SESSION 1

PAPER 4

THE MECHANISM OF HABITUATION

by

DR. W. ROSS ASHBY

BIOGRAPHICAL NOTE

Dr. W. Ross Ashby, born in London, studied medicine at Cambridge and London, took M.B., B.Ch. in 1928, M.D. in 1935.

He has since been engaged in research in psychiatry, specialising in the application of generalised dynamic principles (equilibrium, homeostasis, self-repairing systems).

His publications include Design for a brain (*Chapman & Hall, London; John Wiley & Sons, New York;* 1954) and An introduction to cybernetics (1958).

His present interests are: study of complex equilibria, especially in their topological aspects, as applied to the intelligent and adaptive aspects of the brain.

He is now in the Department of Research of Barnwood House Hospital, Gloucester.

THE MECHANISM OF HABITUATION

by

DR. W. ROSS ASHBY

SUMMARY

THE phenomenon of habituation, in which the response to any regularly repeated stimulus decreases, has not so far received any general mechanistic explanation. Its occurrence in systems of widely differing nature (nervous system, protozoa, bacterial cultures) and its occurrence against stimuli that the species has never previously encountered (smells or toxicities of newly synthesised chemicals) show that it cannot always be due to a specialised mechanism, shaped by natural selection.

It is here shown that when any system is subjected to a regularly repeated stimulus or disturbance, the successive responses, if they change in size, do not in general tend to become larger or smaller with equal probability: there is a fundamental bias in favour of the smaller. This bias holds over a great range of systems and disturbances. A theorem is proved, setting out the exact conditions for the decrease.

Some applications of the theorem are given. It is shown that habituation is specially likely to occur in any system made of parts that are rich in states of equilibrium. The theorem, without further assumption, provides an explanation of why any extraneous disturbance typically causes de-habituation.

To demonstrate the generality of the theorem, an example is given showing how habituation and de-habituation appear even out of a table of random numbers when the appropriate operations are applied.

It is suggested that this asymmetry is responsible for much habituation, including much of that shown by the cerebral cortex.

---

1. INTRODUCTION

ONE of the commonest phenomena in the living world occurs when an organism, given an innocuous and unvarying stimulus or disturbance repetitively, reacts briskly at first, then less actively, and finally perhaps not at

all. Known in behaviour as "habituation" and in perception as "adaptation", it has been recognised from time immemorial yet still lacks explanation. Only recently Sharpless and Jasper (1956, *ref. 10*) could say "Habituation ... has yet to be explained by any known neurophysiological principles".

A review of the subject need not be given here as it has been well reviewed by Humphrey (1933, *ref. 6*), Harris (1943, *ref. 5*), and Thorpe (1956, *ref. 11*). On one important matter they are agreed: habituation of typical form occurs in almost every form of life; in particular it appears as readily in forms having no neural apparatus as in the forms having a well developed brain. *Amoeba* shows it as freely as does the cat. *The phenomenon evidently does not depend on specifically neurophysiological details.* Its origin must lie in some property of much wider occurrence.

The possibility of "fatigue" as an explanation must be rejected. A number of workers (e.g. Humphrey, 1933, *ref. 6*; Danisch, 1921, *ref. 4*) have shown that the more violent the stimulus (with correspondingly larger response) the less does habituation occur and the later is its onset. This relation is just the opposite to what would be expected to happen with fatigue, in which the larger responses would lead to a more rapid decay.

The aim of this paper is not, in any case, to relate habituation to concepts of physiological or psychological type such as "fatigue" but to relate it to basic concepts of mechanistic type. I shall examine the phenomenon in the light of the modern logic of mechanism, so as to make use of the full generality of its methods (Ashby, 1956 *ref. 2*). Its use has several advantages. It enables the discussion to be as rigorous as we please. Yet this rigour is coupled with an extreme generality; for while it makes no assumption (as mathematical analysis so often does) of continuity, it includes this possibility as a special case. Further, by being abstract, *general* concepts need not be artificially restricted to those covered by the vocabulary of the neurophysiologist, or the electronic engineer, or the colloidal chemist, or other specialist. As will be shown below, the basic phenomenon of habituation can be identified over a very wide range of systems, and only a language that can range equally widely is appropriate.

Habituation was originally a physiologist's concept. To discuss it abstractly we must re-define it in abstract terms. It can be defined at several degrees of strength. The weakest degree ($\underline{H}_1$) requires only that the response to the later stimulus is *smaller* than that to the earlier. (Such is the form shown by the homeostat; *ref. 1* S.13/8). A stronger degree ($\underline{H}_2$) demands $\underline{H}_1$ and also that the change from the initial large response to the terminal small response shall pass through a *monotonic* sequence of intermediate sizes. An even stronger degree ($\underline{H}_3$) demands $\underline{H}_2$ and also that the fall shall be of approximately *exponential* form, with a rapid fall at first, flattening out. Related to $\underline{H}_1$, though not identical with it, is

degree $\underline{H}_4$: the *maximal* later responses are consistently less than the *maximal* earlier.

At first we shall be concerned with habituation only in the weak degree of $\underline{H}_1$; in Section 4 we will consider the stronger H_2 and $\underline{H}_3$. $\underline{H}_4$ is considered in Section 3.

Before we enter the arguments of Sections 2 to 4 it should be noticed that most of the arguments and implications are not reversible -- what is necessary is not usually sufficient, and *vice versa*. The implications in the two directions will be given in separate Sections.

SECTION 2

In this Section it will be shown that there is a fundamental asymmetry in mechanism, such that over a very broad class, containing most biological systems, there is a tendency for the response to a repeated stimulus to get smaller rather than bigger.

To see how this is so, let us start at the beginning and assume that the experimenter has before him some unknown system, totally unrestricted in nature; he is going to test whether it shows habituation. We will specify in detail the assumptions that he usually makes and the operations that he usually performs. We shall formulate these as abstractly and precisely as possible, taking special care to state explicitly those assumptions that he seldom mentions but that are usually taken for granted. We shall then see that over a wide class of system, general in the sense that it contains no *ad hoc* mechanism for habituation, he is certain to find habituation of degree $\underline{H}_1$.

(1) First to be defined is the operation Ω that the experimenter will apply to the system to test it for habituation.

Postulate I; The operation Ω is the following sequence of sub-operations:

(1) The experimenter allows the system to come to rest.
(2) He applies the given stimulus or disturbance.
(3) He allows the system, without further disturbance, to display its response, which he records.
(4) He repeats the cycle (1)-(2)-(3)-(1)- etc. until the responses become invariant (or approximately so) at the "terminal" response.
(5) He defines a numerical scale on which to measure "size of response".
(6) He compares the size of the terminal response with that of the initial.

Each of these sub-operations, as we shall see below, is significant.

(II) Next we must take account of the fact that the experimenter does not typically apply the test for habituation to a system that is unpredictable or chaotic in its behaviour. He restricts his tests to systems that are known to be law-abiding, or that may reasonably be assumed to be so. Formally:

Postulate II: The system under test is a machine with input (*ref. 2* S.4/1), i.e. such that its present state and its present surroundings (or values at its input) determine uniquely the state it will go to next. Abstractly, there is a set E of its states and a set A of its conditions or input-values, and its behaviour corresponds to a single-valued mapping, T, of the product set $A \times E$ into E. T specifies what the machine does, how it behaves.

If the system is law-abiding in the probabilities of its transitions rather than in the transitions individually (i.e. if it is a Markovian machine; *ref. 2* S.12/8) the deductions below would have to be re-stated in probabilistic form. Only modifications in detail would be needed; as they are almost obvious I need not give them here as they would only obscure the theme.

The sub-operations of Ω can now be converted to algebraic form (a necessity for rigorous discussion) as follows (with the numbering of Postulate I):-

(1) "Let the system come to rest" means allowing U to operate, where $U = \lim_{n \to \infty} T^n$, and where $T^2(x) = T(T(x))$, etc. If T has no cycles, only states of equilibrium, U maps E into Q, where Q is the set of T's states of equilibrium, i.e. such states x as satisfy $T(x) = x$.

(2) A stimulus or disturbance D, acting impulsively, displaces x to a well-defined state $D(x)$. Thus D maps Q into E, if D operates only when the system has come to rest under T.

(3) The return to rest after D implies the operation of U.

(4) The triple (2)-(3)-(1) is the operation $U \circ D$ a composite operator (Bourbaki, 1951, *ref. 3*), such that $(UD) \circ (x) = U(D(x))$; call it Σ. Σ maps Q into Q, if applied only when the system has come to rest under T. For the responses to become invariant (if Σ has no cycles) the system must have reached a state q such that $\Sigma(q) = q$, i.e. a state of equilibrium under Σ.

(III) Next, the experimenter does not usually test for habituation a system that is undergoing some obtrusive cycle of activity. Formally:

Postulate III: The transformation T is assumed to have no cycles, only states of equilibrium (ref.2 S.5/4).

Another reason for stating this postulate explicitly is because Rubin and Sitgreaves (1954, *ref.9*) have shown that if the set of machines obtained by taking all possible mappings of E into E is taken as a sample space (so that its elements are those of the set E^E -- Bourbaki, 1951, *ref.3*) then as the number of states in E increases, so does the probability of any trajectory ending in a *state* of equilibrium tend to zero. (It becomes almost certain that the trajectory will end in a cycle). If therefore we want to talk of machines with some generality yet wish the class discussed to have equilibrial states rather than cycles, we must make clear that we are speaking of some subset of the class of all machines. The assumption is made here simply because the experimenter usually restricts himself to tests on such a subset.

(This Postulate, and those that follow, are not intended to dogmatise about what will be found empirically in the world of real systems, but simply to state precisely what is being discussed in this paper).

(IV) Next, the experimenter would reject as peculiar any system that produced, under the cycle of suboperations (1)-(2)-(3)- of Ω, an invariant *cycle* of responses R_1-R_2-R_3-R_1- etc., rather than a single response repetitively. So, formally:

Postulate IV: Σ is assumed to have no cycles, only states of equilibrium.

Now Σ is a composite operator, for $\Sigma = U \circ D$, where U is $\lim_{n \to \infty} T^n$. For a state q to be one of equilibrium under a composite operator, a *relation* must exist between how D and U affect it. This relation is the point of the paper. What it is will first be sketched picturesquely, and then formally. (The picturesque statement will show its obvious relation to habituation H_1, but leaves obscure exactly what is being assumed; this fault will be corrected in the formal statement).

Let the system's states E be represented by the points within the area of *fig.1*. For clarity, group together all those states that come to the same state of equilibrium (under T); such a set is a "confluent". The whole set E is thus partitioned into confluents, each containing one state of equilibrium, (shown in the *figure* as a heavy dot). The illustration shows nine states of equilibrium each surrounded by its confluent. From each state of equilibrium draw an arrow to show how the representative point will be displaced when the stimulus or disturbance D acts. (An arbitrary set has been drawn in the *figure*).

Now by merely tracing the sequence of events, it is easily verified that the system, if started from any point in the left-hand two-thirds, must end in either the top left or bottom left confluent. Thus, if started

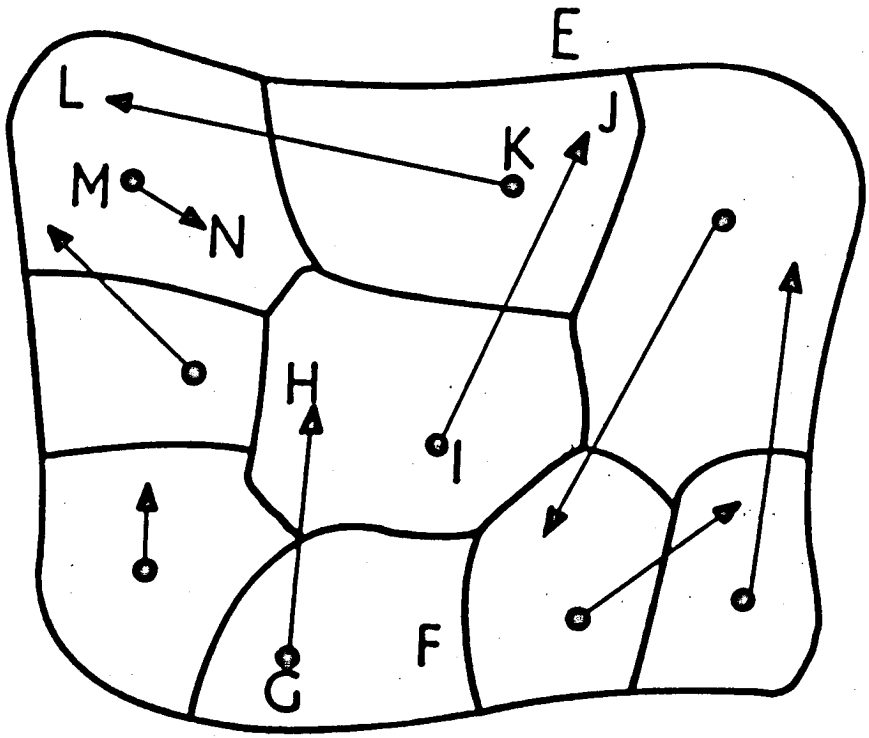

Fig. 1.

at *F*, suboperation (1) of Ω (i.e. *ν*: "let the system come to rest") will take it to *G*; the stimulus *D* then takes it to *H*; then "come to rest" takes it to *I*; and so it goes on, changing from confluent to confluent. When it gets to *K*, however, it is trapped. *N* does not take it outside *M*'s confluent, so it comes back to *M*. Thus the point is trapped in the first confluent such that *D* cannot move it out. Clearly, confluents with short arrows, such as that from *M* to *N*, are more likely to be terminal than those with long ones, such as that from *K* to *L*. So the sequence of arrows, caused by *D*, has a tendency to finish on an arrow *of less than average length*.

Thus the very process of *testing* for habituation, of applying Ω, so acts on the system that by the time the invariant response has been elicited the system's state is no longer an average one but is one such that *D* displaces it less than averagely. Thus, to say "I tested this system for habituation, and I eventually found it to show a diminished response" is to verge on the tautologous.

(Either the representative point ends at such a confluent as M's, or it must enter a cycle made by the disturbance, as on the right of *fig. 1*. The latter possibility has been excluded by Postulate IV).

The necessary theorem will now be stated and proved formally. It will be stated in terms of arbitrary and discrete states, for in this form it has simplicity and the greatest possible generality. It must not be thought, however, that the theorem is true only for systems, such as digital computers, that change by finite jumps; on the contrary, the states may be as close as is wished, and may thus, in the limit, represent the values of continuous variables. (It should also be noticed that the states are not here analysed into components, i.e. the whole system is not regarded as being made of parts; the complication is avoided in this Section as it is not necessary).

Definitions and postulates

(1) A set E of states x is mapped into itself by a single-valued transformation T (Postulate II).
(2) Q is the set of states in E that satisfy $T(q) = q$.
(3) T has no cycles (Postulate III).
(4) D is a single-valued transformation mapping Q into E. (The transition from x to $D(x)$ will be called the "displacement").
(5) U is the transformation $\lim_{n\to\infty} T^n$, mapping E into Q.
(6) Σ is the transformation $U \circ D$, so that $\Sigma(x) = U(D(x))$; it maps Q into Q.
(7) Σ has no cycles (Postulate IV).
(8) Given a state q in Q, the "T-confluent containing q" is the set of all states in E obtainable from q by repeated application of T^{-1}. (It consists of all those states in E that come, under T, finally to q).

Theorem: For a state x to be equilibrial under Σ, it is necessary and sufficient that x and $D(x)$ lie in the same T-confluent.

Proof: (1) Assume x is equilibrial under Σ. Then $\Sigma(x) = x$, and $U(D(x)) = x$, by definition. Now if $U(a) = b$, a must lie in the same T-confluent as b; so $D(x)$ and x must lie in the same T-confluent.

(2) Assume x and $D(x)$ lie in the same T-confluent C. D operates only on states in Q, so x must be a state of equilibrium under T, and it lies in C. $U(D(x))$ is in the same confluent as $D(x)$, so $\Sigma(x)$ lies, with $D(x)$, in C. Σ maps into Q, so $\Sigma(x)$ is also a state of equilibrium under T. But any confluent can contain only one state of equilibrium; so $\Sigma(x) = x$, and x is equilibrial under Σ. (Q.E.D.)

(V) So far, no topology or metric has been assumed over the states, for such an assumption is not necessary within the theorem. In order, however, to introduce the concept of "size", required by (5) of Postulate I, we must

now suppose that some metric, some measure of "distance", holds over the states, so that states have the relation of being "near to" or "far from" each other. (In most practical cases, of course, some metric presents itself as being the obvious one; often it is that in which the system is describable, in three dimensions of space, in units of mass, length, and time).

Given the metric, we must next say, explicitly, how the sizes of the displacements (from x to $D(x)$) are related to the sizes of the responses. We naturally assume that if the displacement is zero the response will be zero (for the system has not undergone any real change), and that an increase in the displacement will show as an increase in the size of the response. Formally:

Postulate V: The size of the response is a positive monotonic function of the size of the displacement, the zeros corresponding.

The point of the theorem (so far as we are concerned) is that whereas the initial response is based on *any* displacement (e.g. on any of the arrows in *fig. 1*), the terminal response can be based only on some displacement that leaves the point inside the confluent. If "in the same confluent as "x" has any implication of "near to x", then the terminal response will tend to be less than the initial, and some habituation (of degree H_1) will tend to occur.

For rigour, the word "tend" requires definition. There are obviously various ways of stating the relation, of various utilities in various applications, and the reader may prefer to formulate his own way. Whichever way is used must be compatible with the fact that the Postulates given so far do not allow the deduction that the terminal response *must* be smaller than the initial (for arbitrary T and D); for nothing prevents the initial displacement from being small yet taking the representative point to another confluent, and then a later large displacement leaving the point in the same (later) confluent. Clearly, our interest is in the fact that the terminal displacements are a subset of the total displacements, biassed in favour of shortness. One way of expressing the relation rigorously is as follows.

To compare the distributions of the initial and terminal displacements, let a "typifying" function f be defined (averaging, taking the mode, taking the maximum, etc.) which maps the set of distributions into the scale of sizes of response, so that each distribution indicates its "typical" size of response. Now if the diameter (topologically) of the largest confluent is l, every displacement in the terminal distribution must be less than l in length. If now, over a sequence of systems (over a sequence of T's) l tends to zero, the terminal displacements must all tend to zero also, and so therefore must the size of the typical response. So whatever the size

of the typical response initially (provided it is non-zero), the typical terminal response will become less than it, and habituation of degree H_1 will be shown.

(This demonstration that there is a fundamental *asymmetry* involved when a composite operation leads to equilibrium is the point of the paper. The theorem, however, leads on to several interesting and easy applications; the remainder of the paper will be concerned with them).

SECTION 3

In the previous Section, we assumed that the experimenter applied the operator Ω to *test* for habituation, and we saw that a very general class of systems, not characterised by possession of any specialised mechanism, would show it. The class was characterised chiefly by having "in the same confluent as x" correlated with "near x". The relation of this property to habituation is further illustrated by consideration of it in the inverse direction. (To make sure that the argument is not circular, and to make sure that we are keeping what is necessary distinct from what is sufficient, we make here a new start).

Starting, then, *ab initio*, let us consider what we can deduce about a system when all we know of it is that it has been tested for habituation and has been found to show it.

This statement must imply that the operator Ω, defined as before in Postulate I, has been applied to it as test. That the test was made implies that the system must have been behaving with some regularity; so Postulate II -- that the system is a machine with input -- is evidently applicable. It implies that the system showed no important cycles when at equilibrium (Postulate III); and it implies that the response to Σ gave a single response rather than a cycle (Postulate IV). (By the preceding Section we could now deduce that *if* T had l less than some value, etc.; but this is not our direction. It is now *given* that the system shows habituation, and we want to know what we can deduce about the particular system given).

Nothing has been said so far about whether the habituation shown by the system is against a particular disturbance D, or against all possible disturbances D_1, D_2, D_3, ... etc. from some class. The two cases will be considered separately.

The first case can be dismissed briefly. It is similar, in the postulates that apply to it, to that discussed in Section 2; and as these Postulates imply that it is likely to show habituation, the information that it does tells us little. So little can be said in this case about T.

The second case, when the system habituates against *any* of a set of D's, is more interesting. This is what is implied when it is said that the

Protozoa show habituation freely, with no specification of the disturbance to be used; for the statement implies that habituation will be shown against any of a wide class of disturbances. An interesting deduction can be made when the habituation is of degree H_4: when all the terminal responses have size less than λ, where λ is much less than the maximal size of the initial responses. (H_1 and H_4 are hardly independent, but will be so treated here, for the sake of rigour). Our assumption is now, formally:

Postulate VI: The system shows habituation (H_4) against each of a set of disturbances that cause, between them, all possible displacements from each state of equilibrium.

We use the fact that if a system is such that all its terminal responses are smaller in size than λ, and if the typifying function f is such that the limitation to λ implies that every displacement is through some distance less than l', then Postulate IV (with the others of this Section) implies that *no confluent in T. can* have a diameter exceeding l'. (For suppose one confluent had a greater diameter; then among the set of displacements would be one (by Postulate VI) longer than l' and lying within the confluent; it would give rise to a response that was both terminal and larger than λ, contrary to hypothesis). So if a system shows habituation H_4 against a wide *class* of disturbances, then, by the theorem, this is evidence that the system's T-confluents are all small. Since, if the number of states does not vary, smallness of confluents implies largeness in their number, the confluents must evidently be numerous. And as each confluent must have a state of equilibrium, it follows that the system's states of equilibrium must be numerous too.

From this fact we can draw an inference about the parts that compose the whole. So far no reference has been made to any parts of which the whole might be made. The "states" referred to so far have been those of the whole system, assumed to be identifiable as such without having to be built up from components. Statements about equilibria in the whole, however, have implications about those in the parts, and *vice versa*. The basic relationship has been described in *ref.2* S.5/12: the whole is at a state of equilibrium if and only if each part is at a state of equilibrium in the conditions provided by the other parts. (The next Section takes up the subject in more detail). Thus if the whole is in equilibrium at a certain state, any particular part must also be in equilibrium at its component-state. So if, over a given sample-space of states, the whole has probability p of being at equilibrium, then the probability π that any given part is at equilibrium cannot be less than p. As we saw above, p is high; so *the π's must be higher.*

This Section thus shows that if a system has been tested for habituation against the set of all possible disturbances and has been found to habituate (H_4) to all, then, with the stated minor qualifications, the

theorem implies that, if made of parts, *the individual parts must have a high proportion of their states equilibrial.*

SECTION 4

As a second application, we can now consider the converse: if a whole is built of parts having a high proportion of their states equilibrial (but not otherwise restricted), will it show habituation when tested by the operation Ω?? (As was said earlier, this converse can by no means be taken for granted).

Let us first make precise what is meant by a part being "rich in states of equilibrium". As we are discussing law-abiding, and not chaotic, parts, each part i will itself be a machine with input (*ref.2*, S.4/1) and will therefore itself behave in accordance with some mapping V_i of the product set $G_i \times B_i$ into B_i, where B_i is its set of possible states and G_i its set of possible input states. When G_i is at a particular state, at g say, there is defined the partial mapping V_{ig} of B_i into B_i generated by V_i and corresponding to g (Bourbaki, 1951, *ref.3*). Then if b is an element in B_i, b is "a state of equilibrium for the input state g" if and only if $V_{ig}(b) = b$. (This definition implies that, if the system is composed of continuous parts or variables y_1, y_2, ..., y_n, behaving in accordance with equations of form

$$dy_1/dt = \phi_1 (y_1, \ldots, y_n)$$

$$\ldots\ldots\ldots\ldots$$

$$dy_n/dt = \phi_n (y_1, \ldots, y_n)$$

then part i is in equilibrium at the state $(y_1, \ldots, y_i, \ldots, y_n)$ if and only if $\phi_i (y_1, \ldots, y_i, \ldots, y_n) = 0$).

For a part i to be "rich" in states of equilibrium, it is implied that the elements in $G_i \times B_i$ satisfying $V_{ig}(b) = b$ are numerous. (Nothing is implied about how they shall be distributed over $G_i \times B_i$). So if the set $G_i \times B_i$, or the domain of the ϕ_i's, is taken as the sample space, a probability can be defined -- that part i should be in equilibrium.

As was said earlier, equilibria in whole and part are related. Thus arises the possibility of relating the richness in equilibria of the parts with that of the whole. Unfortunately, a full treatment of the relation requires attention to how the states are distributed in each part over $G_i \times B_i$, for each i. These combinatorial complexities make the subject hardly worth full treatment here, though the following fact will be required. Suppose each part has a fraction π of its states equilibrial, and that n such parts are coupled (*ref.2*, S.4/6). It can be shown that the whole's fraction p (of states that are equilibrial) can vary from

$1-n$ $(1-\pi)$ (from zero if this function is negative) up to π, according to how the equilibrial states are disposed in the product sets (in the canonical representations) of the parts. A value of p outside this range is combinatorially impossible. If π is near 1, and n is large, this is a broad range, and implies little about the value of p. On the other hand, however large the value of n, a value of π near enough to 1 can force p to be arbitrarily near to 1.

After this preparation, let us start *ab initio* again and assume

Postulate VII: The whole is made of parts each with π so high that the whole has p near to 1, but not otherwise restricted. It is tested for habituation (thus implying the application of Postulates I to V) -- what will be found?

As π tends to 1, so does p. The number of confluents rises, and the average number of states per confluent falls towards one. So the restriction of the terminal displacement to within a confluent (by the theorem) makes its size tend to zero. Whatever the average (or other typical) size of the initial displacement, that of the terminal displacement will become less than it; so the appearance of habituation (H_3 and H_4) certain on the average. Thus, *a whole made of parts rich in equilibria tends to show habituation.* (This result is true even for an arbitrary T-D pair, provided that responses sufficiently small to be terminal exist in D's set of displacements -- i.e. providing D is not everywhere too "strong" in its effects).

Systems made of parts that are rich in equilibria have further interesting relations to habituation. A case of special interest occurs when the parts are not related to one another in some specially arranged way -- when they have been selected, say, as random samples from a distribution of parts, coupled in a way also sampled from a distribution of ways of coupling. Let us then consider the case:

Postulate VIII: In each part the states of equilibrium are distributed independently of one another and of how the equilibria are distributed in the other parts.

As the richness increases (as π tends to 1) the whole tends to be cut into functionally independent subsystems (*ref.1*, S.14/15; *ref.2*, S.4/10); for as the parts on any trajectory, spend more and more time not changing, so do they become unable to transmit variety from part to part within the system.

Suppose now, as is not uncommonly the case, that the disturbance D is actually a vector with components $D = (\delta_1, \delta_2, \ldots, \delta_n)$, so related to the n parts of the system that each component of D acts on some part of the system, so that, e.g.

$$D(y) = (\delta_1(y_1), \ldots, \delta_n(y_n)).$$

This *Postulate IX* is in one sense a simplifying assumption, though it seems more complex algebraically. We are now assuming that how D affects part y_1, say, depends only on the state of y_1, and not also on the states of all the other parts. This means that D affects the whole only so far as each part is affected individually by D. (Often many of the δ's are identity operators, so that D has a non-zero effect only on certain parts; i.e. so that D affects directly only a portion of the whole system).

It is now possible (and the phenomenon will become more evident as π tends to 1) for local regions of the whole to reach equilibrium under Σ while other regions are being changed by it; and they may, for some time, be able to retain their state of equilibrium. Then a final state of equilibrium under Σ (and the invariant terminal response) can be arrived at by degrees, by accumulation of local equilibria (*ref.1*, S.12/5; *ref.2*, S.4/21). If now the system allows

Postulate X: The size of the response is a positive monotonic function of the number of parts changing after each application of Σ, then in the limit (as π tends to 1) the number of parts not in equilibrium can only fall, and the system will thus show the stronger (H_2) degree of habituation.

Finally, if the subsystems go to equilibrium *independently*, so that the number that go is some fraction (approximately constant) of the number of parts still not at equilibrium, then the number not there, and the size of the response, will fall in approximately exponential fashion, thus showing *habituation in the full degree of* H_3.

SECTION 5

The propositions of the previous three Sections have shown that habituation not only can appear but must appear in classes of systems much wider than those with specifically neurophysiological properties. They do not even need to be living, for any system that satisfies the Postulates will show it. In fact, the systems that show it are so general that the process is readily demonstrable on systems formed from random numbers, - a "Monte Carlo" method - all that is necessary being that one must try to adjust the process so that it is neither so simple as to be trivial nor so complex as to be excessively laborious in computation. Here is one that shows the phenomenon readily.

A "part" or "variable" of ten states is given its value by taking a number from a defined place in a table of random numbers. The digit in that place gives the value, and its successive transforms are found by the digits that occur successively below it in the table. Four such variables form a "whole" with 10,000 states. Thus if the table shows 3119 we
8292

interpret it as defining a T-transformation in which state 3119 is changed to state 8292; also that part y_1, at state 3 in conditions $(y_2, y_3, y_4) = (1,1,9)$, changes to 8. (Should 3119 occur again, it must, of course, transform to 8292, not to the number that follows it on the second occasion; but this complication occurs rarely in the process defined below and can be ignored without appreciable error). Thus a T-transformation can readily be defined, devoid of any special structure.

Such a T, however, has on the average only one state of equilibrium, while the T we require must have a fairly high value of p. So the definition is modified. p is made 0.2401 by assuming that all states are stable if they contain no 0,1 or 2. Then the whole progresses to an equilibrium simply by going from state to state (of four digits) until a state occurs having no 0,1 or 2. Thus the start of Kendall and Babington Smith's (1951, *ref. 7*) Table gives

2315
0554
1487
3897

The last state is equilibrial. Thus is defined a T-transformation specialised only in that $p = (0.7)^4 = 0.2401$.

A suitable definition for D was found to be to restrict it to δ_4, and to act on (y_1, y_2, y_3, y_4) so that y_4 was changed by the transformation:

$$D = \delta_4: \quad \begin{matrix} \downarrow 3 & 4 & 5 & 6 & 7 & 8 & 9 \\ 0 & 4 & 0 & 6 & 0 & 8 & 0 \end{matrix}$$

(When at T-equilibrium, values 0,1,2, do not occur). So D, acting on 3897 gives 3890, which is not in equilibrium, and so starts a new trajectory under T;

3890
9731
1174
4336

The last state is an equilibrium for T and for D; it is thus also an equilibrium for Σ.

The response to each application of D is measured by the number of steps taken before equilibrium was restored (corresponding, in *fig. 1*, to the distance from the head of the arrow to the next ensuing equilibrium). Thus the example gave, for its successive responses: 3, 3, 0, 0, ...

The process so far illustrates those of the theorem in Section 2. To show habituation of degree H_3, the processes of Section 4 were followed. A hundred initial states were taken (corresponding to a large system, with many more than 400 parts, being disturbed by D in a hundred places), and to each was applied the operation Ω. The grand "Initial Response" was taken as the average of the hundred individual initial responses; the grand

"Second Response" was taken as the average of the hundred individual second responses; and so on. These average (grand) Responses correspond to what would be observed if a system, much divided into independent local subsystems by a high value of π, were subjected to a D that affected a hundred of them locally, producing a Response that was some function (such as the sum, or average) of the hundred individual responses.

The result is shown on the left half of *fig. 2*. Each column shows the size of the average Response to D, as Σ is applied again and again.

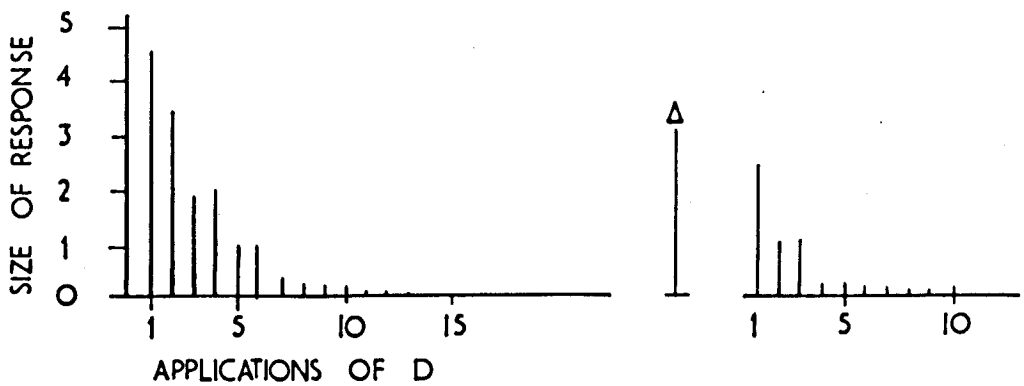

Fig. 2.

It shows typical habituation of degree H_3: the average Response to the first application of D is 4.3; to the second 3.5; and the Responses diminish eventually to zero.

SECTION 6

As ubiquitous as habituation is de-habituation: the fact that a system, habituated to a monotonous stimulus, if given any new stimulus or disturbance, will so change that when the monotonous stimulus is given again it evokes a response greater than those it had shown at the end of the habituation.

This phenomenon occurs widely over the biological world, and is also remarkable in that the interrupting stimulus or disturbance may be of almost any nature. Clearly, in most cases some very general type of process or mechanism is involved. So far, no process or mechanism has been

recognised as being responsible for it. A few explanations have been proposed for particular systems (e.g. an inhibition of an inhibition in the neurophysiological) but these are hardly satisfactory when the same phenomenon appears in systems of quite different type.

If habituation should often be due to the process of Section 2, a reason is readily seen why de-habituation should occur as often. To translate to abstract form, the new stimulus will be a new operator, Δ, mapping the equilibria of Σ into E. That it is "strange" means that Δ (as a set of arrows in *fig. 1*) will be arbitrarily different from D; that it is "strong" means that the arrows will not be so short as to make its effect negligible. Suppose then that the system has become habituated under the repeated action of D (has reached the top left confluent in *fig. 1*, say). An arbitrarily new disturbance Δ is applied -- what will happen?

Δ will displace the system (the representative point) from M to somewhere in the set E of possible states. Should the point fall within M's confluent, the subsequent application of D will evoke only the habituated, small, response; but should the point fall outside the confluent, then the operator Δ has undone the going to equilibrium of Σ, and the new response to D will be, in general, one from the initial, unrestricted, distribution of responses. Thus any new operator Δ does not act symmetrically on the terminal responses: *it tends to free the response rather than to intensify the constraint.*

In biological systems, the processes may often be like those of Sections 3 and 4, the whole having a natural metric and being made of parts rich in equilibria. When this is so, some quantitative relations are likely to appear.

First, Δ can vary on some scale from weak to strong, the weak causing small displacements and the strong large. Obviously, the longer the displacement, the greater is Δ's chance of getting the representative point away to a new confluent and thus of restoring the next response (to D) to its initial value. Thus, the stronger is Δ, the more will the next response return to the initial size.

Secondly, suppose D and Δ have components and a metric of their own, so that the resemblance between them can be measured by the "distance" between them. If Δ is very similar to D it will have similar effects, and will be represented in the phase space by a set of arrows similar to those of D; such an operator will have little tendency to shift the representative point out of a confluent if D cannot. But if Δ is much different from D, Δ's arrow may well take the representative point out of the confluent and destroy the habituation. Thus, the more Δ differs from D, the more effective will it be as a de-habituator.

The theory thus leads, without further modification, to a simple explanation of the phenomenon of de-habituation.

A related property that finds a ready explanation is that the habituated biological system, if left alone, usually recovers its responsiveness "spontaneously". This recovery would occur in systems for two reasons, which are related. First, if being "left alone" means that the system is, in fact, left subject to many little disturbances Δ_1, Δ_2, ... etc., too small to be noticed by the experimenter, then the recovery may be due to the representative point going on a "random walk", under the action of the little disturbances. If it arrives at a new confluent, the response to *D* will be freed from its terminal constraint and will be restored to its initial size. Thus, bombardment by small disturbances will tend to restore responsiveness.

It may also happen that some factors, internal to the system but not noticed by the experimenter, are acting as parameters to the system, are changing slowly in value, and are thus changing its field (*ref. 1*, S.6/3), its *T*-transformation, and the distribution of its confluents. Such, for instance, may happen as an experimental dog grows hungry. When the confluents are changed, a representative point that was previously trapped may be no longer trapped in the new confluent. Thus, secular change in an unobserved parameter may lead to a "spontaneous" recovery of responsiveness.

These facts can be readily illustrated by the "Monte Carlo" method. After the system of *fig. 2* (actually a hundred subsystems) had reached equilibrium under Σ, the operator Δ was applied once:

$\Delta = \Delta_4$:

y_4 =	3	4	5	6	7	8	9
↓ y_4' =	0	0	0	0	9	7	8

(It came after operator *U*, so the state operated on by Δ would be four digits lacking 0, 1 or 2).

As a result of Δ's action, about 4/7 of the habituated subsystems were thrown out of equilibrium under Σ and were forced to follow a trajectory under Σ until stable again. The result is shown on the right of *fig. 2*. The column marked Δ shows the response (under *T*) after it had been disturbed by Δ. The next column to the right shows the response to an application of *D*: previously fallen to zero, it is now restored to the extent of 1.8, and thus demonstrates de-habituation. (Then as Σ is applied again and again, the response falls off to zero).

This Δ displaced only about 4/7 of the hundred subsystems' equilibria. Had it been stronger, e.g. had it been

y_4 =	3	4	5	6	7	8	9
↓ y_4' =	0	0	0	0	0	0	0

it would have disrupted all the hundred equilibria, and would have restored the response to its initial value.

It is not, of course, proposed here that the basic process of the theorem is responsible every time the phenomenon of habituation appears. Showing that Postulates I to V lead to habituation does not exclude the possibility that other processes may also lead to the same phenomenon. Sometimes, especially in sensory adaptation, the phenomenon may be related to survival, either positively or negatively; then natural selection will interfere to give the system what properties are best, departing from those of Postulates I to V. So may develop such specialised, and doubtless more efficient, systems as that of Pringle (1951, *ref. 8*).

On the other hand, when the Postulates hold, habituation follows necessarily. Since the conditions in the cerebral cortex and in protoplasm approximate, from some points of view, to the conditions of the Postulates, it seems likely that much of the unspecialised habituation shown by them is primarily due to this process. With this explanation available, the onus of proof now passes to those who wish to propose that some particular act of habituation is not due to this very generally available process.

The theorem suggests that our attitude to habituation has been basically mistaken. We treated the organism in a particular way, obtained habituation as an outcome, and then asked: what peculiar property in the organism is responsible for this outcome? The theorem suggests that the question itself is wrong, for it assumes that the peculiarity is to be sought in the organism. In fact, the peculiarity, and the reason for habituation, lie at least as much in the particular sequence of operations used by the experimenter. These have sufficient character or pattern to impose some pattern on the response of the system; this pattern shows as habituation, all that is required of the system being that it should be in the class that does not totally destroy such patterns.

REFERENCES

1. ASHBY, W. ROSS. Design for a brain. *London, Chapman & Hall,* (1952).
2. ASHBY, W. ROSS. An introduction to cybernetics. *London, Chapman & Hall,* (1956).
3. BOURBAKI, N. Théorie des ensembles; fascicule de résultats. *A.S.E.I., No. 1141. Paris, Hermann & Cie,* (1951).
4. DANISCH, F. *Zeit. f. allg. Physiol.* 1921, **19**, 133.
5. HARRIS, J. D. *Psychol. Bull.* 1943, **40**, 385.
6. HUMPHREY, G. The nature of learning. *London, Kegan Paul,* (1933).
7. KENDALL, M. G. and BABINGTON SMITH, B. Tables of random sampling numbers. *London, Cambridge University Press,* (1951).
8. PRINGLE, J. W. S. *Behavior,* 1951, **3**, 174.
9. RUBIN, H. and SITGREAVES, R. Tech. Rep. 19 A; Applied Maths. and Stats. Lab.; *Stanford Univ., California,* (1954).
10. SHARPLESS, S. and JASPER, H. *Brain,* 1956, **79**, 655.
11. THORPE, W. H. Learning and instinct in animals. *London, Methuen,* (1956).

DISCUSSION ON THE PAPER BY DR. W. ROSS ASHBY

DR. GREY WALTER: As Dr. Ross Ashby said in his paper, the process rather ineptly called 'habituation' is basic to any consideration of learning. Learning must involve selection, selection implies rejection of more information than is accepted, and this rejection can be called habituation, though this is the reverse of habit-formation. There is a great interest in this field at the present time; at a Colloquium in Moscow some weeks ago this formed the main topic of conversation, though the term used there is "Extinction of the Orienting Reflex".

As you know, there is now ineluctable evidence that habituation is a very active process. This is the question I would like to put to Dr. Ashby. When an animal is subjected to monotonous stimuli there is usually a progressive reduction in the response in the central nervous system. This reduction is due largely to changes at a peripheral level, even in the receptors themselves, yet it seems to depend upon the integrity of rather complex structures in the brain. This should not be confused with Adaptation which is another inherent property of receptors; it is rather a dynamic process of control, of selective blocking from the nervous system outwards. It is susceptible, as Dr. Ashby mentioned, to dishabituation in the presence of another distracting stimulus, and Habituation of this type depends upon integrity of the central nervous system; for example, light anaesthesia may paradoxically augment the brain response to a stimulus which had previously lost its effect by habituation. An anaesthetised animal literally does not perceive the stimulus, yet the electrical response is larger.

This suggests that the structures involved are particularly vulnerable and are therefore probably complex in structure. Does this not imply that habituation in a living organism depends upon a rather elaborate high-level mechanism of selection and control?

Another point is that the degree to which habituation occurs depends on the intensity of the stimulus. If you exhibit a series of auditory stimuli to a human subject, some of which are of moderate intensity and start by being neutral and others are strong enough to evoke an unconditional reflex response, then there may at first be habituation to the moderate neutral stimuli. But if these are associated with the unconditional stimulus so as to become conditioned, the habituation disappears and the response may get larger and larger. The response to the intense unconditional stimulus shows

no habituation however, and in some people may even increase at each repetition.

One can visualise a family of curves plotted on a co-ordinate system in which the ordinate is size or speed of response and the abscissa the number of presentations of a stimulus, the parameter being the intensity of the stimulus. The curves for low intensity stimuli would show a decline with repetition such as Dr. Ashby describes, but those for high intensity stimuli would start off level and then climb away to a limiting value. This sort of relation is seen in the study of nerve fibre excitation where the local potential evoked by subthreshold stimuli declines with time, whereas it rises explosively into an all-or-none propagated impulse in response to supra-liminal stimuli.

Such empirical differences serve in practice to distinguish between neutral stimuli, which show habituation, unconditional ones which show none or even facilitation, and conditioned ones which follow first one rule and then the other. I wonder whether Dr. Ashby has any views about this as a theoretical criterion for identifying the character of stimuli from the point of view of an organism? A classification of this sort would be of great value in experimental studies because it is not at all easy to decide which features in an animal's behaviour are reflex, instinctive, unconditioned, and which are learned, conditioned, exploratory or random.

MR. G. PASK: I have a brief comment to make about this entirely complete paper. It is a request for an extension of the argument to include a more general kind of adaptive behaviour.

Some years ago Eccles *(ref. 1)* suggested that any self oscillatory network of neurone like elements able to interact freely will tend to reject a repeated mode of oscillation in favour of some new mode. If the network also receives stimuli from the external world, and if these sustain a particular oscillatory mode, it is equally the case that this mode will be suppressed, or will be more difficult to sustain, because of the repetition of the stimuli which engender it. The network is thus selective towards novel stimuli and so far as repeated stimuli are concerned it will habituate.

At first sight it seems that all of Dr. Ashby's mathematical arguments are applicable to a structure of this kind and I should be grateful to have a pronouncement on this matter. If such an extension is possible it will provide a calculus for examining not only habituation but the whole process whereby an organism acquires differentiated sensory mechanisms.

REFERENCE

1. ECCLES, J. C. Neurophysiological basis of mind. Waynflete Lecture 1952. *Oxford University Press* (1953).

DR. M. J. BUCKINGHAM: It should be possible to test biological or other systems to see if they behave in a way consistent with that deduced from any particular mechanism for adaptation, such as the interesting one suggested by Dr. Ashby. For example, a system representable by the simple case illustrated in his *figure 1* should show the following property: systems habituated to a stimulus S are subjected to a stimulus of the same type, but with intensity increased just beyond the threshold, until a non-habituated response is elicited. This response should be greater on the average than that shown for the same stimulus by systems which had not been previously habituated. This is a result only for the simple case illustrated, but I would like to ask Dr. Ashby if he has been able to deduce positive predictions of this type, for the more general situation.

DR. W. K. TAYLOR: I believe there is some evidence that the effect of a repeated stimulus can be reduced by an active cancellation process whereby the nervous system learns to cancel the effects of the stimulus. If, for example, the visual scene tends to oscillate, eye movements can compensate for this and the compensation tends to persist after the stimulus is removed. Since many receptors tend to respond to the rate of change of the scene the response will decrease as the compensating mechanism learns to keep the image stationary. This principle also appears to act if we are repeatedly subjected to a force pattern. If, for example, one repeatedly uses the same escalator one gradually builds up an automatic compensation system which reduces the tendency to overbalance on leaving it. Evidence that this is an active cancellation process is afforded by the observation that unbalance tends to occur when one leaves a stationary escalator. It seems to me that there may be a general tendency for living organisms to learn automatically to actively oppose the effects of repeated stimuli.

DR. ROSS ASHBY (in reply): The points that have been raised by Drs. Grey Walter and Taylor are of interest, but hardly concern me, as they refer to mechanisms in which the habituation has been specially developed, usually by natural selection, and I discussed in my paper only a hitherto unnoticed and purely *general* reason for the appearance of habituation. Naturally I do not for a moment suggest that habituation, whenever it appears, is *always* due to this purely general, entropy-like process - I accept freely that sometimes, for reasons of special urgency, natural selection has fostered the development of highly specialised mechanisms that will show habituation with unusual speed and efficacy. The question of what mechanism is at work in any particular case can be answered only by a detailed experimental study of it.

A point I would like to emphasize is that when one uses this type of reasoning about some real "machine" one should realise how free one is to

select *any* meaningful system (as a set of variables) from the infinite number of possibilities that plausibly suggest themselves. A "state", for instance, can legitimately be anything that an observer can recognise with confidence and reliability. It may be one of three typical expressions on an infant's face, or even something like a Fraunhofer line - at which there is actually nothing at all! - no matter, if the observer can recognise it reliably it qualifies as a possible "state". Unless the observer uses freely this freedom of selecting a suitable point of view, he is likely to overlook many possibilities of the theorem's application.

A good example is given by the Eccles' network, referred to by Mr. Pask. We can think of this as changing from state to state so as to *show* the oscillation, or we can ignore these states, think only of the *modes*, and define these as the "states" of a more abstract system, that changes only by going from mode to mode. If the network has plenty of modes at which it can stick, then it will conform to Postulate VII, and will therefore (to answer Mr. Pask's question) certainly tend to show habituation. Thus, though still showing the changes that represent oscillation, it will tend to get to a mode from which the repeated disturbance fails to move it.

Dr. Buckingham's question raises a very interesting extension of my work. Various extensions are being considered at the present time, but I have not yet had time to consider Dr. Buckingham's extension in detail.

It is of interest to notice that these entropy-like processes are all related. Thus, the process of ultrastability that gives coordination and integration is of the same type as this process that gives habituation. In each case the system may go either to a confluent that leaves it immune to disturbance or to one that leaves it still vulnerable. The process is then of the type: Heads, I win; Tails, we toss again - and it tends inevitably to a confluent (or to a set of step-mechanism values) that gives immunity. In general this means habituation; in the particular conditions of the ultrastable system it means coordination and integration.