set up in Ohio, but with the stations two miles

apart instead of one. Quick-responding equipment provided for the measurements of wind gusts and direction, moisture, rainfall and pressure, with readings in most cases available to the nearest minute. In and around these networks there were arrangements for sending up balloons equipped with automatic recording equipment and miniature radio transmitters to report the readings which they obtained, while on the ground there were sufficient radio direction-finding stations to fix the positions of the balloons. In addition, and outside the main network, there were four wind-finding stations, equipped with target balloons, which were followed by radar. Finally, use was made of long-range radar stations to watch the development of storms, to select in advance those thought suitable for study, and to do so with sufficient warning to set the whole elaborate machinery into operation. No one of these procedures was unique to this investigation. What is outstanding was the scale on which they were used, and the co-ordination which enabled the results from so many different sources to be effectively combined.

The essential result is the reduction of the history of a thunderstorm to that of a series of 'cells', each of which is in effect a separate cloud. As an indication of the size of the 'cells', it is stated that in a typical Florida storm, of some 20 miles length and 5 to 8 miles width, there would be 6 or so recognisably separate cells, usually of oblong shape and 3-7 miles across; and, between these, there would be smooth intervals of a mile or less in extent, as shown in aircraft records. The same pattern was shown also in rainfall distribution on the ground. No radar pictures giving a direct view of this structure were included in the initial review which has been published. An example, included in the British publication, showed a cloud divided into two main portions, the nearer of which extended to a height of some 23,000 feet, and the more distant to 40,000 feet. They appeared to be joined at the base, and up to a height apparently of about 17,000 feet, and the taller portion was further subdivided into half-a-dozen or so less clearly defined fingers. A second cloud was shown similarly divided into three main sections, each of which appeared subdivided. With less certain

correlation, it was suggested that the peaks were associated with up-draughts, and the gaps with down-draughts. But on this point, the American picture should be the better substantiated.

Dr Byers' starting-point was the general observation that in a single storm there were usually examples of different types of cell movement which, on more detailed analysis, were found to follow a life cycle which was broadly similar. This seems to have been the critical stage: the transfer of attention from the evolution of the storm to the evolution of the cell which formed part of the storm.

The beginning, as might be expected, is the sudden upwards movement of warm air due to low-level instability. This is described as the 'cumulus stage', and the cloud top is then seldom higher than 15,000 ft. initially. Up-draught is now general throughout the cell, and at rates varying from a few feet per second to nearly one hundred feet per second.

The transition to the second or 'mature' stage begins when the top of the cloud has reached higher than freezing level – this being the necessary condition for the initiation of rainfall. Drops fall through the weaker portions of the up-draught; and, as they increase in number and size, air is dragged along with them, and in that region the up-draught becomes a down-draught. At heights above about 25,000 ft. there is still an up-draught throughout the cloud; but, below this, the up and down movements proceed side by side. Rain at the surface is associated mainly with the down-draught portion, and the accompanying air is responsible for the principal storm gusts at ground level. Balloon records show that this is also true at heights of up to 2,000 to 4,000 ft. By comparison, winds due to in-flow from surrounding air into up-draught regions were of smaller strength. In terms of temperature, the up-draught is warmer, and the down-draught cooler, than is the surrounding air; but the first to a lesser extent than might be expected. This is because drier surrounding air is sucked into the stream, so causing evaporation and consequently cooling.

The third stage of the cell's history is one of collapse. This comes about through the gradual encroachment of the down-

draught area on that of upwards movement. The flow of rising air to freezing level is reduced, and in the end cut off; and with the down-flow continuing, the cell collapses.

The greatest problem, Dr Byers states, is to explain in detail the formation of new cells. One situation which seemed to be favourable is where two existing cells are of the order of 2–3 miles apart, and with their respective down-draughts meeting in opposition. A second, which seemed also to be favourable, is on the downwind side of an existing cell. Given some existing instability, the new upwards movement would be precipitated in either case by the cooler down-draught cutting under what was initially the over-warm air of the surface layers. But there is an unexplained tendency for new cells to grow in clusters almost simultaneously.

Further points are that the greatest turbulence was at 15,000 to 20,000 ft. and that both up- and down-draughts were strongest at this level. Finally, there is the suggestive comment that both lightning and hail – with the latter presenting obvious frictional possibilities – were at their strongest at the same height, about 15,000 ft.

Byers. *Science*, 1949, **110**, 291.
Jones. *Nature*, 1949, **163**, 728.

WINDS IN THE UPPER AIR

At higher levels than those of normal meteorology, wind speeds have had to be measured or estimated by a variety of methods, and the results correlated as well as might be. One of the oldest of these methods, systematised in England under Dr F. J. W. Whipple, of Kew Observatory, is from the travel of sound waves from distant explosions. But the effects of temperature are involved as well as of wind-gradients, which makes the method complicated. A more direct procedure is to observe visually some form of 'drift' in the upper atmosphere. This may be either of the smoke from smoke-shells, or of the trails left by meteors, or of luminous clouds seen at night. From smoke-shells fired over southern England in 1944-5, Sir N. K. Johnson deduced an average winter movement of about 130 km. per

hour (80 m.p.h.) from west to east at a height of 30 km. (19 miles), with a summer average of about 43 km. per hour in the reverse direction. This agreed broadly with Dr Whipple's estimates. Then, at Sonneberg Observatory in Germany, Dr C. Hoffmeister has collected information from the drift of meteor trails and luminous clouds as opportunity arose over a 13 years period. He concluded that east was the predominant wind direction up to about 80 km. At 120 km. height, winds were mainly from the south-west in summer, and during the winter from either south-west or north, with a winter maximum of perhaps 950 km. per hour. But although a good deal of information has been obtained in the aggregate by these different methods, there is an unfortunate resemblance to the advice given by the maker of billiard balls, beginning: – 'First catch your elephant'. For any visual observation of rate of drift, it is necessary first to have something visible the rate of drift of which can be watched. The height of balloon ascents, which provide one obvious answer, has lately been increased, and it is hoped in the United States to make regular observations within the region 30-45 km. A further new method – which will be the main subject of this note – depends on radio measurements and promises regular information about movements in the E Region of the ionosphere at about 100 km. height. This, too, is in a sense a drift method. It has been described by workers at the Cavendish Laboratory, Cambridge, arising out of investigations of which the primary purpose was to study local horizontal variations in the structure of this region. They were described by J. W. Findlay at a conference held by the Physical Society.

In one type of Cambridge experiment, short pulses of radio transmission have been directed vertically upwards, and the strength of the reflected signals compared at different receivers placed relatively close to one another on the ground – for example, at three receivers, placed 100 metres apart in the shape of an 'L', so as to give both a north-south and an east-west separation. The use of pulse transmissions is one of the classical methods used in research on the radio-reflecting regions of the atmosphere. The study of local variations at ground level in the

return signals, and the deduction from these of variations in the reflecting region 100 km. above, is a new development. Measurements have been made over a wide range of radio frequencies, from 7 megacycles per second down to 16 kilocycles, and Mr J. A. Ratcliffe, head of the Radiophysics Section of the Cavendish Laboratory, stated that the picture thus obtained was self-consistent. The deduction of wind-speeds, described by Mr Findlay, depends on the ability to detect 'drift' in this pattern of variation. There are random movements, as between one centre of activity and another, and also a uniform drift which is common to them all. One record which he displayed showed a wind-speed of 80 metres per second, or just under 290 km. per hour. Others, using a different method, showed speeds of 20-140 metres per second. As an indication of seasonal variation, it was stated in discussion that throughout March, 1949, at 11 a.m., the predominant wind-speed was 100 metres per second (360 km. per hour) from east to west. But in June at the same time the wind was from west to east.

Such deductions of windspeed arose as a by-product from the original radio investigation, and there is an interest in the horizontal structure of the ionosphere for its own sake. From this point of view, the Cambridge measurements have been of local detail. Evidence of more extended variations has been obtained by Mr W. Ross of the Radio Division of the National Physical Laboratory, and in the general picture which they suggest these lines of research fit well together. Mr Ross's approach was through the apparent bearing errors found in radio direction-finding at high frequencies, and most of his observations refer to the F Region of the ionosphere which is at a daytime height of some 235 km. The procedure was to take frequent radio bearings, usually at half-minute intervals, on known and stationary transmitters – which in different experiments were at Daventry, Start Point in South Devon, and Fraserburgh in northern Scotland. The direction-finding equipment was either at the Radio Research Station, Slough – distant 90 km., 260 km. and 700 km. respectively from these transmitters – or at other places within the near neighbourhood of the Slough

station. There were, in the first place, the usual rapid variations in apparent bearing with which practical radio operators are familiar. These are put down to local changes in activity in the reflecting region, so that there is a connection to that extent between the Cambridge work and that at Slough. But if these more rapid changes are averaged out, and uncertain bearings omitted which are obtained during fading, then there remain slow fluctuations in bearing which last from a few minutes to half an hour. These fluctuations follow the same course when radio bearings are taken, either between the same two points but on different frequencies, or on a single transmitter but with the receiving stations up to 10 km. apart. Such effects could be produced by a tilting of the F Region by some few degrees from the horizontal, and with the same angle of tilt extending over distances of 10 km. or more in horizontal extent. The suggestion, therefore, is that in its more extended contour the reflecting region should be likened to 'an ocean with a comparatively slight long swell'; and superimposed on this are the more rapid and local variations in activity which have been studied at Cambridge, and which are giving new information on wind speeds in the upper atmosphere.

Review article: Shepherd. *Science Progress*, 1949, **37**, 488.
Direction-finding: Ross. *Radio Research Special Report, No.* 19 (H.M.S.O.)
Cambridge work: Mitra. *Proceedings Institution of Electrical Engineers*, Vol. 96, Pt. III, p. 441

CRYSTAL IMPERFECTIONS

New methods for the measurement of irregularities in crystal structure have been developed by Professor A. Guinier and J. Tennevin at the Laboratoire d'Essais of the Conservatoire National des Arts et Métiers, Paris. Like other methods of the kind, they make use of X-rays; and, although no more than an indication of results will here be given, both are in fact a development of the earliest method used in X-ray analysis: that employed in Munich in 1912, on the suggestion of M. von Laue.

The first of these new methods records the extent to which different but corresponding layers of atoms in a crystal depart from a strictly parallel arrangement. It measures the direction of individual atom layers to within about 10 seconds of arc (or, in round figures, 0.003 degrees), so that a divergence of about 30 seconds can be detected. It is as if a set-square were applied to successive layers of atoms, and the sum of all the observed directions recorded as a pattern of lines. For a perfect crystal, with all corresponding layers exactly parallel, there should be only a single direction; and this should be perfectly sharp within the limitations of the method, that is to within 10 seconds. For a real crystal, there may be either a broadening of the indicated direction, showing that individual layers depart in a more or less random way from the average, or a number of distinct directions, showing that there are small but more or less regular variations of direction. The advance, compared with existing methods, is that the kind of irregularity which is being recorded is precisely known. The usual idea is that such irregularities would take place in definite jumps; corresponding, in Professor Guinier's method, with pictures showing a fine pattern of separate lines. In practice, he finds the continuous types of variation to be more usual – although the type of irregularity, as well as its extent, depends on the conditions of preparation of the crystal. The second new method provides shadow pictures in which areas of irregularity (inclusions) appear as white patches on a dark background, the main outline being that of the physical shape of the complete crystal. Professor Guinier has been awarded the Charles Vernon Boys Prize of the Physical Society.

Guinier and Tennevin. *Acta Cryst.*, 1949, **2**, 133.

GROUP DYNAMICS

GORDON RATTRAY TAYLOR

ERNST CASSIRER once observed that the growth of a science often occurs through a change in the nature of what we regard as 'really existing'. There could be no research on atoms until the idea that they existed was accepted. There could be no science of psychology until will and emotion were accepted as objective phenomena and not just colourful ways of speaking. This is particularly true of social psychology, which is still fighting for recognition of its principal concepts, and is busy trying to formulate additional ones.

Possibly the most fruitful concept in this field is that of the *group*, and the words *group* and *group-dynamics* now appear with great frequency in the technical literature. Nevertheless, outside a small professional field, the idea that the behaviour of a small group of people is anything more than the sum-total of their individual behaviour is scarcely given consideration. To take a specific and rather crude instance, the popular assumption is that if you take ten shy people and form them into a group, it will be a shy group. As a matter of observed fact, this is not true. In certain circumstances, ten normally shy people may be remarkably forthcoming when grouped together.

GROUP PHENOMENA

Perhaps it will be as well, therefore, to start by naming some of the phenomena which require explanation, more or less at random. I have already mentioned one, the fact that people behave quite differently in some groups from how they behave in others: a common domestic example of this is the way a young child will sometimes be difficult and disobedient at home and good as gold at school. A second would be the way groups of the committee type sometimes decide to discuss a particular

We can take the thing a stage further by considering the classic research of Thrasher, who studied 1,313 gangs of adolescents in Chicago in the '20's, and the more recent work of Whyte in the same field. Typically, these adolescents were boys who had failed to become absorbed into some satisfying adult group and had therefore created their own. Briefly, what they needed was a group which would sponsor the sort of activities adolescents find interesting, and in which they could win prestige and acceptance for their skill at such activities. Consequently, Thrasher concluded, it was useless to try and wean boys away from their gang: they needed it too badly. The only thing was to induce the gang as a whole to renounce crime. Once it did this, the intense need of every member for approval which had formerly acted to make reluctant members participate in crime, acted to keep them up to the mark instead. (This is one factor in the effectiveness of group-decision.)

Thrasher was interested chiefly in the social evil. For the purposes of group dynamics generally, the relevant point is that people are much more concerned to retain the approval of fellow group-members, whom they continually meet face to face and who constitute their real psychological environment, than they are to please people outside the group. One needs some first-hand acquaintance with groups to appreciate to the full just how intense and continuous is this preoccupation of the group member with the approval of his fellow members.

HOSTILITY

The new boy on his first days at school spends his time trying to find out the unwritten and the written rules, discovering who it does not pay to offend, who is a sneak and who an ally. Until he has 'explored the field' in this way he feels extremely insecure. The danger of being caught doing what is forbidden is much less alarming than not knowing what is forbidden and what is not. The new member in any group seems to engage in a very similar exploratory process. He wants to know how he stands with regard to the other members, and this is true whatever the official purpose of the group. A man who joins a work gang or a

board of directors spends much of his time in discreetly testing the situation in this way.

This leads us to a more general inference about groups, which may perhaps be regarded as the second basic principle of group dynamics: *every member is always acting in terms of his whole schedule of needs and ambitions*, and not just in terms of his official purpose in joining the group. This statement is phrased rather carefully. For instance, it says 'his' whole schedule of needs, rather than 'man's' whole schedule, to allow for the fact that different people attach varying degrees of importance to different types of satisfaction. And it is deliberately vague about the system of priorities and combinations which determines how the need-schedule regulates behaviour. But before going on to consider it in more detail, it is perhaps worth stressing the radical nature of this assertion. We constantly assume that people are activated only by their official purpose. For instance, employers assume that workmen come to their factory exclusively to earn money and are astonished when they strike on some question of prestige. Employers, too, claim to be interested only in profitable and efficient production, but when it is suggested that this could be better ensured by giving employees a larger say in running the business, it becomes clear that many employers are not concerned to discover whether or not this is true, but are primarily interested in maintaining their autocratic powers. Occasionally, in committee work, for instance, we dimly suspect that Mr X has vetoed the proposal because it would reduce his influence, or because he 'likes being difficult', and we recognise that in certain situations people are motivated by the need to 'save face'. But we try and treat these as exceptions and cling to our belief in the conscious, rational, unemotional nature of behaviour generally.

The principle just enunciated declares that these 'unofficial' factors are always operative, and implies that we should catalogue them and discover exactly how they combine with one another and with official motives to determine behaviour.

In short, it is not too much to say that in almost every group a continuous covert struggle to fulfil needs is going on. Members

are constantly seeking for allies in this struggle and are attempting to undermine or discredit those who seem to be their enemies or who seem to deny them approval. But it is not only a struggle for approval and status. Members also seek outlets for their abilities, and satisfaction for private needs – such as a need for reassurance or a desire to dominate. Besides all this, they often import into the group anxieties generated outside it or use the group to work off hostility generated by events in other groups (such as their family) where they perhaps have less opportunity to assert themselves or to express aggression. These manoeuvres may be rather obvious, but frequently they are extremely indirect and by no means obvious to those who take part in them. At times they may be temporarily forgotten.

In studying the dynamics of the group the social psychologist therefore makes considerable use of the concepts of *hostility* and *support*.

ANXIETY

Another essential concept is *anxiety*, since this is the form in which people betray that their needs are about to be frustrated. We might then say that members of groups act so as to avoid anything which might increase their anxieties and try as far as possible to reduce them – a formula which enables us to dispense with an exact analysis of needs until the psychologist can give us one. But putting it like that brings us back to a problem I I have been postponing ever since I started discussing motivation: how far is this attempt to satisfy unofficial needs conscious?

The problem is partly a semantic one. It is now becoming accepted that the distinction made by Freud between conscious and unconscious was far too simple. There are several varieties of consciousness and unconsciousness. For instance, it has been shown that it is almost impossible to handle intellectually concepts for which we have no verbal label or mental picture. (In one study it was found that patients could easily distinguish, discuss and analyse a variety of kinds of stomach-ache, once they had been taught names for them.) Some of the work of the psycho-analyst consists in helping people to identify emotions which they feel but don't intellectually know they feel, and

providing verbal labels is one of the devices employed. Freud, on the other hand, generally meant by 'unconscious' material, material which was actively repressed or censored, and which could not be uncovered without obstinate resistance. Again, there is material which we can easily handle intellectually when we happen to think of it, but which is crowded out by other events. Probably, too, there is no rigid division between these categories.

So when we ask how far the group is conscious of the unofficial struggle for approval and status, the answer will have to be something complex like this. First from the viewpoint of individuals: some individuals will have censored their need and be unconscious of it; others will be conscious of it from time to time, though it may not be in the focus of their attention at the moment when they are making a bid for prestige. Second, from the viewpoint of the group: few members will be consciously aware of the deeper motives of others, not because of any censorship but chiefly because they have never conceptualised them.

This analysis of consciousness also throws light on the way in which official and unofficial needs interact. When the official purpose is strongly held in the focus of awareness, unofficial motives will exert little effect, but as attention weakens or wanders, they will become more dominant. As a rule, however, both will be at work simultaneously, and the resulting behaviour will be the outcome of both, rather like the parallelogram of forces we used to know at school.

One more gloss on this subject is called for. The more the official purpose of a group calls for a clearly-defined physical activity, the less time there is to worry about prestige or other unofficial considerations. (It may be that repetitive jobs, such as modern industry often provides, are an exception, since they require so little attention they leave plenty of room for wondering about such matters, and this may be an unappreciated factor in industrial disputes.) It is not surprising, therefore, that social psychologists find it easiest to observe anxiety in discussion groups and other groups of this type. This has led Bion to set up groups with no official programme whatever

(except a general desire to study groups) on the grounds that anxieties and hostilities will then rise to the surface and become more obvious, as indeed seems to occur. Therefore the objection which might occur to some, that the kind of groups the social psychologist uses are not typical, is not necessarily relevant, provided due allowance is made for the greatly diluted form in which these phenomena will occur in more mundane groups.

INTERACTION

Such are the principal forces in the group. Now let us see what kind of processes occur as a result.

Professor Mannheim used to relate the story of a number of intensely respectable ladies who volunteered to give their time in aid of the war effort and were asked, individually, to go and wash floors and dust in homes where people were ill – in plain words, to do charwomen's work. Each refused. But the organiser brought them together as a group and put the proposition to them again. After a little discussion they soon agreed to the proposal. Incidentally, this neatly illustrates the point I was making earlier about the greater importance of approval from one's group than from strangers. The ladies were prepared to risk losing prestige in the eyes of strangers but not in the eyes of one another. But if all agreed to the proposal, then of course no one would lose any prestige. However, my reason for telling the story here is to draw attention to the process by which they arrived at agreement. First one lady indicated a provisional willingness of an 'I will if someone else will' character. The others reconsidered their position, and one made a further slight move towards co-operation. The others re-assessed the situation – and so on, until by a series of small shifts they had moved from flat negative to unanimous positive.

It is a fairly safe guess that confession groups organised by movements such as the Oxford Group follow a similar pattern, one confession leading to another until finally all are involved. This kind of interaction is also present in scapegoating. We know from individual psychology that people readily displace their aggression on to innocent parties who cannot retaliate, as a

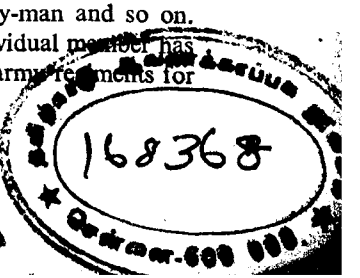
man who has been reprimanded by the boss snaps at his secretary. All we therefore have to explain is the process by which all the group members select the same scapegoat.

Until recently, this kind of process has always been very difficult to treat scientifically, but during the war it became of importance in connection with automatic pilots and A.A. gun predictors. Both these devices continuously reassess the situation and make adjustments to suit. This provided the impetus for the development of a new type of mathematical theory capable of handling such processes, which has recently been described by Professor Wiener in his interesting book *Cybernetics*. This theory should make it possible to analyse interaction in groups in due course.

PROCESS

These are short-term effects, completing themselves in minutes or hours, but at the end the group does not stand where it stood before. How are we to analyse the change of position which has occurred? We must do this in terms of a second set of concepts concerned with long-term changes in the group. (The word 'process' is used to label such developments). The attempt to analyse long-term changes has not been pushed very far, nor has the work of individual schools of experimenters been co-ordinated, but three or four concepts now seem fairly well established.

One of these is *culture*. As defined by the sociologist, the word culture indicates all the ideas, customs, institutions, techniques, traditions and beliefs we learn from the society into which we are born. The briefest definition which has been suggested is 'The learned reactions of a group'. Obviously much of the culture in a small group is identical with the culture of the society of which it is a part. But small groups also develop individual peculiarities: they use a special slang, have a tradition of good or bad work, and so on. In particular they allocate roles to members – one is regarded as the leader, another as the funny-man and so on. These traditions persist even when every individual member has been replaced by a newcomer, as we see in army regiments for instance. But they also grow and change.



A second important concept concerns the pattern of likes and dislikes which members develop for one another. This can be explored by the sociometric technique developed by Moreno and the changes in the pattern can be correlated with events in the group. This approach sometimes even provides a basis for predictions about individual behaviour: Moreno once correctly predicted that two children would run away from school when his enquiries had disclosed that they were liked by none of their schoolmates. Not to be accepted by one's group is an intolerable position, as I stressed earlier. Such patterns also change.

CONVERGENCE

The subject upon which most work has been done, perhaps, is the remarkable tendency of groups to arrive at a unified set of attitudes, a tendency for which the present writer has proposed the term *convergence*.

A particularly interesting experiment on the effect of group membership on attitudes was carried out by Newcomb at an American college. He designed a questionnaire which enabled him to rate the political opinions of students on a scale from 'left' or radical to 'right' or conservative, and administered it to a large number of students shortly after arrival and at intervals during the following three years. On arrival, almost all the students were conservative. The prevailing attitude at the college was radical. He found that all those students who were accepted into school groups adopted radical opinions, while those who remained friendless and isolated became even more conservative. In addition, those who were not quite sure of their acceptance became more radical than the norm; those who failed to win acceptance in the community reacted into extremer conservatism, while those who never sought acceptance remained unchanged.

This kind of experiment reveals something about the way the group affects the individual, but what the social psychologist really wants is information about the group as a whole. He can already rephrase his findings about aspiration levels in the form of a table showing the amount of 'scatter' about the norm in each group. What he would like would be some general index of the degree of scatter or convergence which would cover every type of attitude, opinion and ambition, and preferably one which could be measured continuously.

SECRET AGENDA

One of the difficulties in analysing group behaviour is that we are so often concerned with the individual that we tend to see things from his viewpoint. For instance, in therapy groups comprising people who meet regularly to discuss their psychological troubles, it is often noticeable that as each member recounts a symptom, the group will brush it unceremoniously aside and decline to discuss it. For instance, Mrs A will say: 'I had a peculiar headache yesterday while I was standing in a queue'. Mr B will observe caustically: 'I wish I had time to stand about in queues'. After a silence, Miss C will venture that she sometimes has headaches sitting at home, only to be met by Mr D's comment: 'I'd think myself lucky if I could sit at home.'

From the viewpoint of the individual, the group is being very unco-operative: but, as Bion has pointed out, it is equally true to say that in another sense the group is co-operating very effectively to suppress discussion of neurotic symptoms. There are, in fact, two ways of looking at a group, and for the social psychologist the latter is the more instructive.

Adopting this viewpoint, the observer of groups begins to see that groups sometimes show a determination to behave in a way which every individual member would repudiate. In the case just recounted, every member was prepared to declare his readiness to discuss neurotic symptoms. The fact remained, no one did. Thus there is a valid sense in which we can say that the group has a life of its own, apparently distinct from that of its members. The explanation is to be found in the principle discussed earlier, that people operate in terms of their whole schedule of needs, including those which we rather loosely call unconscious. In this therapy group people were anxious to avoid discussion of their difficulties because to discuss neurotic symptoms is always

productive of deep anxiety – this is a commonplace of psychoanalysis. But this resistance was, of course, unconscious.

This kind of phenomenon also comes up in a slightly less obvious form in groups of a quite normal kind, such as committees. The group will decide to discuss a certain subject, and then (as the observer notices) deliberately neglect its decision and discuss something else. This has led Main to speak of the 'secret agenda' of the group. Here again, the group has tackled the subject which really preoccupies it most, and is causing it anxiety, and not the one the members consciously feel it is their duty to discuss.

CONCLUSION

I started by mentioning that people often behave very differently in one group from the way they behave in another, and before concluding I should like to use the principles discussed to show why this is so.

In one of Thrasher's gangs the boys were regarded by their families and by society as failures, and at home accordingly behaved like failures, conscious of the scorn in which they were held. But in their group they were successes. They had their own standards of success, and it was success in the eyes of group members in which they were alone interested. So they behaved like successes and walked about with their chests stuck out.

But the matter may be more complex than this, as a further instance will show. In a boys' play-group, organised for children with emotional difficulties, the organiser bought a pair of boxing gloves and encouraged the boys to use them. A boy who had always been extremely inhibited and lifeless – his home life was such that all expression of aggression was forbidden – seized the gloves and was soon fighting furiously. Soon after this he became much more normal in behaviour. What has happened here? His behaviour has changed not just because of the atmosphere of the group but because the group provided him with an outlet which he needed, an outlet through which he could work off his feelings of hostility without forfeiting approval – indeed he even won approval by his prowess. In short, groups affect behaviour by virtue of the opportunities they offer or fail to offer. Boxing matches, like other institutionalised patterns of behaviour, are part of the 'culture' of the group, so we can begin to see why the study of the culture of groups is relevant.

This leads to a conclusion which may fairly be said to open a new epoch in psychological thought: that behaviour is not simply a product of personality, as has always been assumed, but is the product of interaction between personality and the culture of the group. The implications are far-reaching. To name but one, it means that if we wish to obtain higher output from industrial workers, we have to study not only individual motivation ('incentives') but also the dynamics of the work-teams and factory groups in which they find themselves.

The second thing which emerges which might also be called revolutionary is that there is a valid sense in which we can say that a group has a life – intentions, anxieties and patterns of behaviour – which is distinct from the life of the individuals composing it. Because of its association with mystical ideas of the 'group mind' conceived as a kind of supernatural entity, social psychologists have always shied away from conceptions such as group-life, group-mentality, group-behaviour and the like. But it is now quite clear that groups' activities *are* more than just the sum of the behaviour of the individuals which compose them, inasmuch as the group influences that behaviour in a way which is characteristic of that particular group.

What is now wanted is not only research into such specific problems as the dynamics of interaction-processes and the relationship between personality structure and group life, but above all an exhaustive analysis of a few representative groups in terms of all the known variables, to establish whether they are adequate to give a complete account of what goes on in the group.

If, finally, the question is asked: What use is group dynamics? one might reply, with Benjamin Franklin, 'What use is a baby?' but perhaps a slightly more specific answer is possible. The character of a group intimately affects its power of tackling the tasks for which it is met – to produce goods, to fly a bomber, to draft a bye-law. Anxious and insecure groups cannot get on with their jobs until they have resolved their more pressing

anxieties. Groups with bad traditions, defective cultures, divergent attitudes, do bad work, fail to elicit the full abilities of their members, find it hard to co-operate. It is only a little picturesque to say groups can be ill, and sometimes need a doctor. Already specialists of this type are emerging and in the U.S. there exists an organisation, supported by a Carnegie Fund grant, which tries to train people in first-aid for groups.

And beside these immediate applications, group dynamics provides the necessary foundation for an analysis of the behaviour of larger social groupings and of society as a whole.

SHORT BIBLIOGRAPHY

Asch, S. E. 'Studies in the Principles of Judgments and Attitudes: II: Determination of Judgments by Group and Ego Standards.' *J. Soc. Psychol.* (1940) **12**: 433.

Bion, W. R. 'Experiences in Groups I, II & III.' *Hum. Rel.* (1948) **1**: 314, (1948) **1**: 487 and (1949) **2**: 13.

Cattell, R. B. 'Psychological Dimensions of Groups.' *Amer. Psychol.* (1947) **2**: 335.

Cattell, R. B. 'Concepts and Methods in the Measurement of Group Syntality.' *Psych. Rev.* (1948) **55**: 48.

Lewin, K. 'The Dynamics of Group Action.' *Educ. Leadership* (1944) **1**: 195.

Main, T., and Nyswander, M. '*Some Observations on the Third National Training Laboratory in Group Development*,' July 1949 (roneoed).

Mayo, E., and Lombard, G. F. F. *Teamwork and Labour Turnover in the Aircraft Industry of Southern California.* Graduate School of Business Administration, Harvard 1944. *Business Research Studies No.* 32.

Murphy, T. M., Murphy, L. B., and Newcomb, T. M. *Experimental Social Psychology.* Harper Bros. 1937.

Newcomb, T. M. *Personality & Social Change.* Dryden Press, 1943.

Scott, J. F., and Fox, J. B. 'Absenteeism: Management's Problem.' Graduate School of Business Administration, Harvard 1943. *Business Research Studies, No.* 29.

Thrasher, F. M. *The Gang.* Univ. of Chicago Press, 1927.

Whyte, W. F. *Street Corner Society.* Univ. of Chicago Press, 1943.

Wiener, N. *Cybernetics.* Wiley, 1948.

THE CHEMISTRY OF PARALYSIS

H. O. J. COLLIER

IT is not uncommon that a substance is known to science long before an effective therapeutic use is found for it. Gelmo, for example, synthesised sulphanilamide about thirty years before the workers at the Institut Pasteur discovered its value against infections. Again, D.D.T. was known as a compound long before it became known as an insecticide. The same is true of curare, but with one interesting difference. Not only has curare been known as a native arrow poison since the days of Sir Walter Raleigh, but its powers of paralysing the body's muscles were carefully investigated by Claude Bernard in the eighteen fifties.

Although the action of curare has been familiar for so long, we find, in 1940, Goodman and Gilman writing in a standard text book of pharmacology:

'Curare is an important pharmacological tool for laboratory investigations, but as yet it has no well-established therapeutic uses.'

It was in the same year as this was published that Bennett announced that carefully prepared extracts of curare, standardised by tests on rabbits, could be used to mitigate the effects of convulsions. Artificial convulsions had of course been given to patients for some years before that as a means of improving certain mental conditions (e.g. shock treatment for schizophrenia). Two years later a more important use for curare came to light. Griffith and Johnson found that it could be used in surgical operations to relax muscles. By giving a small dose of standardised curare extract, these anaesthetists were able to bring about the relaxation of the patient's muscles without having to use a deep degree of anaesthesia. From the surgeon's point of view muscular relaxation is essential if he is to carry out

many operations, and the pain-relieving side of anaesthesia is secondary. It so happens that a very light anaesthesia is sufficient to abolish pain, but a very deep degree of anaesthesia may be needed to get enough relaxation of muscles for the operation to proceed, and the deeper the anaesthesia the more liable the patient becomes to the complications which can follow an operation. The use of curare is firmly established to-day, and it is now common to use the pure substance d-tubocurarine chloride, which is extracted from crude curare.

Every licence issued by the Home Office to permit experiments on animals makes special reference to curare, and requires that the Home Office be notified of any experiments on animals in which this drug is to be administered. These special precautions are taken because curare is not an anaesthetic and the Home Office wishes to make certain that no one administers it to experimental animals in the mistaken belief that it has anaesthetic properties. The precautions taken to protect animals against the misuse of curare underline the fact that in surgical operations curare assists in muscular relaxation, but plays no part in the elimination of pain.

The adoption of curare as an adjunct in anaesthesia provided a spur to a search for synthetic substitutes. There were other factors in the situation which lent force to this. In the first place curare is a mixture of paralysing agents. Since it is generally obtained as a product of plants which grow in Central America, the proportions of the ingredients may vary seriously from batch to batch unless the collection of the material has been carefully supervised. A variable mixture of poisons is obviously unsuitable for administration to man. Consequently, the manufacturer must carefully purify curare and stringently test his product on animals before he can sell it for clinical use. Normally this is done by extracting the pure substance d-tubocurarine chloride. The remaining material may be waste.

It is not surprising, therefore, that manufacturers would prefer to synthesise a substance of similar properties if it could be made from more readily obtainable raw materials, if the process of making were reasonably easy and if its purity could be established by a relatively cheap chemical analysis, rather than by an expensive biological one.

The stimulus for producing synthetic paralysing substances may perhaps then have come from the establishment of a field of clinical usefulness of the compound, together with an apprehension of the difficulties involved in obtaining and preparing curare itself for clinical use. When this situation provided the spur towards producing synthetic paralysing agents, then the scientist who responded to it used as his tools the detailed chemical and biological knowledge of the mode of action of curare, built up by the work of such men as Dr King at the National Institute for Medical Research, Dr Ing and his colleagues, and Sir Henry Dale and his school.

Knowledge of the structure of tubocurarine has provided the starting-point for recent attempts to produce synthetic paralysing agents. It will be seen from Fig. 22 (a) that the formula of tubocurarine is complex. Observers have noticed, however, certain salient features about this structure. First of all there are two nitrogen atoms (N) to which other atoms are linked in a certain way. The nitrogen atoms here, because of their mode of linkage to other atoms, are described as 'quaternary' nitrogen atoms.* A single quaternary nitrogen atom often possesses some degree of curare-like activity; and where two or more of these atoms are linked together, the activity of the whole compound may be greater still.

Students of the formula of tubocurarine have noticed too that a total number of ten atoms separates the two nitrogens, if the distance is traversed by any one of the three short routes. In order to show this fact in the figure I have numbered the con-

* A nitrogen atom can combine with three other atoms by sharing electrons with them and remaining electrically neutral. Thus one nitrogen and three hydrogen atoms form an atom of ammonia gas, NH_3. The nitrogen atom can also combine with a fourth atom, but in a different way, by offering two of its own electrons outright (donating) to the other atom; in which case it becomes electrically unbalanced (ionised) and carries a positive charge, like a metal atom. Thus NH_3 becomes NH_4^+, with a great change in chemical behaviour.

(a)

(b)

(c)

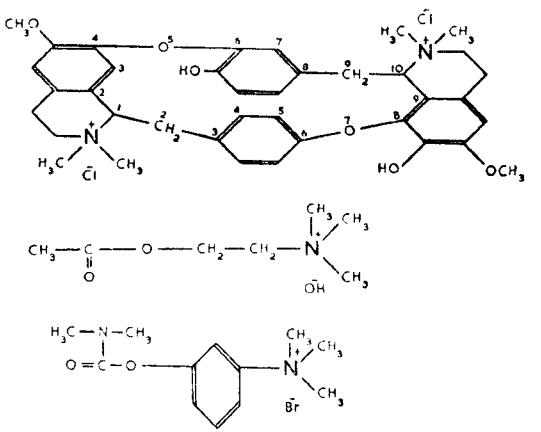


Fig. 22. Substances concerned with the transmission of a message from nerve to muscle, and its blocking. In this and other figures the quaternary nitrogen atoms are in heavy type and the atoms in the chains connecting the nitrogens are numbered. The figures of course do not show the actual shape of the molecule in space, they record merely the way the constituent atoms (C=carbon, O=Oxygen, etc.) are linked together, which tells the chemist a good deal about the chemical behaviour of the substance. To save writing down all the carbon atoms, by convention one reads a CH or C at each unlettered corner of the diagram – for instance at positions numbered 4, 8, in (a) above.

(a) d-Tubocurarine chloride, (b) Acetylcholine, (c) Neostigmine bromide

necting atoms on two of the short routes. Other points of interest exhibited by the formula of tubocurarine are the particular form of each double ring of atoms of which a nitrogen is one member. The attachment of a methoxy (CH_3O) group to each of these rings is of considerable importance.

The importance of the methoxy group is illustrated if the synthetic chemist adds two more groups of this type to the molecule. The resulting derivative, dimethyl tubocurarine, illustrated in Fig. 25 (a), is about ten times as potent as tubo-

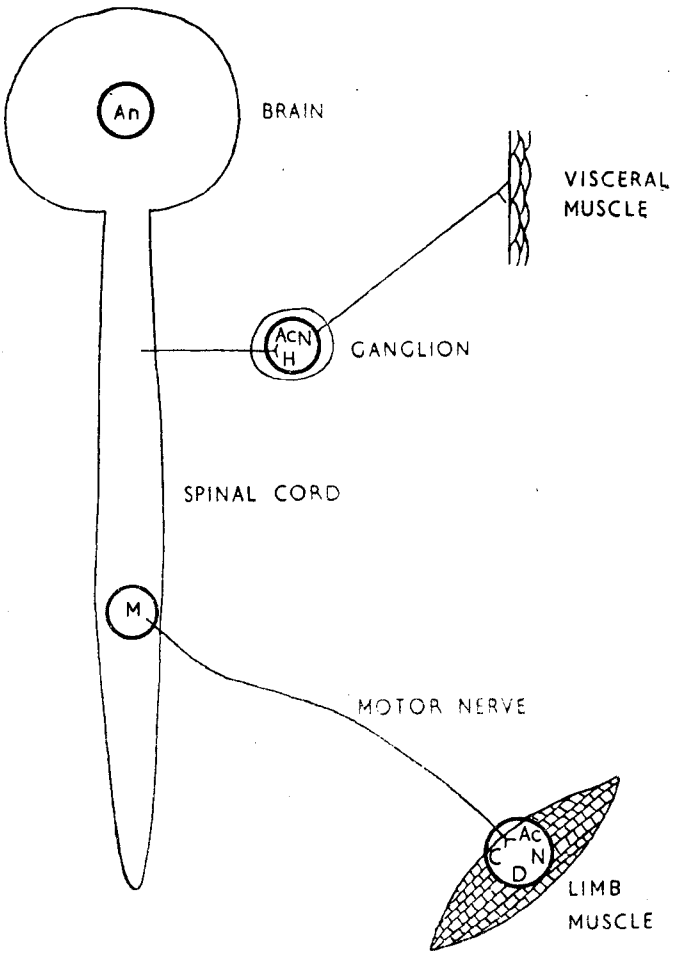


Fig. 23. Highly simplified diagram of nervous system and muscles to illustrate main sites of action of drugs discussed. Circles indicate site of action of drugs. Ac=acetylcholine; An=anaesthetic; C=curare or curare derivatives; D=decamethonium iodide; H=hexamethonium iodide; M=myanesin; N=neostigmine

curarine when tested in the rabbit. In man it is several times more active than the natural compound.

Before discussing how synthetic paralysing substances have been developed, let us consider for a moment the way in which curare exerts its characteristic effect. The numerous muscles of the limbs and of the trunk are each served by motor nerve fibres. The muscle moves in response to a message which has originated in the brain or spinal cord, and which has passed down the motor nerve fibres and over the junction between nerve and muscle. Fig. 23 shows diagrammatically the points where drugs affect the nervous and muscular systems. It will be seen from this diagram that curare acts by blocking the transmission of the message between the nerve and the muscle.

From the biochemical point of view the transmission of a message from nerve to muscle is a very interesting affair. A message travels down a motor nerve at a rate of about 50 metres per second in the form of a rapid volley of impulses. The passage of these nerve impulses can be detected by instruments designed to measure electric currents which last for a very short time. When an impulse reaches the end of the nerve it releases the substance acetylcholine. When acetylcholine touches the adjacent muscle it evokes a contraction of the muscle fibres which spreads over them in a wave. A glance at the formula in Fig. 22 (b) shows that acetylcholine possesses, like tubocurarine, a quaternary nitrogen atom. According to the principle that chemically related substances may antagonise one another, we may suppose that tubocurarine competes with acetylcholine. Perhaps tubocurarine is absorbed by the muscle fibres at the spot where acetylcholine would normally be absorbed.

The blocking of transmission between nerve and muscle by tubocurarine can be described then as a tussle between this substance and the natural transmitter acetylcholine. Another factor comes into the situation. A substance called cholinesterase, which exists in the tissues, destroys acetylcholine when it is liberated in the body. Cholinesterase can be 'poisoned' by various drugs, one of which is neostigmine (Fig. 22c). Administration of neostigmine thus protects acetylcholine and keeps up

its concentration in the body. Hence if we give neostigmine, after giving tubocurarine, muscular paralysis is lessened, because more acetylocholine is available in the struggle. Hence neostigmine serves as an antidote to tubocurarine. As we shall see later, this antidote can be very useful in operations.

In the account which follows we will confine ourselves to the synthetic paralysing substances which also act by blocking transmission between nerve and muscle. This means that we shall not discuss further the substance myanesin, Fig. 24 (a), whose paralysing action was first reported by Berger and Bradley.

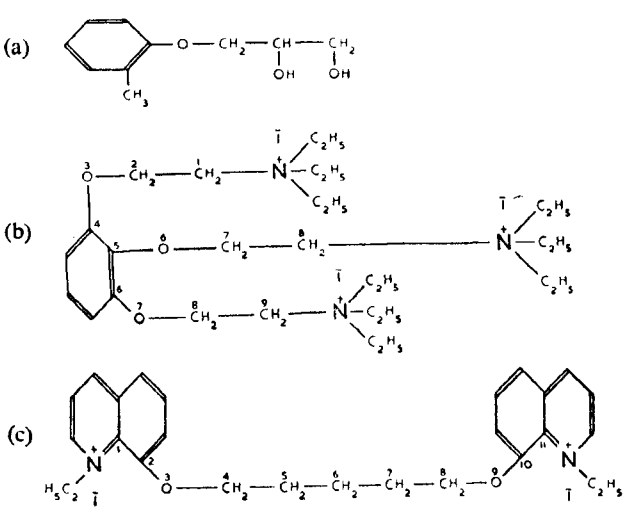


Fig. 24. Some synthetic paralysing agents

(a) Myanesin, (b) Flaxedil, (c) 3381 R.P.

Myanesin interferes with the conduction of messages in the spinal cord and thus paralyses by quite another type of action than curare. Let us consider then the synthetic substances which block the transmission of a message between nerve and muscle.

Professor Bovet and his colleagues at the Institut Pasteur and in Rome have played a leading part in developing synthetic paralysing agents of clinical value. The first of these to be useful in operations, flaxedil, Fig. 24 (b), was made and studied in his laboratory. He has investigated many other interesting compounds also, such as Compound No. 3381 R.P., illustrated in Fig. 24 (c).

A glance at these formulae will show that they contain two or three of the characteristic quaternary nitrogen atoms, linked by chains of atoms from 8 to 11 in number. They thus conform fairly closely to the fundamental pattern of the molecule of d-tubocurarine itself.

As soon as an active synthetic paralysing compound has been discovered, a whole series of questions must be asked before we can know whether it is likely to be of any clinical use. Every drug has secondary, or side, effects as well as its main action, and we must be sure that the side-effects of a potential paralysing agent are not serious. It was especially because of its lack of serious side-effects that Bovet selected flaxedil for trial in the clinic.

Nearly as important as the lack of side-effects is the need for an antidote. If an anaesthetist is going to use a powerful poison, as these drugs are, he wants to be quite sure that he has an antidote ready to hand, in case his patient is unduly sensitive to a normal dose, or in case the action needs to be cut short. For the reasons already given, the drug neostigmine is an effective antidote for curare. And, as might be hoped, the same antidote serves for flaxedil.

A further success in the search for paralysing agents has recently been achieved in England. A few years ago in Oxford Drs Ing and Barlow set about making and studying a number of simple compounds in which they incorporated some of the salient features of the d-tubocurarine molecule. The simplest and most effective paralysing agent among these was decamethonium iodide (originally known as C10) illustrated in Fig. 25 (b). It will be seen that this bears two quaternary nitrogen atoms separated by a simple chain of ten carbon atoms.

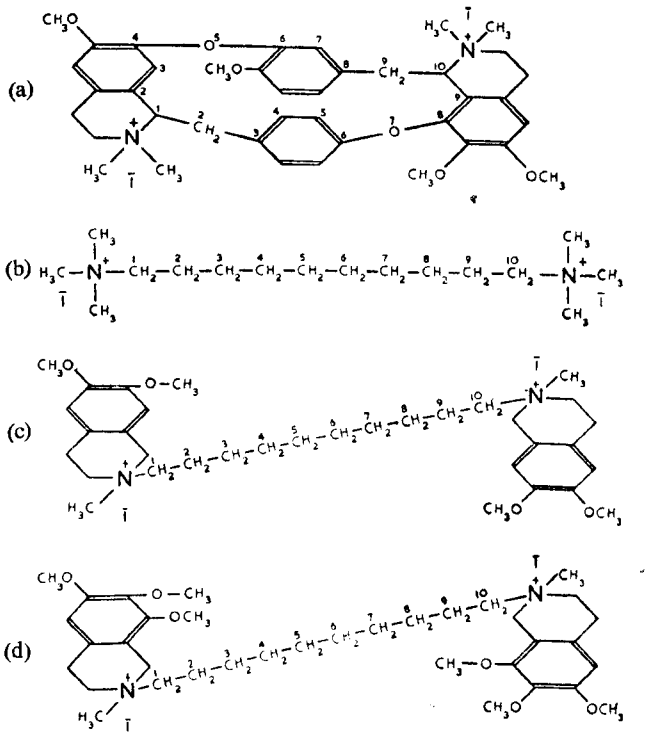


Fig. 25. The chemical 'marriage' of paralysing agents
(a) d-O.O-dimethyl tubocurarine iodide, (b) Decamethonium iodide (C10), (c) Compound 14, (d) Compound 15

At the same time as Ing and Barlow were working in Oxford, Dr Paton and Mrs Zaimis, following an observation in another experiment, were working in Hampstead on the same types of compounds.

Both teams of workers made compounds in which the quaternary nitrogen atoms were separated by chains of carbon atoms of various lengths, and independently each team arrived at the conclusion that the paralysing and other activities of

compounds of this type was closely associated with the length of this linking chain.

It will be seen from Fig. 23 that the spinal cord is linked with the muscles and glands of many organs, such as the viscera, the eyelids, the pupil of the eye and so on, through groups of nerve-cells called ganglia. Paton and Zaimis studied how far their compounds blocked conduction at the junctions between (a) nerve and muscle and (b) nerve-cells inside a ganglion. They obtained the graph shown in Fig. 26 by plotting the potency

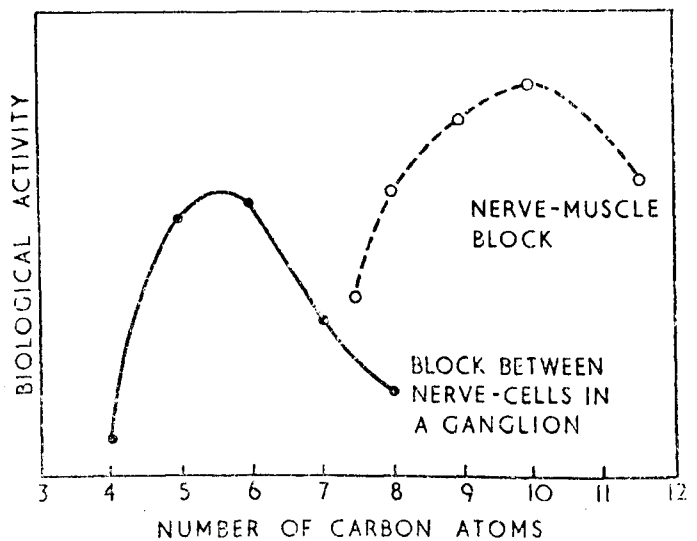


Fig. 26. Relation of biological blocking activities of methonium compounds to the number of carbon atoms in the linking chain (after Paton and Zaimis)

of each compound in blocking these two types of conduction against the number of carbon atoms in the linking chain. It will be seen that the compound most active in interrupting transmission in the ganglion contains six carbon atoms in its chain (hexamethonium iodide). On the other hand ten carbon atoms give the best length of chain for blocking the nerve-muscle

junction. Acetylcholine mediates the passage of nervous messages within the ganglion as well as between nerve and muscle. Thus we have the rather paradoxical situation that acetylcholine is antagonised best in two different sites by compounds of different chain-lengths.

Detailed study of the active paralysing substance decamethonium iodide, discovered in this way, has shown that it is suitable for use in operations. An effective dose has few undesirable side-effects. Neostigmine does not act as an antidote to this compound, but Paton and Zaimis found that the compounds with 5 or 6 carbon atoms in the chain antagonised the 10 atom compounds. However, this antidote has its own powerful effects on the body and the situation regarding an antidote is not finally settled.

Recently Dr Taylor and I have sought to produce active compounds by a 'marriage', as it were, between decamethonium iodide and the tubocurarine type of compound. In Fig. 25 (c) and (d) two of these compounds are illustrated. Each might be regarded as the offspring of decamethonium iodide and the dimethyl derivative of tubocurarine, illustrated in the same figure. It will be seen that in dimethyl tubocurarine the quaternary nitrogen atoms are placed in not quite similar rings. Compound 14 inherits the left-hand ring of dimethyl tubocurarine and Compound 15 the right-hand ring. Both compounds are very active curarising agents. Compound 15 is about half as active in the rabbit as its natural parent and about seven times more active than its synthetic parent. Compound 14 possesses about one-half of the activity of Compound 15. Both compounds and other relatives await further tests before it can be known whether or not they are likely to be suitable for clinical use.

The story of synthetic curarising substances gives hope to the scientist in two ways. First it makes him realise that he has gained a little useful knowledge in this field, for he can sometimes foretell that a new compound will possess curarising activity and he can even foresee a little how active it will be. Secondly the subject gives hope because it is full of unsolved contradictions.

And so there is the possibility of some new observation or interpretation solving these contradictions and lifting our knowledge to a higher level.

REFERENCES

Goodman, L., and Gillman, A. *The Pharmacological Basis of Therapeutics*, p. 495. Macmillan, New York, 1941.

Bovet, D., and Bovet-Nitti, F. 'Curare.' *Experientia*, Vol. IV/9, 1948, p. 322.

Paton, W. D. M. 'The Pharmacology of Curare and Curarising Substances.' *J. Pharm. & Pharmacol.*, 1949, **1**, p. 273.

Paton, W. D. M., and Zaimis, E. J. 'The Pharmacological Actions of Polymethylene Bistrimethyl Ammonium Salts.' *Brit. J. Pharmacol.*, 1949, **4**, 381.

Taylor, E. P., and Collier, H. O. J. Paralysing Activity of some further Heterocyclic Decamethylene - $\alpha\omega$ - bis Quaternary Ammonium Compounds. *Nature*, 1950, **165**, 602.

MAKING CHEMICALS FROM PETROLEUM

C. L. GILBERT

IN years gone by, man relied on products obtained directly from living sources such as animals, fish, trees, plants and fruits for his drugs, dyes, fats, solvents, soaps, and fabrication materials for clothing, etc. With the increasing complexity of life to-day, these sources are unable to supply materials in the necessary quantities, so that synthetic organic chemical industries have sprung up to supplement or improve upon natural products and, in many cases, to provide entirely new materials: rubbers, plastics, drugs, flavourings, dyes, etc. A particularly important development has been the industries based on nitrocellulose and cellulose acetate, which require enormous quantities of chemicals for the manufacture of lacquers, fibres, explosives, etc.

RAW MATERIALS FOR ORGANIC CHEMICALS

Apart from the products obtained by direct extraction from natural sources, the main raw materials from which chemicals are prepared are:

(i) Wood
(ii) Carbohydrates
(iii) Coal
(iv) Petroleum

(i) From the destructive distillation of wood are obtained methyl alcohol and acetone and, from these, derivatives such as acids, esters and ethers are prepared. We also have the distillation of the resinous exudation of pine trees which provides turpentine, pine oils, resin, etc.

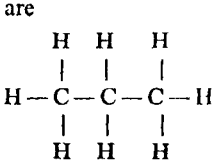
(ii) The fermentation of carbohydrates (the action of yeast and bacteria on sugars and starches) provides ethyl, propyl and

butyl alcohol, acetone, acetic acid, glycerine, etc. This process is now applied to a variety of materials, such as corn husks to give furfural, and grain to give butylene glycol. Agricultural products are increasing in importance as chemical raw materials, as shown by the development of plastics and synthetic fibres from, for example, soya beans and groundnuts.

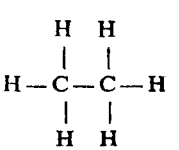
(iii) Coal provides chemicals by a variety of methods; the oldest is the distillation of coal tar to obtain the aromatic hydrocarbons, benzene, toluene, xylene, naphthalene, etc., and phenols and these form the basis of a wide range of chemical derivatives – aniline, nitrobenzene, dyes and drugs. A further method is the treatment of coke with steam – the water/gas reaction – which produces methyl alcohol and, from this, formaldehyde. In Germany, there have been very considerable developments of modifications of this process, such as the Fischer-Tropsch, which produce a wide range of fuels and alcohols, etc. From coal, we must also include the acetylene produced by the action of water on the calcium carbide formed by heating coke and lime in an electric furnace. In areas where electric power is cheap, this is the basis of large industries producing acetone, acetic acid, acetaldehyde and vinyl chloride (used for plastics manufacture).

(iv) Petroleum, being a mixture of hydrocarbons* of different types, can produce by chemical reactions a wide variety of organic

* It should perhaps be explained that hydrocarbons are made up of carbon and hydrogen atoms and that each carbon atom can combine with a maximum of four other atoms, which may be all carbon atoms, or may be hydrogen, etc. Carbon atoms linked by single bonds into a long chain are called *paraffins*, into a ring *naphthenes*. Examples of paraffins are



propane



ethane

and they are chemically rather inert. But if instead of being singly

chemicals. The main sources of hydrocarbons used for chemical manufacture are:

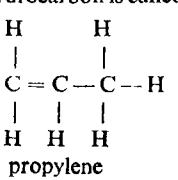
(a) Natural gas, and
(b) Gases from cracking plants

(a) Natural gas, which occurs in the ground with crude petroleum, is a mixture of light (small molecule) paraffin hydrocarbons such as methane, ethane, propane and butane and, as will be seen from Fig. 27, can be converted to a variety of chemical products. Thus, by combustion, carbon black is produced or, by thermal decomposition (cracking), another type of carbon and also hydrogen, and the latter can be converted to ammonia. In spite of the stable chemical nature of paraffin hydrocarbons, processes have been developed for their oxidation, chlorination and nitration to a wide variety of derivatives. Natural gas can also replace coke in a modification of the water-gas reaction ($CH_4 + H_2O \longrightarrow CO + 3H_2$) to give 'synthesis' gas for motor fuel and chemical manufacture. It is possible to decompose it in an electric arc and produce acetylene, and hence the usual derivatives obtained from calcium carbide.

(b) Crude petroleum itself is a mixture of paraffin, aromatic and naphthene hydrocarbons and certain of these are extracted

---

linked to four different atoms, a carbon is twice-linked to another carbon, the hydrocarbon is called an *olefine*



propylene

ethylene

and this double-link, a form of distortion, is under strain and readily broken, so that olefines are chemically reactive and can readily be converted into alcohols, ketones, oxides, glycols, amines, acids, esters, etc.

Carbon atoms linked by double bonds into a ring instead of a chain are called *aromatic* (e.g. benzene, toluene), corresponding to the single-bond naphthenes, as an olefine compares with a paraffin.

and used directly for chemical production, but the main source of chemicals is a further type of hydrocarbon, the olefine or unsaturated type, which is produced by processes such as cracking. This process consists of heating a petroleum product to

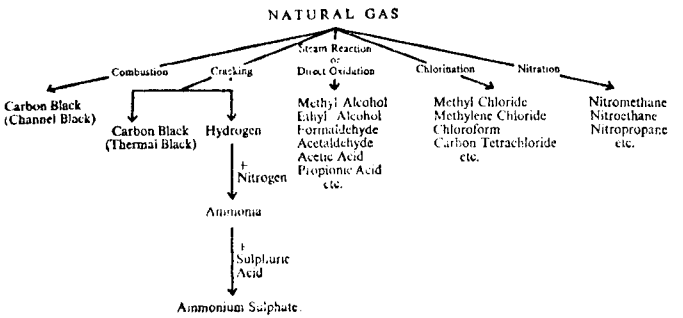


Fig. 27. Derivation of chemicals from natural gas

a high temperature and pressure, causing the large molecules to break up into small ones. It is actually a very complex reaction, but a simple picture of the first step in cracking a paraffin hydrocarbon may be given as follows:

$$\underset{\text{decane}}{C_{10}H_{22}} \longrightarrow \underset{\text{ethylene}}{C_2H_4} + \underset{\text{octane}}{C_8H_{18}}$$

The gases obtained from these cracking processes contain high proportions of olefines such as ethylene, propylene and butylene and these, after separation by distillation, are the main raw materials for chemical manufacture (see Fig. 28).

It should also be noted that certain cracking processes such as the Catarole (which uses a copper catalyst) produce high proportions of aromatic hydrocarbons and these, in turn, form the basis of a similar range of derivatives to those from the coal tar industry.

As already mentioned, some of the hydrocarbons actually present in crude petroleum are frequently used for conversion to chemicals. These include a naphthene hydrocarbon, cyclohexane, which is separated from gasoline by fractionation

(see later, p. 138) and is oxidised to provide adipic acid for nylon manufacture. During the War large quantities of naphthenes such as methylcyclohexane and dimethylcyclopentane separated

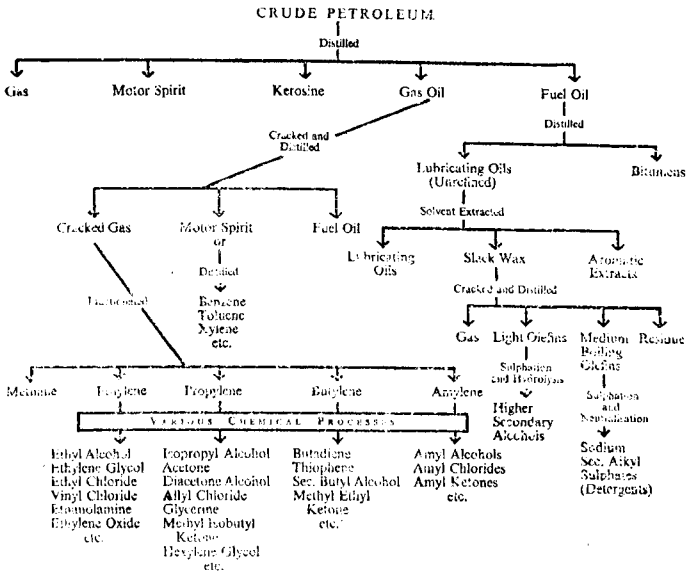


Fig. 28. Derivation of chemicals from crude petroleum

from gasoline were converted to toluene for explosives manufacture. Of the paraffin hydrocarbons, the simpler members present in natural gas, such as methane, ethane and propane, form the main chemical raw materials, but mixtures of paraffin hydrocarbons from special crudes (comprising the kerosine fraction) are also chlorinated (treated with chlorine) for use in the manufacture of synthetic soaps or detergents. Similarly, the very long chain paraffin hydrocarbons which form paraffin wax, separated from petroleum by chilling or by solvent precipitation, can be cracked to give olefines. A special type of the latter process is the production of higher olefines which are converted to synthetic detergents by sulphation (treatment with sulphuric

acid) and neutralisation (these are termed 'sodium secondary alkyl sulphates' and form the basis of 'Teepol').

An aromatic hydrocarbon, *ortho*xylene, which is a component of a fraction produced in certain cracking processes, is also used instead of naphthalene from coal tar to make phthalic anhydride, a component of 'alkyd' synthetic resins.

It will be seen, therefore, that although the petroleum chemical industry is of comparatively recent growth, it is already producing very many of the chemicals originally derived from other sources, such as coal and agricultural products, and is capable of producing practically any organic chemical.

GROWTH OF THE PETROLEUM CHEMICAL INDUSTRY

The petroleum chemical industry is of small size compared with the petroleum industry itself (about 0.5 per cent in volume of products). How it has grown up can be visualized from a quick backward glimpse into the early days of petroleum. In 1859, when Colonel Drake drilled the first well for crude oil at Titusville, Pennsylvania, the demand was for lamp oil and lubricants to supplement the animal and vegetable oils then used. The distillation of this crude oil produced light gasoline fractions, for which there was no demand, so that they were thrown away or burnt.

In the early part of the twentieth century, however, the development of the motor car provided a demand for these light fractions and, by 1912, it had grown so great that it could not be met by straightforward distillation of crudes without leaving at the same time more residual fuel oil (of high boiling-point) than could be used. This difficulty was overcome by development of the 'cracking' process in which heavy residual fuels were decomposed at high temperatures and pressures to yield motor spirit. This solved the problem of the motor fuel shortage, at the same time creating another disposal problem, that of the gases produced during cracking.

It was not long before the industry was busy experimenting with the conversion of these gases to chemicals, and a modest start was made in 1925 with the production of 70 tons of isopropyl alcohol from propylene, one of the hydrocarbons present in cracked gases. This extended rapidly, particularly with the growing demand for acetone, derived from isopropyl alcohol by dehydrogenation (removal of two hydrogen atoms from each molecule), for use in the artificial silk industry. This was soon followed by the production of secondary butyl alcohol and methyl ethyl ketone from butylene. From then on, the reactivity of the olefines was studied intensively and a wide variety of derivatives were produced, including alcohols, ketones, glycols, oxides, acids, amines, etc. The big opportunity came during the 1939–45 War owing to the shortage of essential materials imported from abroad, such as rubber, and to the enormous demands for all kinds of chemicals for explosives, etc. It was necessary, for example, to double the pre-War production of carbon black* from natural gas and to build up an entirely new production of about a million tons of synthetic rubber from butadiene (obtained from ethyl alcohol and from the dehydrogenation of butane and butylene) and styrene (from coal tar benzol and from petroleum ethylene).

The War years also saw considerable extensions in the production of chemicals from natural gas by oxidation, nitration and chlorination, etc. Ethylene, in its turn, has now become an important raw material through the development of processes to yield ethylene oxide, ethylene glycol (an anti-freeze), ethyl alcohol and ethyl chloride (used for making tetra ethyl lead, an 'anti-knock' in motor-car petrol).

It may be remarked that this interest in petroleum chemicals led to two important discoveries:

(1) That the paraffin hydrocarbons originally considered most unreactive could be oxidised, nitrated and chlorinated: and (2) that it was possible to chlorinate olefines at high temperatures without affecting the double bond (i.e. substituting a chlorine atom for one of the hydrogen atoms, instead of adding it on in the usual way). The latter process, based on propylene,

* Carbon black is used as a 'filler' for rubber motor tyres, which may contain as much as 40 per cent carbon black, which greatly increases their lifetime.

has led to commercial production of glycerine in the U.S.A.

An indication of the growth of the U.S.A. petroleum chemical industry is given by the following figures:

1925	67 tons
1942	224,000 ,,
1943	700,000 ,,
1944	1,260,000 ,,
1945	1,480,000 ,,
1946	1,700,000 ,,

By now, the figure is certainly over 2,000,000 tons and this is one-third of the total production of organic chemicals in the U.S.A. It includes 75 per cent of the acetone production, 100 per cent of the isopropyl alcohol, 70 per cent of the ethyl and methyl alcohol (and of most of the other alcohols) and 20 per cent of the glycerine. It has been estimated that a billion dollars has been, or is being, spent on the construction of petroleum chemical plants in the U.S.A. and sizeable industries are growing up in Canada, France, Holland and Russia as well as in Britain.

To-day, the petroleum chemical industry is providing in increasing quantities the raw materials for synthetic rubbers such as Thiokol, Buna and Butyl, for plastics such as P.V.C., methacrylate, polystyrene and polyethylene and for fibres such as nylon and polyvinylidene chloride. The last few years have also seen rapid developments in the synthetic detergents industry, which is now approaching half a million tons in the U.S.A., of which petroleum products provide the largest proportion of raw materials.

The following factors can be put forward to explain the rapid growth of the petroleum chemical industry:

(i) It is based on hydrocarbons which are cheap, available in large quantities and for the most part highly reactive.

(ii) The petroleum industries spend very large sums of money on research.

(iii) There is an increasing demand for raw materials for the synthetic and semi-synthetic industries.

(iv) It is economically important for the petroleum industry to make maximum use of its by-products.

In this country, the first large-scale manufacture of chemicals from petroleum was that of the Shell synthetic detergent Teepol in 1942. Since then, there have been plans for the construction of large petroleum chemical plants by Shell Chemical Manufacturing Co. Ltd., Petrochemicals Ltd., British Petroleum Chemicals Ltd. and Imperial Chemical Industries. The first-named are already in production of a range of solvents, and it is this plant which will form the basis of the remainder of this article. The other plants are expected to come into full operation during the next few years.

FROM THE LABORATORY TO THE PLANT

From the discovery of a new chemical reaction or the manufacture of a new chemical in the laboratory to the bringing on stream of a commercial plant is a long process with difficulties at every step. For every new product that reaches commercial production, dozens fall by the wayside on some count or other. The primary requirement is, of course, that the manufacture of the product should show a reasonable return for the capital expenditure – and erecting a chemical plant is a costly affair. If it is a question of a new process for producing a known product, it is very much easier to determine the profitability of the venture. When, however, the product is a new one, it is necessary to go through a number of stages. Firstly, its value as a potential raw material or for direct use in a variety of applications must be determined, then it must be established that the manufacture of the product does not present undue difficulties in the way of toxicity, low yields, etc., and that the patent coverage is adequate. Samples of any promising materials are sent to customers, research laboratories, etc., for their opinion, and some attempt is made to determine the potential size of the market. If the interest is maintained, the next step is to produce samples on a larger scale for a fuller market evaluation. This will involve the construction of a pilot plant, from which the optimum process conditions can be determined. If interest is still maintained in the product, and the cost of its manufacture appears reasonable and it is decided to go ahead, the next step will be to draw up a

detailed process flow chart from which the type of plant necessary to produce the required amount of product can be worked out. This involves breaking the design down into so many pieces of equipment, of which blueprints of all but standard units, such as pumps, will be required. The construction of equipment is then generally left to specialists in this field.

SHELL CHEMICAL PLANT

Thus the petroleum chemical industry produces a great variety of chemicals from a number of raw materials, mainly olefine (cracked gases) or paraffin (natural gas) hydrocarbons. As it will not be possible to discuss the process side of all these in any detail, it is intended to concentrate on the chemicals made at the Shell Chemical Plant at Stanlow (Cheshire). The products obtained and how they are derived are shown in Fig. 29. Here the raw materials for chemical processing are propylene and butylene, and these are obtained by cracking gas oil and fractionating the resulting gases. Each hydrocarbon forms the starting point of a series of reactions, a different chemical being produced at each stage. Thus, propylene is converted by stages into isopropyl alcohol (and isopropyl ether), acetone, diacetone alcohol, mesityl oxide and methyl isobutyl ketone. Similarly, butylene is converted into secondary butyl alcohol and methyl ethyl ketone.

The nature of the chemical reactions involved will be seen from Fig. 30. Reduced to their simplest terms, the reactions consist of adding a molecule of water to a molecule of the olefine (hydration) or, in a later stage, splitting off a molecule of water from the particular chemical – as in the dehydration of diacetone alcohol to mesityl oxide. Similarly, one reaction, the formation of methyl isobutyl ketone, involves the addition of hydrogen (hydrogenation) and another, the formation of acetone, the removal of hydrogen (dehydrogenation). In forming diacetone alcohol, two molecules of acetone are made to join together (condensation).

The cracking reaction itself is a very highly complex chemical process, involving both decomposition to give olefines, and

dehydrogenation of naphthenes to give aromatic hydrocarbons. Further, the olefine hydrocarbons produced require fractionation to obtain relatively pure base materials, and chemical

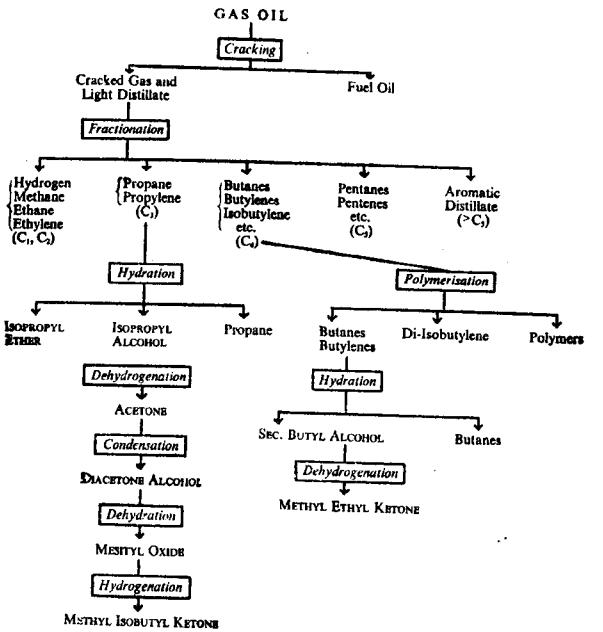


Fig. 29. Scheme of processes in chemical plant

treatment to remove sulphur compounds as well as isobutylene, which is not required for chemical processing. This is performed by 'polymerising' the isobutylene, that is causing two molecules to join together in the presence of sulphuric acid, from which the resulting di-isobutylene can be readily removed.

It will be appreciated that many further reactions could be carried out, such as the chlorination of propylene to give allyl chloride and hence glycerine, or its treatment with hypochlorous acid to give propylene chlorhydrin and hence propylene oxide and propylene glycol.

HOW CHEMICALS ARE MADE

How are these chemical reactions carried out in practice? It will be observed from Inset, Nos. 1–6 that the plant involved at Stanlow is of a highly complex nature, whereas a reaction such as the following is apparently simple:

$$CH_2 = CH - CH_3 + H_2O = CH_3 - CHOH - CH_3$$

(propylene + water = isopropyl alcohol)

In point of fact, few organic chemical reactions are simple and numbers of by-products are formed. Further, the starting material is seldom *completely* converted to the required end product and part of it remains unchanged. The plant must, therefore, be capable not only of carrying out the chemical reactions, but of separating the end product from unreacted constituents and from these by-products.

Let us consider in more detail the important factors in a chemical reaction. If, for example, we are converting isopropyl alcohol into acetone and hydrogen, during the reaction there will be a tendency for the reaction to go in the reverse direction, that is to hydrogenate the acetone formed, back to isopropyl alcohol. In point of fact, isopropyl alcohol is still sometimes made from acetone by hydrogenation. For this reason, a balance is set up between the initial and final products, i.e. the reaction reaches a state of equilibrium. In a chemical formula, this is expressed in the following way:

$$CH_3 - CHOH - CH_3 \rightleftharpoons CH_3 - CO - CH_3 + H_2$$

(Isopropyl alcohol — Acetone + Hydrogen)

The extent of conversion will depend on factors such as concentration of the reactants, temperature, pressure and time. A further important point is the rate at which the reaction takes place, and this obviously has to be borne in mind in practice by controlling the time of residence of the reactants in the system. It will be noted that practically all of the reactions described involve the use of a catalyst. This is a material which is used to increase the *rate* of reaction, without itself being chemically changed during the reaction. Sometimes a reaction proceeds so slowly that the catalyst behaves as if it initiates the reaction, but

it is really a question of acceleration. It causes no change in equilibrium, i.e. the *extent* of the chemical reaction, but does enable the reaction to be carried out under an entirely different set of conditions, in particular at a lower temperature, which is frequently advantageous, as for example in hydrogenation where at a high temperature there would be a much greater tendency for the reaction products to decompose.

The functions of a catalyst are highly complex and not completely understood. Catalysts are, in general, though not always, metals, metal salts or metal oxides. They are prepared by chemical methods in a fine state of division so as to provide a large surface area and are frequently deposited on supporting materials of a clay-like nature. It is believed that the atoms on the surface of these catalysts are in a highly active state and that the reaction being catalysed takes place at their surface. Some reactions are catalysed by liquid catalysts such as sulphuric acid, and a further variation is the use by the petroleum industry of very finely divided solid catalysts, which are transported round the reaction system of certain catalytic cracking processes, as if they were liquid. (Fluidised catalyst processes.) The main point is that the use of a catalyst increases the rate of the reaction and enables us to select more advantageous conditions for its promotion.

On the question of side reactions, it will be realised that we are dealing with, in the first place, highly reactive olefines which will not only undergo the chemical reactions we want, but will also react with themselves (polymerisation). Further, the chemical derivatives being produced are also reactive and may combine with the starting materials or with themselves to form unwanted products.

To sum up, therefore, the plant required for carrying out a chemical process will involve, firstly, the reaction equipment and, secondly, distillation equipment to separate the end product in a high state of purity from the other constituents present in the reaction mixture. In addition, there will have to be product tanks, pumps and lines to bring the reacting chemicals to the plant and take them away again, heating or refrigerating equip-

ment to provide and control the reaction temperature, compressors and refrigerating equipment for bringing the gaseous products to the liquid state when required for process work or distillation, mixing equipment, units for the preparation and regeneration* of catalysts, boilers to provide steam for heating or reaction control, etc. Finally, there are the highly important meters, indicators, recorders and controllers which are for convenience centralised at particular points in the plant. Each process is operated as far as possible on a continuous basis, that is the products pass from one stage to the next continuously, so that full use is made of controllers to keep the temperatures, pressures, rates of feed, etc., constant and ensure smooth running of the plant.

PROCESS PLANT

We can now consider the type of plant necessary for the carrying out of the chemical reaction proper, and for the purification by distillation.

Before the type of plant to be used for the chemical process can be decided upon, it is necessary to make a careful study of the effect of the reaction variables – concentration of chemicals, time and rate of reaction, temperature, pressure, type and life of catalyst, etc., and to select the optimum conditions. A consideration of Table 1 will show the extent to which these factors vary in the processes being considered.

Of these factors, time and temperature are the most important in deciding the type of plant to be used. Further, it is necessary not only to provide the optimum temperature of reaction but to maintain it. The latter point is important because many reactions produce heat (these are termed 'exothermic'), so that cooling is necessary, whereas others absorb heat ('endothermic'), so that additional heat has to be provided.

Whereas many chemical plants operate on a 'batch' basis, the petroleum chemical industry generally uses a continuous basis, so that instead of filling a large vat or vessel the reaction mixture

* After a time in use, a catalyst becomes less effective and requires 'cleaning'.

TABLE 1: REACTION CONDITIONS

Starting Material	*End Product*	*Nature of Reaction*	*Phase*	*Heat of Reaction*	**Temp. of Reaction*	*Rate of Reaction*	*Catalyst*
1. Gas Oil	*Olefins, etc.	Cracking	Vapour	Negative	High	Fast	None
2. Iso-Butylene, etc.	Di-isobutylene, etc.	1. Sulphation 2. Polymerisation	Liquid Liquid	Positive Positive	Low Moderate	Slow Fast	Liquid Liquid
3. Propylene	Isopropyl alcohol Isopropyl ether	1. Sulphation 2. Hydrolysis	Liquid Liquid	Positive Positive	Low Low	Slow Slow	Liquid Liquid
4. Butylene	Sec. butyl alcohol	1. Sulphation 2. Hydrolysis	Liquid Liquid	Positive Positive	Low Low	Slow Slow	Liquid Liquid
5. Isopropyl alcohol	Acetone	Dehydrogenation	Vapour	Negative	High	Fast	Solid
6. Sec. Butyl alcohol	Methyl ethyl ketone	Dehydrogenation	Vapour	Negative	High	Fast	Solid
7. Acetone	Diacetone alcohol	Condensation	Liquid	Positive	Low	Moderate	Solid
8. Diacetone alcohol	Mesityl oxide	Dehydration	Liquid	Positive	Moderate	Moderate	Liquid
9. Mesityl Oxide	Methyl isobutyl ketone	Hydrogenation	Vapour	Positive	Moderate	Fast	Solid

* Low temperature means below 100°C. High temperature means above 400°C.

runs through a narrow tube or series of tubes, heated or cooled where necessary. If, however, the reaction is a slow one, this will not give sufficient time for the conversion, and use is made of a larger vessel, generally a wide, vertical tube, through which the products pass slowly. This is termed a 'time-tank'. When the reaction mixture consists of two liquids which do not mix, such as a hydrocarbon and dilute sulphuric acid, it is difficult to ensure sufficiently close contact between the two phases, and the outgoing reaction products are therefore recirculated to the time tank, generally via a cooler, to give even longer for the reaction to be accomplished.

In the case of reactions 1, 5 and 6 in Table 1, fast, high temperature reaction conditions are involved, so that the reaction tubes are set in a furnace. This is the method generally used in the petroleum industry for heating crude oil, etc., prior to distillation and for carrying out all kinds of high temperature reactions, particularly cracking. A diagram of a typical furnace is shown in Fig. 31, page 136. It will be seen that it consists of a brick chamber lined on the sides and roof with horizontal pipes connected at the ends to form a continuous coil. In the front of the chamber will be a set of burners, fed by gas or fuel oil, the tubes and burners being arranged so that the flames do not impinge on the tubes, as this would cause overheating. The tubes are thus heated by radiant heat from the flame. The gases of combustion, after leaving the 'radiant' section of the furnace, then pass to another section, which may be on top of the furnace or at the side of it, separated by a wall. In this are set further rows of tubes comprising the 'convection' bank. Usually, the incoming feed will pass through the convection bank first, which is at a lower temperature, before going to the radiant bank.

In the case of reactions 5 and 6, a smaller furnace is used than for reaction 1 (since the quantities involved are much smaller) and the tubes are filled with catalyst. Where the reaction evolves heat, as in 9, direct heating cannot be used, so a nest of reaction tubes filled with catalyst and kept at the necessary temperature by hot oil circulating around the outside of the tubes is used. For reaction 7, a catalyst-filled vessel is used, and the temperature

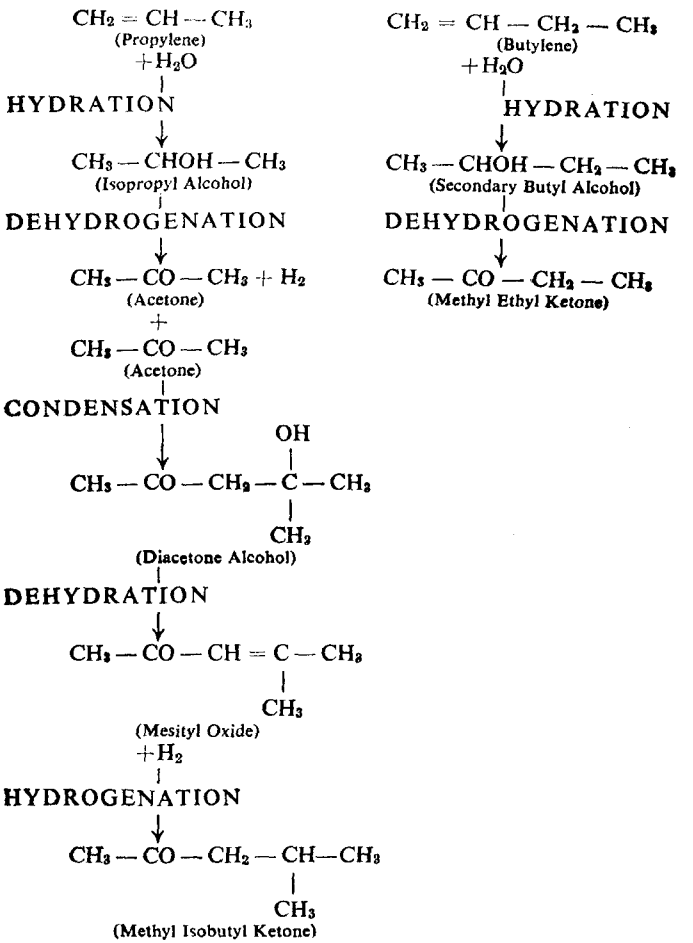


Fig. 30. Chemical reactions involved

is controlled by the use of external cooling, since the reaction evolves heat.

Most of the other reactions are slow and operate under moderate temperature conditions, and accordingly 'time tanks' are used, followed by coolers where heat is evolved in the reaction. In the case of the second stage in reaction No. 2 (polymerisation), the reaction is fast and operates at a higher temperature than the sulphation stage of this reaction, so that a single reaction tube, without recirculation, is used. An exception to the general rule is reaction 8, which involves handling a product which readily polymerises so that use is made of a steam-heated horizontal vessel, termed a kettle, from which the polymer can be drained off. (It would, of course, otherwise rapidly plug the tubes.)

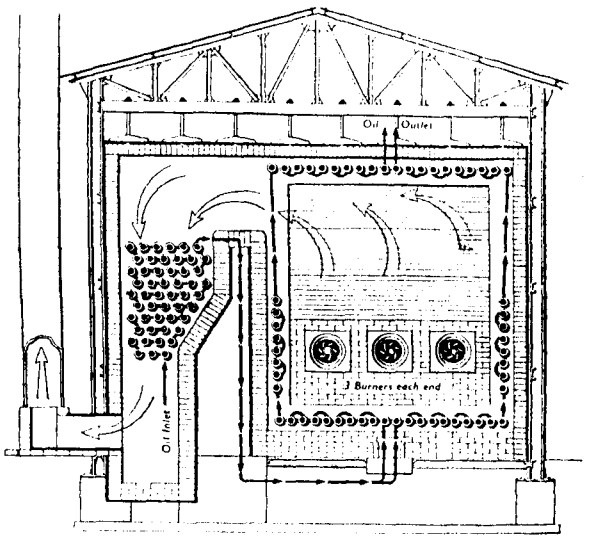


Fig. 31. Diagram of cracking furnace

This will give an idea of the reason for the selection of the different types of plant for a chemical reaction. Then the size

can be readily determined from a knowledge of the time factor and of the quantity of the chemical which it is required to produce. Further points have to be taken into account, viz. the type of metal to be used; this is mainly steel, but steel alloys are used for high temperature furnace tubes and for low temperature service, and corrosion-resisting metals or lead linings for equipment in contact with acid, etc. For the equipment to be used at high pressure, a knowledge of the strength of metals is necessary to determine the safe wall thickness, but this is generally available from engineering data.

DISTILLATION PROCESS

As already mentioned, the distillation process is used for the purification both of the gases used for chemical processing and of the chemical products obtained. Distillation columns form the basis of all petroleum processes, so that the design of those used in petroleum chemical plants follows that used for the distillation of crude oil and its products. (See Fig. 32). The columns consist of steel towers fitted with up to 60 trays of special design which enable a mixture of a number of products to be separated into its components. It should be explained that distillation in its simplest form is a method of separating substances which have different boiling-points by gradually warming up the mixture and condensing the vapour separately as each liquid in the mixture boils in turn. In the case of the mixture of hydrocarbons which comprise petroleum, it is only possible to separate them into fractions consisting of hydrocarbons having a wide range of boiling points, e.g. 35°-200°C for motor spirit, 150°-300°C for kerosine, and so on. In the case of chemical processing, where a pure material is required, each distillation column is used to separate a mixture into two components, a lighter and a heavier component, and, if these are still mixtures, they are distilled again to enable a further separation to be carried out. The conditions in the distillation column will then be such that one component emerges as liquid and the other as vapour, and the purpose of the trays is to obtain a clean separation between the two fractions. (See also Inset, Nos. 4 and 5.)

Reference to the diagram of a typical tray (Fig. 33), will show that it contains a number of so-called 'bubble caps' A, which surmount the vapour uptakes B, and overflow pipes ('downspouts') C taking off excess liquid to the tray below. The purpose

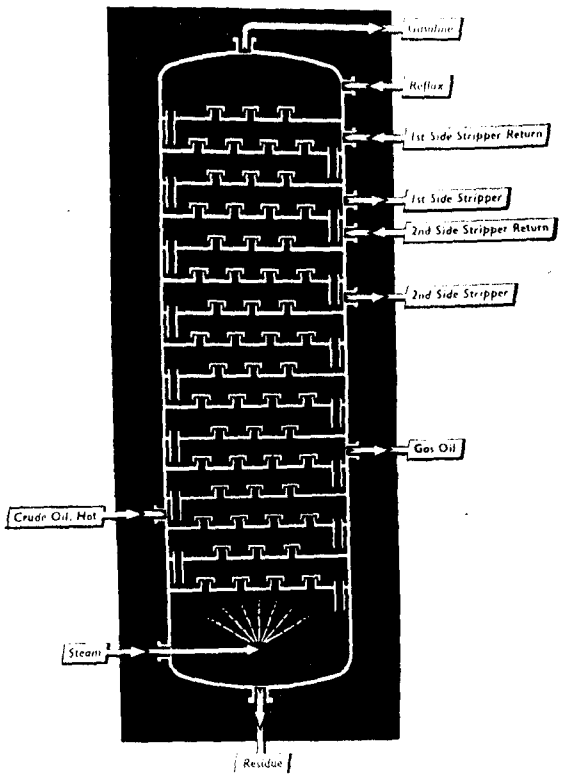


Fig. 32. Diagram of distillation column

of the caps, which have serrated edges, is to ensure that the vapour which rises through the uptakes and under the bubble caps into the liquid will pass into the liquid contents on the tray in streams of very small bubbles, so as to become thoroughly mixed with the liquid.

It would obviously be possible, theoretically, to separate a mixture of, say, six constituents continuously by heating the mixture in one still to such a temperature that the lightest constituent will distil off, and then passing the residue to a second

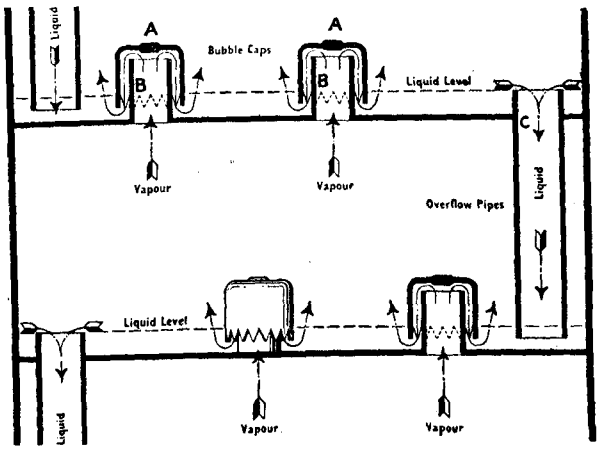


Fig. 33. Diagram of typical tray of fractionation column

still kept at a higher temperature, so that the next constituent distils off and so on, until five constituents are successively separated as vapour, and are condensed, the sixth remaining as liquid in the last still.

In the fractionation column the stills and condensers are, so to speak, set on top of each other so that distillation takes place from one tray, the vapour passing to the tray above, where some of it is condensed by the liquid already there, the rest passing to the tray above, and so on.

Let us imagine that a mixture of a low boiling point liquid x and a high boiling point liquid y is to be separated into its components. The bottom of the distillation column will be kept hot, by means of indirect steam heating, and the top of the column will be relatively cold, so that a temperature gradient exists down the column. When the mixture of x and y is first

introduced into the column, about half-way up, condensation takes place on the trays, so that they become full of liquid. As more feed is introduced, it comes into contact with the comparatively warm liquid in the tray immediately below the inlet point, and this causes it to distil, the vapours rising up the vapour intakes B and under the bubble caps A into the liquid on the next tray above, while the liquid overflows the tray via the downspouts C into the tray below. In the tray above, the fresh vapour will be in close contact with the liquid on the tray, and a further distillation will take place, but at a lower temperature than that on the tray below, so that the vapour rising from this will be richer in the light component x. Similarly, the condensed liquid on the tray will overflow to the tray below, where it will join the feed being distilled. Thus, a series of distillations and condensations takes place in the column, the vapour being progressively richer in x as it approaches the top of the column, and the liquid richer in y as it approaches the bottom of the column. With a sufficient number of trays, pure x will be recovered as vapour from the top of the column and pure y from the bottom as liquid.

This is an extremely simplified picture of distillation, and the whole equipment will also involve the use of what is termed 'reflux', i.e. a part of the condensed vapours from the top of the column will be sent back to the top part of the column again to increase the extent of fractionation. In the case of the fractionation of gases, high pressures are used to increase the boiling points of the liquid, whereas in other cases distillation is carried out under vacuum in order to reduce the temperature at which the products boil and hence to reduce risk of decomposition during heating. This is always applied to the recovery of lubricating oils.

A further type of distillation which is met with in the petroleum chemical industry is 'azeotropic' distillation. Some mixtures of liquids form a loose type of association of constant composition during distillation and cannot be separated by fractionation. For example, isopropyl alcohol and water form a constant boiling mixture containing 87.7 per cent of isopropyl alcohol and

12.3 per cent of water which boils at 80.2°C. In order to separate these, it is necessary to add a third component, which forms a three-component constant boiling mixture with these two components but which, on condensation, separates into two layers, the added liquid and isopropyl alcohol in one layer (which can be separated by distillation) and water comparatively free of isopropyl alcohol in the other. A similar principle is used for the production of anhydrous ethyl alcohol from fermentation processes.

It is not proposed to discuss the particular distillation problems and processes in connection with each of the chemical manufacturing processes described earlier, as these follow broadly the same pattern of separating, stage by stage, unwanted constituents. Thus, if product B contains a light impurity and a heavy impurity A and C respectively, it would be distilled twice, first to take off product A as a distillate, while B and C are drawn off from the bottom of the tower, and secondly the mixture B and C is redistilled, taking off B as a distillate and C as a residue.

The fractionation unit, however, which is used to separate the olefine streams for chemical processing is the largest unit in the Chemical Plant and will be described briefly.

It will be recalled that the feedstock for this plant consists of a mixture of gases and light distillate from the cracking unit. As will be seen from the simplified flow-sheet (No. 1, Fig. 34 on p. 143), the first step is to pass this feed into an absorber (A) which is fed with an absorption oil (actually the heaviest stream taken from the end of the plant). The absorber is a type of fractionation column in which the absorption oil is maintained at such a temperature that it will remove the heavier components from the feed by solution, whereas the lighter components are distilled over as vapour.

Before going further, it would be as well to refer to the shorthand system we are using to describe the different constituents present in the feed. We are using the term 'C_1' to indicate the hydrocarbon with one carbon atom, viz. methane (and for convenience also includes hydrogen). The term 'C_2' means hydro-

carbons consisting of two carbon atoms, viz. ethane and ethylene, and so on. For details of the constituents, see the legend on the simplified flow-sheet described.

In the absorber, the feed is separated into the vapour stream, consisting of C_1, C_2 and some C_3, and the liquid stream, consisting of some C_3, C_4, C_5 and heavier. The final aim of the plant is to produce C_3 and C_4 streams comparatively free from other constituents, so that the process will consist of taking off in turn lighter (C_1 and C_2) and heavier (C_5 and heavier) streams. It will be noted that the vapour from the absorber A is passed to a demethanizer B (this is a unit for distilling off methane), leaving C_2–C_5 as a residue which is transferred to a de-ethanizer C, taking off C_2 and leaving C_3. Now considering the liquid stream from the absorber, the first step is the debutanizer D, which takes off butane (C_4) and propane (C_3) present. This stream is combined with the bottoms from the de-ethanizer C and fractionated in the depropanizer E into the separate C_3 and C_4 streams required for chemical processing. Finally, the bottoms from the debutanizer D pass to the depentanizer F, where they are separated into a vapour stream, the C_5 fraction, which is recycled to the cracking furnace, and a liquid fraction termed 'aromatic distillate', which is partly returned to the system as absorption oil in absorber A, the balance being used for motor fuel.

It should also be noted that the gaseous streams at various parts of the plant are treated with chemicals to remove sulphur compounds. These include a Girbotol Unit, in which ethanolamines are used to remove hydrogen sulphide, and a soda 'scrubber' to remove other sulphur compounds. The gases are also dried carefully to avoid freezing at the low temperatures which have to be maintained in the distillation columns.

With regard to the distillation which is carried out as a means of purifying the chemical products from the plant, the general principles have already been outlined and it is sufficient here to emphasize that considerable redistillation is necessary to obtain products with the required degree of purity.

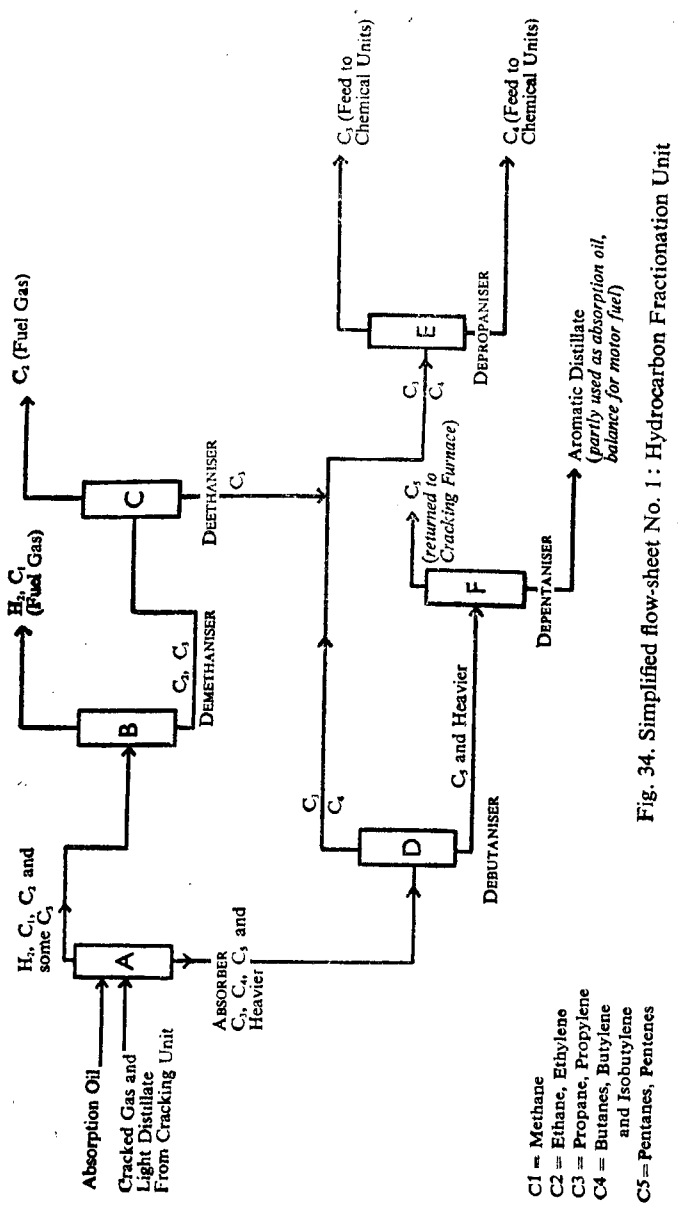


Fig. 34. Simplified flow-sheet No. 1: Hydrocarbon Fractionation Unit

A further point of process interest in the plant is the necessity for reconcentration of the waste acids, which leave the alcohol units at about 50 per cent strength, to the concentration required for processing. This is effected by evaporation of the water, the particular difficulty being the removal of carbonaceous material formed by the decomposition in the sulphuric acid of various reaction products present.

One other point which might be stressed with regard to the whole of the process design is that the plant has available streams of products, some at very high temperatures and some at very low, and full use of these is made throughout the plant for refrigeration or heating of feedstocks by heat exchange, in which the hot (or cold) liquid is passed through a series of tubes in a vessel filled with the liquid to be heated or cooled.

CARRYING OUT THE CHEMICAL PROCESSES

We can now discuss the way in which the various processes listed in Table I are carried out.

1) *Cracking*

The cracking process is the method used for the production from gas oil of olefines for chemical processing. As will be seen from the simplified flow-sheet (No. 2, Fig. 35) the essential pieces of equipment are the furnaces and the distillation columns. In brief, the process consists of heating the gas oil in a vaporizer furnace (A), and subjecting it to a simple distillation in the evaporator (B), in which the heaviest fractions are removed in order to reduce coke formation during the subsequent cracking. The heavy fractions which are subsequently used as fuel oil are drawn off the bottom of the evaporator column (B), while the vaporized gas oil from the top of the column is passed to the cracking furnace proper (C), where it is heated to a high temperature.

At this point, steam is injected into the plant, in order to suppress the recombination of olefines into higher molecular weight polymers, and to reduce coke formation. The resulting

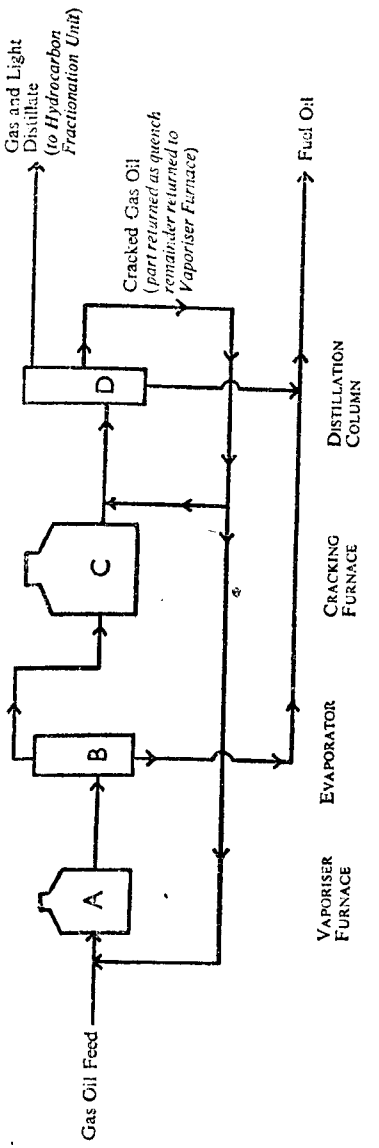


Fig. 35. Simplified flow-sheet No. 2: Cracking Unit

mixture of gases and liquids is passed to the distillation column (D), where it is separated into fractions, gas and light distillate being taken from the top of the column and heavy fuel oil from the bottom of the column. The gas and light distillate are sent to the hydrocarbon fractionation unit for the recovery of the particular gases required for the chemical processing. It will be noted that some of the intermediate fraction, cracked gas oil, is returned to the outlet of the cracking furnace as a 'quench' to cool the products rapidly and prevent further reaction taking place. The remainder is returned in the gas oil feed to the vaporizer furnace, so that it can be recracked to provide more olefines.

The process plant includes, in addition, the necessary pumps to feed the raw material into the process and to the different units, and steam generators and superheaters to provide the steam for the reaction and maintain the required temperature in the bottom of the evaporator. There are also storage tanks to hold the products from the distillation column, and the necessary process control and recording instruments.

(2) *Removal of Isobutylene (Cold Acid Polymerisation Process)*

Reaction:

$$CH_3-C(CH_3)=CH_2 + CH_3-C(CH_3)=CH_2 \longrightarrow CH_3-C(CH_3)_2-CH_2-C(CH_3)=CH_2$$

isobutylene + isobutylene → di-isobutylene

The mixture of cracked gas and light distillate from the cracking plant is pumped to the hydrocarbon fractionation unit. In this unit, two streams are obtained for the manufacture of chemicals, the C_3 stream containing propane and propylene, and the C_4 stream containing butane, isobutane and isobutylene, in addition to the butylene required for chemical processing. In the case of the C_3 stream, propane does not interfere with the subsequent production of alcohol, nor, in the case of the C_4 stream,

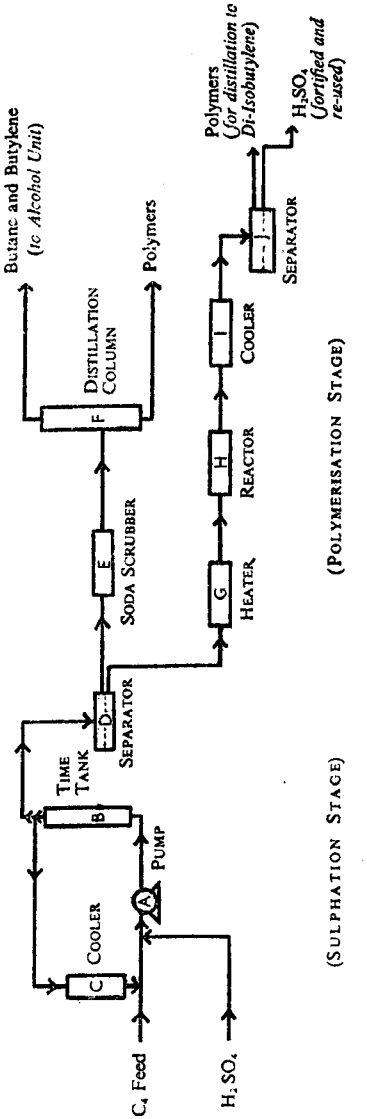


Fig. 36. Simplified flow-sheet No. 3: Cold Acid Polymerisation Process (for removal of isobutylene from butylene)

secondary alcohols from olefines (though in the U.S.A. ethyl alcohol is made from ethylene by direct hydration with steam in the presence of a solid catalyst). The reaction is complex and consists of two stages: (a) treatment of the propylene or butylene with 96 per cent sulphuric acid to form isopropyl or sec. butyl sulphates; and (b) the breaking down of these with water to the alcohol and sulphuric acid (hydrolysis).

In the case of isopropyl alcohol, the reaction is carried out (see Simplified Flow Sheet No. 4, Fig. 37) by mixing the feed (propane-propylene) with sulphuric acid in pump A and passing the mixture into time tank B. The greater part is recycled through a cooler C (since the reaction evolves heat) and returned to the time tank to ensure thorough contact. The remainder is fed to a separator D, where the unreacted propane is drawn off and washed with soda in a scrubber E. The liquid in the bottom of the separator (termed 'fat acid') contains a mixture of di-isopropyl sulphate and isopropyl hydrogen sulphate, and this is 'hydrolysed' in vessel F with added water, giving isopropyl alcohol and sulphuric acid. A complication in the reaction is the formation of isopropyl ether, as shown in the formula above. The mixture of isopropyl alcohol and isopropyl ether is removed from the acid after hydrolysis by steaming out in a steam stripper G (simple type of still). The acid is drawn off from the bottom of the stripper and sent to the recovery unit, while the mixture of alcohol, ether and water vapours is washed with caustic soda in neutraliser H in order to remove any entrained acid. The vapours are then condensed and are sent to the distillation unit for separation and purification.

A similar process is used for the production of secondary butyl alcohol from butylene.

(5) and (6) *Dehydrogenation of Isopropyl Alcohol to Acetone and of Secondary Butyl Alcohol to Methyl Ethyl Ketone*

$$CH_3—CHOH—CH_3 \xrightarrow{\text{catalyst}} CH_3—CO—CH_3 + H_2$$

isopropyl alcohol — acetone — hydrogen

$$CH_3—CHOH—CH_2—CH_3 \xrightarrow{\text{catalyst}} CH_3—CO—CH_2—CH_3 + H_2$$

sec. butyl alcohol — methyl ethyl ketone — hydrogen

The reaction involved here consists of the removal of two hydrogen atoms from the alcohol, and the process is carried out by passing the alcohol through a series of furnace tubes filled with a solid catalyst. The resulting mixture of acetone, unreacted alcohol and high boiling constituents, resulting from polymerisation of the acetone, is distilled, and the unreacted alcohol returned to the furnace for processing. Equipment is also required for the preparation of the catalyst and for its regeneration after removal from the furnace tubes. The hydrogen produced in the process is of a high state of purity and is used for subsequent hydrogenation (Reaction No. 9).

(7) *Condensation of Acetone to Diacetone Alcohol*

Reaction:

$$2CH_3—CO—CH_3 \xrightarrow{\text{catalyst}} CH_3—CO—CH_2—\underset{\displaystyle CH_3}{\overset{\displaystyle OH}{\overset{|}{\underset{|}{C}}}}—CH_3$$

acetone — diacetone alcohol

In this reaction two molecules of acetone combine in the presence of a catalyst to form diacetone alcohol. This is a low temperature reaction and the process is carried out by passing the acetone through a cylindrical vessel containing the catalyst. The heat of reaction is taken away by external cooling. It will be appreciated that the high chemical reactivity of acetone leads to the formation of a number of by-products which have to be removed by distillation.

(8) *Dehydration of Diacetone Alcohol to Mesityl Oxide*

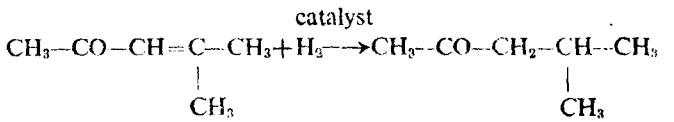
$$CH_3—CO—CH_2—\underset{\displaystyle CH_3}{\overset{\displaystyle OH}{\overset{|}{\underset{|}{C}}}}—CH_3 \xrightarrow{\text{catalyst}} CH_3—CO—CH=\underset{\displaystyle CH_3}{\overset{}{\underset{|}{C}}}—CH_3 + H_2O$$

diacetone alcohol — mesityl oxide

This reaction consists of the removal of a molecule of water from diacetone alcohol to give mesityl oxide. The process is

carried out by passing the diacetone alcohol into a reaction kettle together with the catalyst and keeping the contents hot by steam-heated coils, this also enabling the mesityl oxide to be distilled out of the kettle. Mesityl oxide is highly reactive and readily forms polymers, so that its purification by distillation is again a difficult matter.

(9) *Hydrogenation of Mesityl Oxide to Methyl Isobutyl Ketone Reaction:*



mesityl oxide hydrogen methyl isobutyl ketone

Hydrogenation is an extremely important reaction in chemical industry, being used for the hardening of fats, production of motor fuels from coal and the manufacture of ammonia and methanol. In this case, the reaction is carried out at much lower temperatures than in most hydrogenation processes. The equipment used is the usual nest of tubes filled with catalyst, surrounded by hot oil to keep the temperature constant. Hydrogen and the mesityl oxide are pumped into the reactor tubes and the resulting reaction mixture finally distilled to purify the ketone.

APPLICATIONS OF SOLVENTS

The chemicals produced at the Shell Chemical Plant are used either as solvents or as raw materials for chemical processing. It will be noticed that three of them are ketones (acetone, methyl ethyl ketone and methyl isobutyl ketone), and with these should be included diacetone alcohol, which is both a ketone and an alcohol. These have excellent solvent properties for gums, resins, nitrocellulose, etc., and their main use is in the formulation of compositions based on nitrocellulose, such as lacquers, varnishes, coatings for leather cloth and gas mantles, etc. They are used as solvents in the manufacture of transparent paper, in the spinning of cellulose acetate fibres, and for fabric processes such as trubenising. Acetone is used as raw material for the

manufacture of transparent plastics of the Perspex type, of perfumes such as ionone, and of drugs such as chloroform and sulphonal. Methyl ethyl ketone is used in the de-waxing of lubricating oils, in the production of resins for application to fabrics and for the manufacture of diacetyl (butter flavouring) and of rubber vulcanisation accelerators.

Both isopropyl alcohol and secondary butyl alcohol are used as 'latent' solvents in nitrocellulose lacquers, and the former is of value in the extraction of essential oils, for massage, in the synthesis of thymol and Paludrine (anti-malarial drug), and in de-icing and anti-freeze compounds. Diacetone alcohol is a valuable component of brake fluids, being a good solvent for castor oil.

From this, it will be seen that petroleum chemicals make a valuable contribution to the requirements of many important industries.

HUMAN COLOUR VISION
(PART II)

PROFESSOR H. HARTRIDGE, F.R.S.

(The first part of this article appeared in the previous issue)

SECTION VIII: THE POLYCHROMATIC THEORY

Section II (*Science News* 15, p. 16) contains a brief description of the pioneer work of Morton and his associates on the syntheses of pigments from Vitamin A_1 and from its aldehyde Retinene–1. These pigments have been shown to have the following characteristic properties. Since they are derived from Vitamin A_1 they are related to the important retinal pigment called visual purple; they have absorption bands which occupy a number of different positions in the spectrum; these absorption bands do not extend widely over the spectrum but are limited to a small part of it; like visual purple itself, they are bleached by ordinary white light.

The work of Granit and his co-workers* has shown that there are receptors in the eyes of many animals which have the following properties:

Their response curves occupy a number of different positions in the spectrum.

Their response curves do not extend widely over the spectrum but are limited to a small part of it.

The comparison of these facts with those discovered by Morton suggests the conclusion that pigments similar to those described might very well be the chromophores which are absorbing light energy to produce the responses of the receptors which Granit and his co-workers have found in the retinae of animals.

So far as human vision is concerned the situation as understood at present is as follows: there are seven types of retinal receptor;

* Described in Section III, p. 18 of *Science News* 15.

three of these belong to the important tricolour unit – orange, green and blueviolet; two belong to the less important Y – B unit, namely yellow and blue; and two to the less important R – BG – R unit, namely red and bluegreen. The red receptor has two responses, one in the extreme red and one in the extreme violet part of the spectrum. The positions of the crests of the receptors of the tricolour and other units are shown in Table I, and these are compared with the wavelengths of the receptors as found previously by the fixation method, which has been referred to in Section V, and by the subjective colour method, which has been referred to in Section VI.

TABLE I

Receptor	*Unit*	*Wavelengths of crests of receptors*		
		*Recent Methods**	*Fixation Methods*	*Subjective Method*
Red	R – BG – R	0.66μ	0.66μ	0.64μ
Orange	Tricolour	0.60	0.62	0.61
Yellow	Y – B	0.578	0.58	0.575
Green	Tricolour	0.547	0.54	0.54
Green	?	—†	0.52	—
Bluegreen	R – BG – R	0.50	0.50	0.51
Blue	Y – B	0.476	—	0.48
Blueviolet	Tricolour	0.45	0.46	0.45
Violet	R – BG – R	0.42	0.42	0.42

* See text.
† Required by retinal direction effect at about 0.52μ.

In the third column are given the wavelengths of the crests of the receptors of different type as calculated from such important data as the hue discrimination and the spectral mixture data of normal and of colour-blind subjects. In the fourth and fifth columns are given the wavelengths of the crests as found about three years ago by the fixation and the subjective colour methods. A comparison of these three columns will show that there is fair agreement between them, and this lends support to the view that these different methods have given reliable results.

The situation with regard to the receptors with wavelengths of 0.52μ is by no means clear. One of these was definitely found in the fovea* by means of the fixation method. It was also found by the writer to be required in order to account for the retinal direction efforts referred to in Section IV previously. But so far it has not been identified either by the subjective colour method, or as forming part of one of the three visual units about which there is information at present.

The situation with regard to the response at 0.42μ which has been referred to towards the end of Section VII (*Science News* 15) is not entirely clear either. At first this response was thought to occur at a wavelength of about 0.36μ, on the grounds that its overlap with the bluegreen receptor of 0.54μ produces a neutral point, which Willmer and Wright (1945) found at 0.41μ when they were investigating the properties of the fovea at small visual angles. Now experiments by the author, to be published in full elsewhere, suggest that this neutral point should be nearer to a wavelength of perhaps 0.45μ or 0.46μ and this would permit the short wavelength receptor to have a wavelength of about 0.42μ instead of 0.36μ. The following facts may be found to have a bearing on this point. (1) A definite fixation point was found for this wavelength, see Section V. (2) A specific subjective colour has been found at or near to this wavelength, see Section VI. (3) A well-marked hump has been found by Thomson (1948) in the luminosity curve of the fovea at small visual angles. (4) A receptor has now been found in the cat's retina by Granit at this wavelength. (5) This wavelength is in better agreement with the data on congenital tritanopia (disease in which yellows and blues are confused). There would therefore appear to be grounds for the assumption that there is a receptor with a crest in the human fovea at this wavelength 0.42μ, rather than at 0.36μ.

This theory accounts for the changes in colour vision which are found to occur when light intensity is reduced or the visual angle made small, which have been described in Section VII, in the following manner: the first change which occurs is the disappearance of the responses of the receptors which comprise

* See *Glossary*.

the Y – B unit. Since these contribute their share towards the total luminosity of the spectrum their disappearance causes a reduction in luminosity both in the yellow and in the blue parts of the spectrum. The former produces the notches in the yellow which have been described by Sloan and by Wright. With a further reduction either of visual angle, or of light intensity, the tricolour unit also ceases to be active, so that finally the R – BG – R unit is alone in a functional state. This is responsible for the dichromatic red–bluegreen vision which has been investigated by Willmer and Wright, and by the author.

When light intensity is greatly increased, or peripheral vision is substituted for foveal vision, then the reverse changes occur, and first the R – BG – R unit becomes inactive and then this is followed by inactivity on the part of the tricolour unit, the consequence being that the Y – B unit is alone effective and this brings about a form of dichromatic vision in which yellow and blue are the two components. In this manner the polychromatic theory accounts for the Bezold-Brucke phenomenon which was investigated by Purdy (1931) and also for the sizes of the perimetric colour fields, which were examined by Abney and Festing (1892).

In every one of these cases an essential feature of the theory is that there should be polychromatism not only of the retinal receptors, and of the nervous pathways, but also of the brain centres themselves; and the same statement applies to the phenomena of colour matching and of hue discrimination to which reference will now be made.

SECTION IX: COLOUR MIXTURE EXPERIMENTS

The fundamental fact on which the three-colour theory was based was the ability to match all colours including white by means of three selected spectral colours. But it must be realised that the match is not exact because the spectral colour is almost always noticeably more saturated* than the mixture which most

* Saturation or purity of a colour depends on the amount of whiteness mixed with it: thus *red* is more saturated (contains less white) than *pink* of the same brightness and hue (wavelength).

closely matches it. According to adherents of the three-colour theory this discrepancy is due to the response curves of the receptors spreading over a considerable portion of the spectrum. On the other hand on the basis of the polychromatic theory the difference is due to the activities of the additional receptors comprised in the Y – B and R – BG – R units. Thus spectral yellow appears to be more saturated than a mixture of red rays and green rays, because the latter stimulate the red and the green receptors strongly and the yellow receptors feebly, whereas the opposite is the case when spectral yellow is employed instead. Similar arguments apply to the extreme red, the bluegreen and the extreme violet parts of the spectrum.

Let us examine these two rival views in order to see which is the nearer to the truth. As stated above, the single spectral colours are not found to be matched exactly by means of mixtures of spectral red, spectral green and spectral blue because the former are always purer, that is are more saturated and contain less white than the latter. Therefore in order to obtain a precise match it is necessary to add to the spectral colour a desaturating colour which is red, green or blue according to the part of the spectrum which is being matched. Thus to orange, yellow or yellowgreen is added blue of 0.46μ wavelength; to bluegreen or peacock blue is added red of 0.56μ wavelength; and to red or violet is added green of 0.53μ wavelength. Now the addition of blue, 0.64μ, to the single spectral colour is physiologically equivalent to the addition of yellow of about 0.57μ to the mixture which matches that spectral colour, because blue and yellow are complementary to one another, and adding the one to one half of a matching field is equivalent to adding the other to the other half of that field.

In a similar manner the bluegreens and the peacock blues may be matched by mixtures of blue rays and green rays either by desaturating the former by adding red to them or by improving the saturation of the mixture by adding bluegreen of 0.54μ wavelength to it. This follows from the fact that red and bluegreen are complementary.

Lastly the reds and the violets may be matched by red (0.65μ) and blue (0.46μ) mixtures if they are desaturated by the addition of green rays. Now green has no single spectral complementary but is complementary to violet or purple, that is to a mixture of red and blue. In order to match all the spectral colours precisely it is therefore necessary to be able to mix together (1) extreme red (2) red (of 0.65μ) (3) yellow (of 0.57μ) (4) green (of 0.53μ) (5) bluegreen (of 0.50μ) (6) blue (of 0.45μ) and (7) extreme violet. Thus we see that seven component spectral colours are required in order to obtain precise matches with the spectral colours.

The fact that mixtures of only three spectral colours do not match the other spectral colours exactly has been accounted for on the basis of the three-colour theory by the supposition that the response of each of the three types of receptor is a wide one so that they overlap one another in the spectrum. For this supposition there is no confirmatory evidence and, moreover, it is found to lead to fallacious conclusions when it is applied to cases of colour blindness.

The situation may therefore be summarised as follows: the statement that all colours, including those of the spectrum, can be matched by means of mixtures of three spectral colours, red, green, and blue, is incorrect, because more than three colours are required. This fact has not been accounted for on the three-colour theory, but can readily be accounted for on the polychromatic theory.

SECTION X: HUE DISCRIMINATION

When subjects who have normal colour vision look at the spectrum of white light they see as a rule seven distinct colours, and lying in between these there are a series of intermediate hues. Very different is the case with subjects who are suffering from some form of colour deficiency. Thus the completely colour blind do not see any variations of hue throughout the spectrum, but only variations of brightness, and dichromats, that is to say people who have two types of receptor only, see the spectrum divided into three zones; a short wavelength zone, uniform in colour but varying in brightness, which is due to the short wave-

length receptor mechanism of their retina; a second region of uniform colour, but varying in intensity, at the long wavelength end of the spectrum, the hue of which depends on the activity of their long wave receptors; and a third region in between these two, which not only varies in intensity but also in hue, there being a so-called neutral point near the middle of the latter region, that is to say this part of the spectrum suitably adjusted in intensity may be caused to match ordinary white light. The different appearances of these three parts of the spectrum may be accounted for in the following manner. At the short wavelength end of the spectrum only one type of receptor is receiving adequate stimulation, the other receptor mechanism remaining in a quiescent state. At the long wavelength end of the spectrum similarly, this receptor mechanism is stimulated while the other one is inactive. But in the middle spectral region between these uniform areas there is differential stimulation of the two types of receptor, so that a change from a particular wavelength towards a longer wavelength causes a relative decrease in the intensity of stimulation of the short wavelength type of receptor, and at the same time an increase in the stimulation of the long wavelength type of receptor. Thus hue discrimination depends on the differential stimulation of the receptors which are active at a particular spectral region.

If a subject possessed three types of receptor, one responding predominantly to the short, another predominantly to the medium and a third predominantly to the long wavelength rays, then we should expect to find three spectral regions with nearly uniform hue, one at the short, one at the medium and one at the long wavelength end of the spectrum, but lying in between these there would be two spectral regions where a difference of hue is to be observed, this difference being caused by the differential stimulation of the receptor mechanisms. The one situated in the bluegreen between the short and the medium wavelengths would be due to the differential stimulation of the receptors responding to the long and medium wavelengths, and the other in the yellow would be due to the differential stimulation of the receptors for the medium and long wavelengths. This type of subject possesses

what is known as protanomalous trichromatic vision. The term trichromatic signifies that he possesses three types of receptor.

A subject who in addition to the two regions of good hue discrimination in the bluegreen and in the yellow described above also possesses good hue discrimination near the ends of the spectrum in the blue and in the orange would have to have two additional types of receptor, one most effective in the violet and the other most effective in the extreme red, so that he would have five types of receptor altogether. Referring now to the middle of the spectrum, the anomalous trichromat sees this as a nearly uniform green area, because his medium wavelength receptors respond to it strongly, whereas his long and his short wavelength receptors hardly respond to it at all. Now this is not the state of affairs in a subject who has normal colour vision, for he sees a number of different hues in the green, not so many as there are in the yellow and in the bluegreen, but still many more than those seen by an anomalous trichromat. In order to account for the improved hue discrimination possessed by a subject who has normal colour vision it is necessary to postulate that there is present in his retina an additional type of receptor whose response curve corresponds to the green. Such a receptor might be situated with its crest in the yellow or in the bluegreen or there might be receptors of both kinds. Thus in order to account in a satisfactory manner for the hue discrimination of subjects with normal colour vision it is necessary to suppose that there are six types of receptor at least, one in the orange, one in the green and one in the blue-violet, these being the receptors possessed by the anomalous trichromat, one in the extreme red, one in the extreme violet in order to give the improved hue discrimination towards the ends of the spectrum, and one in the yellow or in the bluegreen in order to improve the hue discrimination in the green.

SECTION XI: SOME OTHER VISUAL PHENOMENA

In the previous section of this paper we have considered some of the more important phenomena of colour vision, but there are

others which might well be considered before summing up the position so far as theories of colour vision are concerned.

When a small test object which has been examined by the fovea is slowly moved to right or left, keeping the gaze fixed so that the image of the object falls on progressively more peripheral parts of the retina, it is usually found to undergo changes of colour or to lose its colour entirely and then to appear colourless. The precise angle with the optic axis at which the colour is lost varies with the light intensity, the size of the test object and with the colour of the test object. Thus some colours, for example yellow and blue, are seen much further into the periphery than are other colours such as red and green.

Now, as pointed out in the first section, a yellow light can have many different spectral compositions. It may consist of yellow rays only; for example the yellow light of the sodium lamp is nearly monochromatic. But a yellow light may also consist of a mixture of red rays and green rays and contain no yellow rays at all. According to Young's theory a pure spectral yellow and a red-green mixture stimulate essentially the same receptors, that is the red receptors and the green receptors, except that the red-green mixture stimulates to some extent the blue receptors owing to the fact that the green rays which it contains fall within the response of the blue receptors and thus feebly stimulate them. If, however, a small quantity of blue light be added to the pure monochromatic yellow light, this mixture may be made to match the redgreen mixture precisely, the explanation on Young's theory being that the two yellows under these conditions stimulate the three receptors in exactly the same way. If these two yellows be used as test objects for examining the colour vision of the periphery of the retina, as described above, it is found that the monochromatic yellow with its small admixture of blue retains its colour far into the periphery. On the contrary the redgreen mixture which matches it loses its colour when it subtends a very much smaller perimetric angle. Thus the size of the field over which the pure yellow can be seen is far larger than that over which the mixed yellow can be seen. It is very

difficult to account for these facts, which were first observed by Abney, on the basis of the three-colour theory.

The explanation on the polychromatic theory is that pure yellow stimulates the yellow receptors principally, and the green and red receptors feebly. Now the yellow receptors are distributed right out to the extreme periphery of the retina, whereas the red ones and the green ones are not. Thus it would be anticipated that the field for pure yellow would be large, as in fact is found to be the case. The redgreen mixture on the contrary stimulates the red receptors and the green receptors strongly and the yellow receptors feebly. Now the red receptors and the green receptors are principally found in the fovea and in the part of the retina which immediately surrounds it, and are not found in the periphery of the retina. Thus the field for the red-green mixture would be expected to be a small one, as in fact is found to be the case.

When the sizes of the fields for other colours are examined it is found that, whereas red, green and violet have small perimetric fields, the yellow and blue have large ones. The large field for blue and the small field for violet are very difficult to account for on the three-colour theory, which supposes that both stimulate a single type of receptor which has its response at the short wavelength part of the spectrum. According to the polychromatic theory there are separate receptors, one variety having a maximum of response in the blue and another variety having a maximum of response in the violet, and these have different retinal distributions, the latter being limited to the centre of the retina and the former extending into the periphery.

Another visual phenomenon which should be mentioned concerns some of the effects which follow the exposure of the eye to a coloured light, which is called 'specific' adaptation in order to differentiate it from the 'general' adaptation which is produced by exposing the eye to a white light. On Young's theory it would be expected that adapting the eye to a pure monochromatic yellow light, or to a redgreen mixture which matched it, would produce identical effects provided that the intensities and durations of exposure and other conditions were

the same; but when the experiment is performed it is found in fact that they have very different effects. Thus after exposing the eye to pure yellow, the visible spectrum is found to be altered in the following ways: red and orange are replaced by crimson; yellow looks grey or even a greyish violet; the yellowgreen and green parts appear to be bluegreen, while the blue and violet parts are greatly darkened and can hardly be recognised. The appearance of the spectrum after adaptation of the eye to a redgreen mixture is much less altered; the red, orange, yellowgreen and green are hardly affected in hue and the blue and violet much less darkened than they are by adaptation to a pure yellow light.

Now the differences produced by the two adaptations are very difficult to account for on Young's theory. According to the polychromatic theory, on the other hand, the differences are brought about by the effects of the two yellows on the retinal receptors. Thus the pure yellow stimulates the yellow receptors predominantly, and this has an inhibiting effect on their subsequent activity, but it also affects the blue receptors in a similar manner, since the yellow and the blue form one functional unit, the activity of which is enhanced or depressed as a whole. The redgreen mixture on the contrary hardly stimulates the yellow receptors at all, and therefore produces much less depression on the unit which comprises the yellow and the blue receptors. In this way the supporters of the polychromatic theory account for the very slight changes which adaptation to the redgreen mixture produce on the visible spectrum.

Another phenomenon which might be mentioned is the discrimination of saturation of the different spectral colours. By saturation is meant the quantity of white light which each of the spectral colours appears to contain. This is usually measured by ascertaining by experiment the number of just distinguishable steps which can be found between a given spectral colour and white. This number is few for yellow, but considerably more for green and for red. Thus the saturation of yellow is spoken of as being low, and the saturation of red and green as being high. These differences in apparent saturation are very difficult to

account for on the three-colour theory because if it be true that yellow rays stimulate red receptors and green receptors, why has not yellow a high saturation just as red and green themselves have? The polychromatic theory finds itself confronted with no difficulties in accounting for the saturation discrimination either of subjects with normal vision or of those who are colour blind.

Many other phenomena of vision might be considered in a similar manner, but perhaps enough has been said in this connection in favour of the statement that supporters of the three-colour theory continually find themselves confronted by experimental facts for which they cannot account, or for which they account by means of additional hypotheses for which there is no experimental evidence, and which are frequently found to contradict one another. On the contrary, the polychromatic theory accounts in a satisfactory manner for all these phenomena.

SECTION XII: HUMAN COLOUR PERCEPTION

Some further details of the polychromatic theory will now be given, in particular of the way in which hue, luminosity and saturation are perceived under various conditions of lighting, of field size and of retinal situation. Thus, suppose the fovea to be used for the examination of a monochromatic red test object, which subtends at the eye an angle of two degrees and which is illuminated by means of white light at an intensity of one hundred foot-candles, then from Section VIII it would appear that the tricolour unit is in an active state and that this is assisted by the Y – B and by the R – BG – R units, both of which play, however, a subsidiary part under these conditions. The incidence of the image of the red test object on the retina will bring about stimulation (a) of the red receptors of the R – BG – R unit, (b) of the orange receptors of the tricolour unit, and (c) of the yellow receptors of the Y – B unit, to some extent. The combined effects of the nerve impulses which reach the brain as the result of this stimulation will be a sensation of red.

If a yellow stimulus be employed instead of a red one the receptors which receive stimulation will be (a) the orange receptors of the tricolour unit, (b) the yellow receptors of the T – B

unit, and (c) the green receptors of the tricolour unit, and the consequence will be the perception of yellow, but owing to the activity of the yellow receptors this will appear to have a greater degree of saturation than would be the case if the orange and the green receptors were the only ones to receive stimulation.

If a dichromatic yellow light consisting of red rays and of green rays be employed instead, then the red component stimulates the red, the orange, and the yellow receptors and the green component stimulates the yellow, the green, and the blue-green receptors, so that the dichromatic yellow light stimulates both the red and the bluegreen receptors of the R – BG – R unit, and since these are complementary in colour to one another the result will be that a considerable white light component will be present in the dichromatic yellow, causing it to be noticeably less saturated than a monochromatic one from which such a white light component is absent. It is in this manner that the difference in saturation is produced between a monochromatic yellow on the one hand, and a mixture of red rays and green rays with which a match is attempted on the other, and which requires in consequence the pure yellow to be desaturated by some admixture of blue light.

If, instead of the fovea, a part of the dichromatic zone of the retinal periphery be investigated, then in this case both a red and a yellow stimulus will cause activity of the yellow receptors of the Y – B unit, the result being the perception of a sensation of yellow. A big increase of light intensity will cause the fovea to possess a similar type of dichromatic vision, with the result that both red and yellow stimuli will look yellow.

The employment of the fovea at a low light intensity, or at a small visual angle, will introduce a different type of dichromatism with activity on the part of the R – BG – R unit by itself. In this case a red stimulus will act on the red receptors alone and will cause a corresponding sensation; but a yellow stimulus will act on both the red and the green receptors, causing, according to its wavelength, a sensation of neutral grey, pale pink or pale green.

Hue discrimination will attain its greatest perfection when receptors of all types of colour response are in a full state of

activity. The precise conditions for achieving this state of affairs have not so far been investigated, but present indications are that light intensity should not be excessive because a high light intensity has an adverse effect on the R – BG – R unit. Probably a favourable intensity would be 10 to 100 foot-candles and the eye should be fully adapted to this intensity. The field size should be quite large, for example 30 to 40 degrees, so as to include as large a population of receptors as is conveniently possible. This would include a number of the Y – B receptors which appear to predominate in the more peripheral parts of the retina. Under these conditions it would be expected that a monochromatic colour, for example yellow or bluegreen, would differ in saturation from a red-green or green-blue mixture by a greater amount than is found to be the case when a field size of two degrees is in use, as in the usual trichromatic colorimeter. A special large-field colorimeter is now being designed by the writer for the investigation of this point.

White light is perceived, as in the classical three-colour or four-colour theories, by the simultaneous stimulation of the receptors of different colour response in a certain ratio, which is subject to readjustment from time to time in order to allow for variations in the spectral composition of the light which is illuminating objects which the subject knows to be white. The perception of white is assisted, on the basis of the polychromatic theory, by the complementary relationship between the receptors comprised in the Y – B and the R – BG – R units. The appreciation of desaturation by the fovea depends then on at least two separate things: (a) the stimulation to a certain extent of all the receptors in the ratios which are usually associated with the incidence of white light on the retina; (b) the stimulation of the receptors of the Y – B and of the R – BG – R units in the ratios which cause them to act as complementaries to one another. In the case of an unsaturated colour there is also of course the specific stimulus of the colour itself which affects certain colour receptors only. The greater the general stimulation, and the less the specific stimulation, the lower the saturation of a given colour, and vice versa.

This brief review must not be thought to suggest that all major problems connected with human vision have now received an adequate explanation, because this is far from being the case. In one way the problem is made more difficult by this new theory, because, while the three-colour theory supposed only three separate visual mechanisms, the polychromatic theory postulates several more. We still know little about the way that light rays stimulate the retinal receptors, and we know even less about the conversion in the brain of nerve impulses into perceptions.

(Some papers have been published recently by Dr Motokawa (1949) which are of great interest from the point of view of colour vision theory. He first stimulates the eye by light and then investigates its sensitivity to an electrical stimulus. He finds evidence for a specific yellow receptor which is not noticeably present in the fovea, but is clearly marked in the more peripheral parts of the retina. In this respect and many others his work supports the polychromatic theory of human colour vision. All readers who are interested in this subject should consult these important papers, which deal not only with normal colour vision but also with colour blindness.) (*This final paragraph* was added April 1950.)

REFERENCES

Abney, W., and Festing, E. R. 1892. *Phil. Trans. Roy. Soc.* A., **183**, 537.

Ball, S., Collins, F. D., Morton, R. A., and Stubbs, A. L. 1948. *Nature*, **161**, 424.

Franklin, 1894. *Sitzungsb. d. k. Akad. d. Wissensch.* Berlin, p. 589.

Granit, R. 1947. *Sensory Mechanisms of the Retina.*

Hartridge, H. 1947. *Phil. Trans.* B. *Roy. Soc.*, **232**, 592.

Holmgren, A. 1889. *Skandin. Arch. f. Physiol.* Leipzig, **1**, 152; **3**, 253.

Lord, M. P., and Wright, W. D. 1948. In course of publication.

McKeon, W. M., and Wright, W. D. 1940. *Proc. Phys. Soc.*, **52**, 464-479.

Motokawa, K. 1949, *J. of Neurophysiol.*, XII, **5**, 291 and XII, **6**, 465 and 475.

Pitt, F. H. G. 1944, *Proc. Roy. Soc.* B., **132**, 101.

Purdy, D. McL. 1931. *Amer. J. Psychol.*, **43**, 541.

Sloan, L. L. 1928. *Psychol. Monogr.*, 38, No. 1.

Stiles, W. S. 1944, *Proc. Phys. Soc.*, **56**, 329.

Thomson, L. C. 1948. In course of publication in the *British Journa of Ophthalmology.*

Wilbrand. 1896. *Die Erholungsausdehnung des Gesichtsfeldes*, p. 42.

Willmer, E. N., and Wright, W. D. 1945. *Nature*, **156**, 119, 121.

Wright, W. D. 1946. *Researches on Normal and Defective Colour Vision* (London: Henry Kimpton).

ABOUT OUR CONTRIBUTORS

STANLEY BYARD is a Londoner and is in the Royal Naval Scientific Service. He is interested in Social History and has made a special study of the automaton.

H. O. J. COLLIER was born in Natal (Brazil) in 1912. After reading biological and chemical subjects at Cambridge in the early 1930's, he engaged in biological research, with a growing trend towards its medical aspects, at Cambridge, Belfast, Manchester and Liverpool Universities. In 1945 he designed and set up for Allen and Hanburys a pharmacological research and control laboratory which he now directs. He is the author of numerous papers; of two books, *An Interpretation of Biology* (1938) and *Chemotherapy of Infection* (in the press): and of a recent television programme. His wife is a radio actress, and they have three children.

D. W. DAVIES is a graduate of Imperial College, London and worked during the war on 'atomic energy' at Birmingham University and I.C.I. He is now employed on the A.C.E. electronic computor project. Recently he constructed as a novelty a machine which plays noughts and crosses (see Inset, No. 16).

C. L. GILBERT qualified in chemistry at the Royal College of Science, London in 1925 and has since worked with the 'Shell' Group. He was a Technical Adviser in Shell Petroleum Company Limited until recently, when he was transferred as Assistant to the Manager of Shell Chemical Manufacturing Company. One of his particular interests on the technical side has been the development of special products and by-products from petroleum. He is a Fellow of the Institute of Petroleum, and has contributed to their publications, as well as being Chairman of the Committee responsible for the production of abstracts.

A. R. HALL was educated in Leicester and at Cambridge, where he read history and became interested in science (and scientists). After the war he began work on the History of Science and is now a Research Fellow of Christ's College, Cambridge.

H. HARTRIDGE, F.R.S., is director of the Medical Research Council's Vision Research Unit and Gresham Professor of Physic at the ancient Gresham College in the City of London. He was formerly Professor of Physiology at the medical college of St Bartholomew's Hospital for twenty years from 1927, and before this he was Senior Demonstrator and Lecturer in the Special Senses at the Cambridge Physiological Laboratory, where much of his research was on the functioning of the senses. He has recently been appointed Emeritus Professor of Physiology in the University of London.

A. W. HASLETT is a Cambridge graduate, who has made scientific journalism his career. He is the Editor of *Science Today*, and author of a number of books, of which *Science in Transition* was recommended by the Book Society.

TERENCE PRICE is twenty-nine years old. He read physics at Cambridge, and then joined the staff of Admiralty Signal Establishment. When the war ended he became a member of the Atomic Energy Research Establishment at Harwell. His main hobbies are music and sailing.

G. RATTRAY TAYLOR originally set out to be a biologist but realised he was more interested in human relations, and now has a full-time job evaluating and editing the social research done by the Acton Society, of which Seebohm Rowntree is one of the trustees. He spent the last summer in America, studying group dynamics under the guidance of the Research Centre in Group Dynamics of the University of Michigan. He is now 39, married, and has published *Conditions of Happiness* (Bodley Head) and *Are Workers Human?* (Falcon Press).

V. B. WIGGLESWORTH, F.R.S., is Director of the Agricultural Research Council Unit of Insect Physiology, Reader in Entomology in the University of Cambridge, and Fellow of Gonville and Caius College. He was educated at Cambridge and St Thomas' Hospital (London), and later became Lecturer in Medical Entomology at the London School of Hygiene and Tropical Medicine.

BOOKS RECEIVED

MOLECULES IN MOTION by T. G. Cowling, F.R.S. (Hutchinson's University Library, 1950), 7/6.

This is rather a hum-drum book, and it is difficult to imagine for whom it has been written. It deals with the kinetic theory of gases, and how such properties as viscosity, diffusion, and the conduction of heat and electricity are quantitatively related to the motions of gas molecules, and though containing a fair amount of mathematics manages to stay within the confines of school algebra and avoid the calculus. It does also as the author admits, require acquaintance with elementary physics and chemistry, particularly the basic ideas of mechanics (force, momentum, etc.). This being so, the reader must clearly have some scientific education, and if he is sufficiently interested in classical gas theory to want to read a book about it, one would expect him to come either with a proper mathematical equipment, or with no mathematical interest at all, but not with the half-prepared mind Professor Cowling has written for. The non-mathematician in particular will be disappointed with the lack of excitement in the treatment, which savours of the textbook and appears at its worst in the earlier historical passages. These are written in such a level, distant, bored tone that they fail to strike the reader or to impress upon him the great turning points in the evolution of physics, and in particular why the application of mathematics produced the enormous advance it did. There was an opportunity here to illumine the whole relationship between physics and mathematics, and incidentally to elucidate the value and limitations of quasi-mathematical concepts like Force, Energy, and Momentum. But this opportunity has not been taken.

BIOLOGY AND MAN by Francis G. W. Knowles (Harrap, 1950), 10/6.

This is one of a new series of Torch books, designed for the general reader interested in up-to-date science. It talks first of the Biology of our surrounding, with a neatly-illustrated classification of the Animal Kingdom, followed by chapters on insects, parasites, the potato and

wheat, and a study of the underlying structure of cells and their functions.

The second part of the book deals with the Biology of Man: the anatomy, functions and history of the human body, with discussions on bacteria and disease, and a chapter on evolution.

There are a large number of very clear line diagrams, in some cases most effectively reproduced in reverse, white on black.

If you want to start an aquarium at home, or see a recently caught fish of a type thought to have been extinct for millions of years, or read of the curious ideas about reproduction held by the ancients, or even observe a three inch fly footprint, here you will find your answers.

ELEMENTS OF INTERNAL COMBUSTION TURBINE THEORY, by H. T. Adams (Cambridge University Press 1949), 16s.

A technical book for advanced students and engineers, requiring a good knowledge of mathematics, and some aerodynamics.

GLOSSARY

BLIND SPOT: Nerve fibres are insensitive to light rays, and only specialised cells in the retina of the eye are affected by light. The retina is connected to the brain by the optic nerve fibres, which come into the eyeball in one mass and splay out over the retinal surface to make contact with the receptors. Clearly there must be a 'hole' in the retina to let these optic nerves through, and where there is this hole there are no receptors and consequently a *blind spot*, as you can discover for yourself if you fix your eye on a distant object and move a tiny bit of paper held at arm's length in front of your eye at the same time: there is one position where it suddenly vanishes.

Since light entering the eye has to pass through splayed-out nerve fibres before it can reach the retinal receptors, and these fibres are quite thick, there is one very sensitive region of the retina, about half a millimetre across, called the *fovea*, where the fibre layer is thinned down to a minimum. When both eyes are used together, they are moved so that the object to be observed is focussed on the fovea of each eye: this movement is called a *fixation*, and the point in space where the axis of one eyeball would meet the other if prolonged is called the *fixation point*.

DETERGENT: A wetting agent, a substance which reduces the tension between oil and water and enables them to mix. Soap is an example of such a substance.

PROVENTRICULUS: The 'stomach' or front part of the gut (intestine) of an insect, specialised as a gizzard for grinding up the swallowed food.

THORAX: The chest.

VISCERA: A general name for all internal organs - heart, lungs, intestines, liver, kidneys, bladder, etc.

INDEX

Amateurs and science, 33
Anxiety and the group, 98
Approval by the group, 95
Astrolabe, 35, 36
Atomic nuclei, size, 16
  Theory of structure, 23
Azeotropic distillation, 141
Bacteria, isolated walls of, 79
Beat of insect wings, speed, 26
Bee's use of sugar, 27
Brahe, Tycho, 36
Chemical plant, 132
Chemical reactions in Shell plant, 129, 133, 135, 146
Chess-playing machines, 70, 74–7
Coal-tar distillation, 120
Coincidence unit, use, 18
Cosmic ray telescope, 19
Cosmic ray tracks, study, 10
Cracking of petroleum, 122, 125
  Furnace for, 134
Crystal irregularities, 91
Culture, nature of, 101
Curare, 107
Cyclotron at Berkeley, 15, 16
Delayed coincidence unit, 19
Design of instruments, historical periods in, 35
Distillation, 137, 140
  Azeotropic, 141
  Design of apparatus for, 138–9
Energy and oxygen consumption, 27
Flaxedil, 114
Fruit fly, different sugars and, 31
  Rest and flight capacity, 30
  Starch and fat in, 28
  Use of fuel in, 32
Furnace in petroleum processing, 134
Game, mathematical description of, 50 *et seq*
Games, theory of, 62
Games and economics, 64
Gas, natural, 121, 125
Geiger counter, 18
History of petroleum, 124
Hostility, 96
Hydrocarbons discussed, 120
Influenza virus, recent work, 82
Instrument makers in London, 38
Jacquet-Droz robots, 69
Juvenile delinquency, 96, 104
Labour relations, 97, 99
Laws of physics in subatomic transformations, 21
Lifetime of particles, 18
Mathematical approach, essence of, 63
Mesons, mass of, 17, 22
  Properties of, summarised, 24
  Radioactivity of, 13, 17
Microscope, George III's, 38
μ-meson, 12
Myanesin, 113
Natural gas, 121, 125
Neostigmine, 112
Neutrino, 21

Nuclei, atomic, theory of, 23
Observatory, idea of an, 36
Orrery, 38
Oxygen need and energy, 27
Penicillin, radioactive, 79
Petroleum, composition of, 121
  History of, 124
Photography in atomic research, 10
π-meson, 12
Poliomyelitis research, 83
Protein synthesis in cells, 84
Psycho, 71
Quaternary nitrogen, 109
Radar study of weather, 85
Radio study of winds, 89
Radioactive penicillin, 79
Rest and flight capacity in fruit flies, 30
Royal Institution, forerunners of, 37
σ-meson, 13
'Stationary flight' in insects, 30
Strategy in a game, 53
Sulpha drugs, mode of action, 78
τ-meson, 23
Thunderstorms, 86
Time Tank, 134
Tree as mathematical model for puzzles and games, 43 *et seq.*
Tycho Brahe, 36
Values in a game, 55
Vaucanson's robots, 66
von Kempelen's chess player, 70
von Knaus, writing machine of, 68
Winds, radio study of, 91
Winds in ionosphere, 88
Wing beats, speed in insects, 26

1 — 0 58

SCIENCE NEWS

Outstanding articles from recent issues include:

SCIENCE NEWS 12

The Light of Glow-Worms and Fire-flies	*Dr V. B. Wigglesworth*, F.R.S.
Infectious Diseases and the Country Doctor	*Dr W. N. Pickles*
Palaeomagnetism	*H. Manley*
Viruses	*K. M. Smith*, F.R.S.

SCIENCE NEWS 13

Ancient Peoples of the Andes	*G. H. S. Bushnell*
The Problem of Cancer	*Prof. E. C. Dodds*, F.R.S.
Experimental Embryology To-day	*Michael Abercrombie*

SCIENCE NEWS 14

Recent Progress in Television	*R. Howard Cricks*
The Origin of the Solar System	*Sir Harold Spencer Jones*, F.R.S.
The Electrical Activity of the Human Brain	*Dr William A. Cobb*

SCIENCE NEWS 15

Rockets	*D. Hurden*
The Blood of Water Fleas	*Prof. Munro Fox*, F.R.S.
Aspects of Crystal Chemistry	*P. R. Rowland*
Human Colour Vision	*Prof. H. Hartridge*, F.R.S.