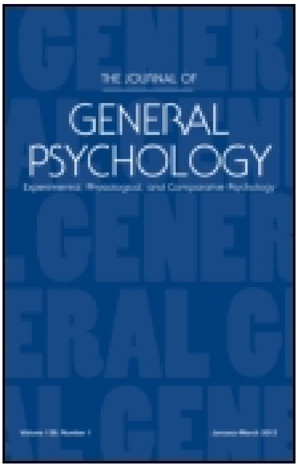

The Journal of General Psychology
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/vgen20

The Physical Origin of Adaptation by Trial and Error

W. R. Ashby [a]
[a] Northampton, England

Version of record first published: 04 Nov 2012.

The Journal of General Psychology, 1945, **32**, 13-25.

THE PHYSICAL ORIGIN OF ADAPTATION BY TRIAL AND ERROR*

Northampton, England

W. R. Ashby

A. Introduction

The type of behavior known as "adaptation by trial and error" is of wide occurrence, being demonstrable with most clarity and simplicity in the lower organisms.

It is the purpose of this paper to suggest that this type of behavior is in no way special to living things, that it is an elementary and fundamental property of all matter, and that it finds its origin and explanation in the concept of a machine "breaking." It is demonstrated, that if any environment starts to break down any machine, then that machine is bound to show the essential features of adaptation by trial and error, if by "adaptation" we understand "reaching equilibrium."

The proof proceeds in three stages: First it is shown that "adapted" from the physiological point of view is equivalent to "in equilibrium" from the physical point of view, at least in the simpler cases. Secondly, it is shown that all dynamic systems, whether living or dead, keep breaking until they reach some equilibrium, reaching it by a process identical with that of trial and error. (So far no special hypothesis is required). Thirdly, if we add the hypothesis that the breaks are available in large numbers and are consistent, then it is shown that the machine is bound automatically to reach an equilibrium within certain essential limits.

A difficulty of much psychological theorizing is vagueness in the terms employed. In this work, the above ideas have been studied in mathematical form throughout, the definitions and proofs being given corresponding precision.

B. Adaptation by Trial and Error

We must first be agreed on the essential features of this mode of behavior. They seem to be as follows (listed here for convenience): (1) It is called forth only when the environment becomes in some way unsatisfactory. (2)

*Received in the Editorial Office on June 12, 1943.

Each system of behaving, i.e., each "trial," is persisted in for some finite time and is changed only if and when it is proved wrong. (3) The persistence of an unsatisfactory environment calls forth an almost endless succession of trials. (4) The next trial is selected at random and is not necessarily any improvement on the previous one—it is only different. (5) If, and when, a trial results in "success" it is persisted in and does not alter.

But so far there is no objective definition of "success." To study this further, I quote two typical authorities. Jennings (1906) writes:

> Organisms do those things that advance their welfare. If the environment changes, the organism changes to meet the new conditions. . . . If the mammal is cooled from without, it heats from within, maintaining the temperature that is to its advantage. . . . In innumerable details it does those things that are good for it.

Many other writers have emphasized the same theme: that every animal has a certain number of essential variables and that the behavior must be such, in relation to the environment, as will keep these variables within certain (i.e., "physiological') limits. But this is the definition in physics of an equilibrium (e.g., Courant, 1936). Haldane (1922) sums up with: "Biology must take as its fundamental working hypothesis the assumption that the organic identity of a living organism actively maintains itself in the midst of changing external circumstances." It may, therefore, reasonably be assumed (Ashby, 1940) that "adapted" behavior is equivalent to "the behavior of a system in equilibrium."

Our problem, then, is to obtain trials which shall ensure the arrival of the animal at an equilibrium within physiological limits.

C. Machines and Organization

We start with a "dynamic system" or "machine." The essential character of this is that it is a collection of parts which (*a*) alter in time, and (*b*) which interact on one another in some determinate and known manner. Given its state at any one moment it is assumed that we know or can calculate what its state will be an instant later. Consideration seems to show that this is the most general possible description of a "machine." The parts will be referred to as "variables" since in all cases they must be capable of being specified by measurement. It is assumed throughout the paper that the system is complete, i.e., that all the interacting parts are included. (It is not in any way restricted to mechanical systems with Newtonian dynamics.)

A typical and clear-cut example of a dynamic system is given by a frame with a number of heavy beads in it, the beads being joined together by elastic strands to form an irregular network. Such a network, if pressed out of position and then released, will oscillate with a pattern of movement which depends on (*a*) the pattern of the original displacement and (*b*) on the arrangement of the elastic strands, their tensions, and the masses of the beads. The network is a dynamic system by the definition since the positions and velocities of the beads (the "variables") alter with time, and since the movement of any one bead at a given instant depends on where the other beads are at that instant. It is "complete" if we include all the beads. In this illustration the variables are all spatial (or velocities) but this is not in any way necessary: temperatures, electric potentials, concentrations of solutes, etc., may be included if necessary.

The variables are those items which alter with time. But behind these are the constants of the network: the masses of the beads, the lengths of the strands, their arrangement, etc. These constants are the "organization" by definition. Any change of them would mean, really, a different network, and a change of organization.

D. Equilibrium

The subject is commonly mentioned as if it needed no discussion. The usual elementary definitions (e.g., "that the vector sum of the forces are zero") are of no use here, for we are studying much more general systems. A suitable exact definition is given later, but here we may use that of Lorentz (1927) as sufficient: "By a state of equilibrium of a system we mean a state in which it can persist permanently."

But this definition makes possible all sorts of curious special cases, physically rare but mathematically possible. From all these we separate off a restricted type characterized by simplicity and common occurrence to which is applied the term "normal" equilibrium. It occurs when a dynamic system has a configuration *C* with the properties that (*a*) if started at *C* and released, it does not move from *C*, and (*b*) if started at any configuration near *C*, the system changes in time towards *C*.

Normal equilibrium has some special properties which we must notice. Firstly, the system tends to the configuration *C;* so if it is disturbed slightly from *C*, it will *automatically* develop internal actions or tendencies bringing it back to *C*. In other words, *it opposes any disturbance from C*. Further, if we disturb it in various ways, it will develop different tendencies with different disturbances, the tendencies being always adjusted to the disturbances so as to oppose them. It should be noted that this is characteristic of any system in normal equilibrium, whether living or dead.

Finally, it must be noted that an equilibrium configuration is a property of the organization and has nothing to do directly with the variables. No matter how the latter may vary the equilibrium configuration is quite unaltered. But a change of organization will change the equilibrium configuration at once. *The equilibrium states of a machine are defined by the organization only.*

E. Breaks

Let us imagine a machine has "broken." The first observation is that no matter how chaotic the result, it is, by our definition, still a machine. But it is a different machine. *A break is a change of organization.* A simple and clear example is given by the elastic network. If one of the strands breaks we obtain the same machine as if that strand had been omitted from the beginning. In the first machine the elastic strand had a pull per cm. stretch of, say, *k* dynes, while in the second machine, or in the first after breaking, the pull is *O* dynes per cm. A break, therefore, occurs in a machine when some constant of the organization changes suddenly from one value to another (constant) value. (Other examples are that in the break of a cup the cohesive force drops to *O,* while in the fusing of an electric wire the resistance jumps to infinity).

In all these cases the break has occurred because something has "gone too far." Thus the elastic strand between beads will break if the distance between the two beads exceeds some given distance; the cup breaks if the force exceeds a certain amount; and the wire fuses if the current exceeds some fixed value. We can see clearly in the elastic network example that this means that every break has associated with it a set of configurations of the machine such that if the machine should happen, in its changes, to reach one of these configurations, then the break will occur. Every break must have this set of configurations specified.

Not every change of a variable can be considered as a "break." It is *defined* here that a machine may be said to "break" if and only if some variable which was at one *constant* level, changed to some other constant level due to the machine assuming one of the given set of configurations.

F. Machines and Time

What happens to machines, as defined above, in time? The first point is that, in practice, they all arrive sooner or later at some equilibrium (in the general sense defined above). Thus, suppose we start with a great number of haphazardly assembled machines which are given random configurations and then started. Those which are tending towards equilibrium states will arrive at them *and will then stop there*. But what of the others, some of whose variables are increasing indefinitely? In practice the result is almost invariable—something breaks. Thus, quicker movements in a machine lead in the end to mechanical breaks; increasing electric currents or potentials lead inevitably to the fusing of wires or the break-down of insulation; increasing pressures lead to bursts; increasing temperatures lead to structures melting; even in chemical dynamics, increasing concentrations sooner or later meet saturation. The sole case I have been able to discover where a variable can actually increase indefinitely is that of a comet, which can wander indefinitely far without altering the equations which specify its motion. But this rule, that machines break if their variables increase too far, certainly seems to hold for all terrestrial machines.

After a break, the organization is changed, and therefore so are the equilibria. This gives the machine fresh chances of moving to some new equilibrium or, if not, of breaking again. So the machine goes on: either moving to an equilibrium, or moving and, sooner or later, breaking again. (If it goes on moving but always stays within a certain region, it still counts as being in equilibrium.) It is as if a gambling game were being played with the curious rule of "Heads—I win; Tails—we toss again!" Clearly there is only one end. We may state this principle in the form: dynamic systems stop breaking when, and only when, they reach a state of equilibrium. And since a "break" is a change of organization, the principle may be restated in the more important form: *all dynamic systems change their internal organizations spontaneously until they arrive at some state of equilibrium.*

It is important to appreciate that up to this point we have made no special hypotheses at all. The properties described are necessary to all machines whatever.

We now reach the point of the whole paper. We ask what will happen if (*a*) there are large numbers of breaks possible and (*b*) if the configurations of the various breaks are closely similar to one another and have some common feature. Suppose such a machine started. If it reaches an equilibrium not passing any of the special configurations, nothing breaks, and it settles at the equilibrium. But if it should move towards, and reach one of the configurations, a break at once occurs, the machine changes its organization and starts a new line of behavior. If this takes it away from the special configurations to an equilibrium, it will go there and stop there. But if it continues to move towards the special configurations another break will occur leading to another line of behavior. Clearly, if the number of breaks is sufficient the machine is bound, sooner or later, to arrive at an equilibrium which is in some configuration *away from the special configurations*. Thus the special configurations control, to some extent, the position of the final equilibrium configuration. The elastic network will provide an example. Suppose it to be connected up so that many of the strands are much over-stretched and that all strands are such that they break if they exceed 10 cm. in length. We let the machine go. The beads may oscillate wildly and various strands may or may not break but at any rate we can predict with certainty that the final organization will be such that there will be an equilibrium configuration which has not a single strand exceeding 10 cm. in length. The machine finds this organization automatically if it is allowed to break freely. In this form the theorem may seem a little obvious but its importance is that it is still true in cases where it is impossible to see it intuitively.

The precise statement of this theorem is given later. Here we may sum up by stating that: a machine which has available an indefinitely large number of breaks depending on configurations closely similar to one another will inevitably change its internal organization spontaneously until it arrives at an organization which has an equilibrium with the special property that it avoids those configurations.

G. Application to Living Organisms

We now apply the result of the previous section to living organisms, remembering that our problem is "to obtain trials which shall ensure the arrival of the animal at an equilibrium within physiological limits."

First we must assume that we have all essential information about a given animal and its internal mechanisms at a given moment—a large hypothesis but one hampered only by technical difficulties. (This, however, applies only to the fixed elements of its construction and nature; the changeable organization, i.e., the "breaks" are not included here.)

The next point is that the animal by itself is incomplete. The environment is an essential part of the dynamic system. When we have both animal and environment properly specified and all knowledge of how they interact, our system is complete. (We now make no essential distinction between "animal" and "environment": both contribute to the organization of the whole, both act on themselves and on each other, any equilibrium must stabilize both, and as physical systems the same principles must govern both.)

The sole special hypothesis required is that we must assume that the animal is provided with a sufficiency of breaks, i.e., variables with the following properties: (*a*) They must remain constant at one value and change to some other constant value on certain contingencies. (*b*) Such change must affect the animal's behavior. (*c*) Such change must depend upon, or be associated with, some threat to the animal's existence. Thus, decreasing oxygen supply is to cause manifold changes of the type described). The association must be provided by heredity: here we assume it done. It specifies which configurations are to be avoided in the final equilibrium, thereby ensuring that the final equilibrium shall be within "physiological limits" for that animal.

We hypothesize, therefore, simply that if the animal should leave a state of being within "physiological limits," then something is to break. But what the break is, or how it affects the animal is of no importance at all. It need not be in any sense "adaptive" or specially chosen.

The application of the theorem to living organisms is now direct. The essential features of adaptation by trial and error are now paralleled in detail by those of the dynamic system as described above. Suppose we set up and observe such a system. If it is started when not in equilibrium the many variables will start to change in accordance with the internal organization of the whole machine. As long as the variables do not get near the special configurations mentioned so will the behavior continue. All this counts as one "trial." If the system arrives at an equilibrium it will stay there. Here it will oppose any disturbance, always returning to the central state. The system, in other words, will appear to be "adapted." But if it does not arrive at an equilibrium it must tend, sooner or later, to cause a break. The organization is thereby changed and the behavior consequently changes, the organism appearing to be making a new trial. If this new organization should lead to an equilibrium avoiding the special configurations ("within physiological limits") then it stays there. If not, another break will occur later, and so on. The machine, therefore, keeps changing its internal organization until it arrives at an equilibrium avoiding the special configurations. No other ending is possible to it.

The only other point to mention at present is that the development of a nervous system will provide vastly greater opportunities both for the number of breaks available and also for complexity and variety of organization. Here I would merely emphasize that the difference, from this point of view, is solely one of degree and not of principle.

H. Dynamic Systems

To describe a dynamic system we first label all necessary variables x_1, x_2, . . . , x_n. A configuration of the system (at a given moment) will be the set of numerical values of the variables. The behavior of the system will be defined by the sequence of configurations. If the properties hold, the behavior must be defined by equations of type

$$\left.\begin{aligned} \frac{dx_1}{dt} &= f_1(x_1, x_2, \ldots, x_n) \\ \frac{dx_2}{dt} &= f_2(x_1, x_2, \ldots, x_n) \\ \frac{dx_n}{dt} &= f_n(x_1, x_2, \ldots, x_n) \end{aligned}\right\}$$

(This may be written more briefly as

$$\frac{dx_i}{dt} = f_i(x_1, x_2, \ldots, x_n) \quad (i = 1, 2, \ldots, n)$$

or even as

$$\left.\frac{dx_i}{dt} = f_i(x)\right).$$

After much testing I am satisfied that this method of specifying a dynamic system is the best for the present purpose. It describes what the system *does,* not what it is. The "organization" is defined by the f's. (The latter must be single-valued but no restriction of continuity or differentiability is implied.) The final definition of a dynamic system as used in this paper is that it refers to any set of numerical variables whose alterations in time are capable of being described by such a set of equations.

It is required that the time shall not appear explicitly in the f's; for if it did, it would mean that the system was subject to some arbitrary external interference. Such systems are not considered here.

This specification of a given dynamic system may be converted to an equivalent geometric form. (This is essential for later proofs. The n

variables x_i may be treated as coördinates in an n-dimensional space. The point $x_1, x_2, \ldots, x_n$ specifies one configuration, or instantaneous state of the machine. It is called the "representative point" of the machine. As is well-known, such a set of equations defines a field, filled with paths (except perhaps in some regions which are not physically possible). If the representative point is started on a path it must follow it.

With regard to equilibrium, it seems clear that the essential idea is that a path ends in equilibrium if from some configuration onwards the values of the variables are always bounded. Anything less general is demonstrably unsatisfactory. "Normal" equilibrium occurs when a simply-connected region of the x-space is bounded by a surface such that all paths which traverse it do so from without inwards, and all terminate finally at one, and only one, point.

So far we have dealt with the features of one given organization. A change of organization involves a change of equations from, say,

$$\frac{dx_i}{dt} = f_i(x) \text{ to } \frac{dx_i}{dt} = \phi_i(x).$$

A more convenient method is to use parameters. If

$$\frac{dx_i}{dt} = F_i(x_1, x_2, \ldots \; x_n; p_1, p_2, \ldots, p_r) \quad (i = 1, 2, \ldots \; n)$$

then a change of the p's will change the F's as functions of the x's, and this change, by definition, is a change of organization.

I. Step-Functions and Breaks

A "step-function" is defined as a variable which can take only certain discrete numerical values. Excluding wholly discontinuous types we get a graph of the form shown in Figure 1. (It is often convenient to consider instead a continuous function like *tanh* qx which can approach the step-function form as closely as we please while remaining continuous).

A step function of more than one variable is defined similarly except that it depends on whether a given function $v(x_1, x_2, \ldots, x_n)$ is greater or less than a given quantity (which can be made O without loss of generality).

We can now give an exact mathematical form to the concept of a "break." Consider a system $dx_i/dt = f_i(x)$ $(i = 1, 2, \ldots, n)$, starting from a given point $x_1^0, x_2^0, \ldots, x_n^0$. Suppose one of the variables, x_n, say, is a step-function of the time, and that it changes from x_n^0 to x_n^1. When x_n is

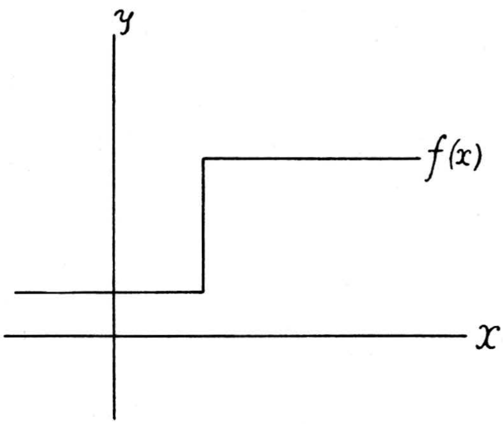

FIGURE 1

constant at either value, we may, so long as this holds, remove x_n as a variable, and thus simplify the functional forms f, getting, say,

$$\frac{dx_i}{dt} = g_i(x_1, x_2, \ldots, x_{n-1}) \quad (i = 1, \ldots, n - 1).$$

Similarly, while it is at the other value, we may simplify getting

$$\frac{dx_i}{dt} = h_i(n_1, x_2, \ldots, x_{n-1}) \quad (i = 1, \ldots, n - 1).$$

A system containing a step-function of the time may therefore be regarded from two points of view: (a) we may think of it as one machine $dx_i/dt = f_i(x)$ which continues indefinitely, is composed of n variables, and undergoes no change of organization, the step-function x_n being considered as an ordinary function; or (b) we may think of it as starting as a machine $dx_i/dt = g_i(x)$, composed of $n - 1$ variables, which at some time changes suddenly and forms a new machine $dx_i/dt = h_i(x)$ in the same $n - 1$ variables, there being a change of organization. In this case the first machine is said to "break." Although the difference between (a) and (b) above is slight with only one step-function, it must be noted that it leads to great differences if the number of step-functions is large. In that case "the system" has a very different meaning in (a) and (b).

The final definition of a "break" can now be given. The occurrence of a "break" is identical with a step-function changing value. On the other hand, it is optional, at the change of value of a step-function, whether this is regarded as a break or not.

J. Break-Surfaces

Every step-function, by definition, depends on whether a given function v of the arguments is positive or negative. If v is a continuous function of the x's, the surface

$$v(x_1, x_2, \ldots, x_n) = O$$

(keeping the notation uniform with that of the last section) will divide up the x-space into regions so that on one side of the v-surface the step-function has one value and on the other side of the surface another. This surface is called a "break-surface."

If x_n is the step-function, with two values, $x_n{}^0$ and $x_n{}^1$, say, then in n — 1-dimensional space there will be two break-surfaces associated with each step-function namely

$$v(x_1, x_2, \ldots, x_{n-1}, x_n{}^0) = O$$

and

$$v(x_1, x_2, \ldots, x_{n-1}, x_n{}^1) = O$$

In n-dimensional space there is only one surface; in n — 1-dimensional space there are two. The latter case is the more important here, for the concept of a "break" belongs only to the machine composed of n — 1 variables.

The special property of a break-surface is that it specifies the conditions under which x_n will break, i.e., change from one value to the other. Thus, suppose the machine starts with x_n at $x_n{}^0$, and suppose that the configuration of the machine (in n — 1 variables) is such that the representative point is on the appropriate side of the break-surface

$$v(x_1, x_2, \ldots, x_{n-1}, x_n{}^0) = O.$$

Then as long as the representative point does not cross that surface, so long will x_n remain at $x_n{}^0$, i.e., "unbroken." Should it cross, x_n will at once break to $x_n{}^1$. (What happens after that depends on the position of the other break-surface; if, as often happens, it is entirely at infinity, then the break is irreversible.)

The last theorem of Section F is specially interesting when the many configurations causing breaks are represented by many break-surfaces almost, but not quite identical, and which surround a region. If the representative point touches a break-surface, the field changes and the new path from that point of contact may, in general, run in any direction. If the new path goes away from the layer of break-surfaces another break will occur and a fresh path formed. Clearly, enough break-surfaces will ensure that sooner or later the point will be turned back from the layer. *A layer of break-surfaces acts as a barrier to the representative point.* If the point at every break is equally likely to move in any direction, then the chance of it getting through a wall of m break-surfaces is of the order of 2^{-m}. In this way we state the theorem much more precisely. In particular, if we start within some assigned region (corresponding to "physiological limits") and if there are great numbers of breaks available if the point tries to leave this region, then this will ensure that the machine will find an organization with an equilibrium within the given region. We have thus solved the problem as it was stated.

K. Summary

An outstanding property of the nervous system is that it is self-organizing, i.e., in contact with a new environment the nervous system tends to develop that internal organization which leads to behavior adapted to that environment.

In an examination of the principles underlying this phenomenon the author was led to a mathematical study of organization in dynamic systems, and in particular, to an examination of what happens when a machine "breaks." It was found that a "break," far from being haphazard, has highly characteristic properties, some of which are described here. It is found, in particular, to be closely connected with the subject of equilibrium. It is demonstrated that *a simple breaking-down of a completed machine is sufficient, by itself, to produce the objective features of adaptation by trial and error,* at least in the simpler cases; i.e., every machine breaking-down automatically readjusts its own internal organization until adaptation is achieved. This type of adaptation (by trial and error) is therefore an essential property of matter, and no "vital" or "selective" hypothesis is required.

References

1. Ashby, W. R. Adaptiveness and equilibrium. *J. Ment. Sci.,* 1940, **86**, 478.
2. Courant, R. Differential and Integral Calculus. London, 1936.

3. HALDANE, J. S. Respiration. New Haven: Yale Univ. Press, 1922.
4. JENNINGS, H. S. Behavior of the Lower Organisms. New York: Columbia Univ. Press, 1915.
5. LORENTZ, H. A. Theoretical Physics. London, 1927.

Green Ridges
Church Way
Weston Favell
Northampton, England