Göteborg Psychological Reports

University of Göteborg Sweden

HELGE MALMGREN

On the nature of reinforcement

Number 3 Volume 15 1985

ON THE NATURE OF REINFORCEMENT

Helge Malmgren

Malmgren, H. On the nature of reinforcement. Göteborg Psychological Reports, 1985, 15, No. 3. - An abstract analysis of reinforcement for randomly composed finite deterministic automata is proposed. Negative reinforcement is explained as being a result of a temporary rise in the diversity of the automaton's input. Such a rise tends to de-stabilize the automaton, and successive such de-stabilizations may amount to a kinesis in the internal state space of the automaton. It is shown that randomly composed automata of a certain kind (inert automata with randomly assigned state-determined output) will with a high probability learn to perform correctly if all outputs except those belonging to a certain subset are negatively reinforced in this way. A realization of the abstract model in terms of a hierarchical, arousable nervous system is then proposed, and a computer simulation of a simplified such realization is presented.

Key words: Automata, learning, operant conditioning, reinforcement, stochastic processes.

In two previous papers (Malmgren 1980, 1984) I showed that phenomena analogous to habituation and simple forms of classical conditioning will tend to occur more often than not in randomly composed finite deterministic automata. I also argued that these probabilities can be turned into reliable performance by means of "mass action", if the inputs and outputs of a large collection of such randomly composed systems are organized in certain ways. In this essay I instead address the question whether operant conditioning can be given a similar kind of explanation. By a "similar kind of explanation" I here refer to an explanation with the following properties:

1) It is framed in purely non-intentional (non-cognitivistic) terms. This disqualifies a number of candidates which essentially refer to states of belief, expectations etc.

2) It is formulated in abstract causal terms, i.e. without reference to this or that realization of the system in question. This excludes explanations in terms of, e.g., special synaptic mechanisms.

3) Its aim is to show the very possibility of learning under certain extremely simple assumptions - not to predict actual parameters of learning phenomena.

Of course, this explanatory aim is not incompatible with a belief on my behalf that we also need intentionalistic explanations (at least in certain cases), concrete physiological explanations and detailed, parametric models in order to completely understand the phenomena of learning. However, explanations of the kind explored here may possibly

a) facilitate the reduction of intentional to non-intentional discourse;

b) explain why learning phenomena are ubiquitous in the living world, and not confined to certain kinds of organisms;
c) exhibit the common structure of several more concrete and more detailed explanations.

Reinforcement in classical and operant conditioning

When we say that a certain process involves the "reinforcement" of a response R (or a central state C), we mean that it alters the probability that R (or C) will occur. The experimental paradigms of classical and operant conditioning seem to demonstrate that at least three kinds of reinforcing processes are at work in associative learning. In classical conditioning, the contingency between CS and UCS positively reinforces (strengthens) responses to CS which are similar to the organism's fixed response to UCS. In the "operant" paradigm, the occurrence of a certain kind of event, a so-called ("positive" or "negative") "reinforcer" is made contingent on the occurrence of a selected response R on behalf of the organism. This arrangement often strengthens (or, in the case of a "negative" reinforcer, weakens) R even if R is not similar to the fixed or "unconditional" response to the reinforcer. In this latter way one can, e.g., learn a rat to respond with "approach behaviour" in order to avoid a shock signalled by a discriminative stimulus D, although in this case the simultaneous classical conditioning of "escape" reactions to D tend to interfere heavily with the learning of the adequate response. (For details, the reader is referred to Mackintosh 1983).

In Malmgren 1984, I argued that the reinforcement involved in a simplified form of classical conditioning may have a very straightforward explanation in terms of certain stability properties of finite deterministic automata under restricted input. In brief, I define a randomly composed automaton as an automaton the transition matrix of which is construed by randomly assigning states to places in the matrix. If a mixture of such randomly composed automata is equipped with a certain "unconditioned response" R to UCS and given UCS at irregular intervals, the consecutive responses to the input preceeding UCS will more and more tend to be identical to R.

In order now to give a possible abstractly causal explanation of operant reinforcement, I first want to propose a certain automaton-theoretic analysis of the working of negative reinforcers. The idea is simply this: since the finite automaton tends to behave in a more stable way under a restricted input than if given a more diversified input, any device which raises the diversity of the input will de-stabilize the automaton. This, in turn, means that the automaton's general tendency to behave in the same way in the future as it did before will be smaller than if the "diversifying" device had not been operating. One reason for choosing this interpretation of the negatively reinforcing event is the physiological fact that there is, at least in the higher animals, a device which works in this de-stabilizing way: namely, the arousal system, which certainly is very much at work in shock reactions, and which tends to raise the information in-

flow into the CNS - i.e., to diversify its input.

I will not attempt a detailed analysis of positive operant reinforcement; suffices it to say that in my opinion it must be given a quite similar treatment in terms of lowering of the information inflow (restricting input). Concerning the relations between classical and operant reinforcement, see the last paragraph.

Operant conditioning as a kinesis in abstract state space

It is well known that many lower organisms succeed in finding food, shelter etc by means of so-called "kinesis", i.e. repeated, pseudo-randomly directed movements, the frequency and/or size of which are controlled by some parameter of the environent. Abstractly, the mechanism by which the appropriate direction of movement is found by means of pseudo-random changes of direction, which cease when a "good" direction is found, is very similar to kinesis proper, and it is sometimes classified as a kind of kinesis ("klinokinesis", in contradistinction to the simple "orthokinesis"; for terminology and examples see Carlile 1975). In any case, both are good examples of "selection by consequences" (Skinner 1981), and it is useful to compare them with operant conditioning.

Imagine a kinesis which takes place in the inner state space of the organism. Suppose that there is some - as yet unexplained - pseudo-randomizing device RND, such that RND is switched on every time the organism is in a state belonging to a region R of its state space. It will then go to any state with the same probability in the next moment. It is easily seen that something like kinesis away from R will occur if an only if the organism´s tendency to stay in R in the absence of RND is greater than that expected from a random selection of states.

Now, it is a basic property of finite deterministic systems that they tend to restrict their behaviour to relatively small sets of states; consequently they tend to return to "same state" with a probability exceeding that which is given by the RND device. (For some illustrations of this principle, see Malmgren 1980). Therefore, a randomly choosen such system with the RND coupled to a region R will probably tend to leave R more often than will the same system without RND. By the same token, if once "randomized out" of R it will also tend to be back in R - after a given period of time - with less probability than if left in R. As a whole, it will therefore tend to avoid R.

The situation is, however, different from ordinary kinesis in that external space has a certain topology and metric which influence the character of adaptation by kinetic movements in space. For example, it is certainly easier for organisms to learn to stay in a spatially connected "good" region than in a disconnected one. In the randomly composed automaton there is no preassigned "nearness" or "connectedness" relations between states, and different automata may have very different topologies. Whether there is, for example, two different regions which are not accessible from each other under a given input depends on the result of the random construction of the relevant transition

function.

Let us then investigate to what degree kinesis in state space will actually be similar to kinesis in external space.

An illustrative special case is given by the random system (in the sense of Malmgren 1984) with n states and a constant input, trying to avoid k of its states (the "R-states") by means of the RND device. If and only if an automaton belonging to this system has a possible basin (equilibrium set of states) not containing any R-state, it will with certainty eventually learn to stably avoid them. This is so because:

1) in such an automaton, if it is in a basin not containing any R-states it will stay there, while if in a basin including some R-state it is certain that it will sooner or later leave it; also, in a finite number of steps it must reach a "good" basin; and

2) in the other automata, there can only be basins containing R-states, and the automaton will forever keep returning to such states with irregular intervals.

Now the probability that there is in the n-state automaton a basin not containing any of k pre-selected states is = (n-k)/n.

We prove the equivalent proposition: The probability $P_{k,n}$ that a set of k states in an n-state automaton contains some basin is = k/n. This is true for k=1, since the probability $P_{1,n}$ that any state is a basin is = 1/n. Suppose that it is true for k = s. Take any set of states with s elements and add one element a. Then the probability $P_{s+1,n}$ that this new set will contain some basin is = s/n + 1/n - (s/n)(1/n) + Q, where Q is the probability that neither the s-set nor the 1-set contains a basin but that one is formed by the union of them. Note that the probabilities of basins in two disjoint sets are independent, which explains the third term. To evaluate Q, note that there are s!/t!(s-t)! ways of choosing the set with t elements from s which combines with a to make a basin, and that this basin can be formed in t! different ways, each with the probability $(1/n)^{t+1}$. In order to countenance the possibility that there is no other basin in the s-set, we also have to multiply with $1 - P_{s-t}$, which is = 1 - (s-t)/n. Q must then be the sum from t = 1 to t = s of $Q_t = s!/(s-t)! \times (1/n)^{t+1} \times (1 - (s-t)/n)$. Q_t can also be written as $(s/n^2) \times (s-1)!/(s-t)! \times (1/n)^{t-1} \times (1 - (s-t)/n) = (s/n^2) \times Q'_t$. Each term in Q'_t contains two parts, stemming from the two terms in the last parenthesis. It can be seen that the second part of each term is cancelled by the first part of the next term; also, the second part of the last term is = 0. Hence, what remains is the first part of the first term, which is = 1. Therefore the value of Q is = s/n^2 which, if substituted in the expression for $P_{s+1,n}$ above completes the proof.

This result can be generalized in the following way. Suppose that to each state in a randomly constructed automaton there is randomly assigned one of a number of outputs, and that the output o is assigned with probability q. Then the probabi-

lity P'_q that the set of states with output o contains some basin is = q.

The proof is simple. P'_q is arrived at by summing over all alternatives of the form: exactly s states have output o, using the previous result. There are n!/s!(n-s)! ways of selecting s states, the probability that they but not any other state have output o is $q^s(1-q)^{n-s}$, and the probability that the set contains a basin is = s/n. From this it can be seen that the expression for P'_q will be exactly the expression for the expectancy of the binomial distribution, which is = q.

The conclusion to be drawn from these calculations is that the random automaton will not learn very efficiently according to the model presented above. If the "built-in" probability of R, defined as a state-space region of a given size or as a randomly assigned output, is = q, then with probability 1-q the random automaton will learn to reliably avoid R if punished in R; equivalently, with probability q it will learn to perform R if "punished" outside R. This means that only very "easy" tasks like for example avoiding one of n states will be learnt with anything approaching certainty. (However, the random system certainly tends to perform better than if the RND device had not been operating. Without it, the randomly composed automaton tends to give the correct response with probability q and with irregular intervals; with learning, a proportion q of automata will certainly learn always to perform the correct response, and among the other ones many will still give the correct response from time to time.)

At this stage, let us note that the character of operant conditioning makes it difficult to exploit the "mass action" of a collection of automata in the way it was done in Malmgren 1984 for primitive classical conditioning, where I hypothesized a mechanism with the function of selecting the most common response of the automata as the response of the organism. If such a mechanism (for example, a pattern of competitive convergence on effector pathways) produces the output of a collection of automata, which is exposed to a negative operant contingency, then all automata in the collection will be punished for the sins of the majority, and the minority will not be punished when the majority perform well. This situation does not seem to be conducive to efficient operant learning. Hence, it is probably advisable to try to find other theoretical ways of improving performance.

The inert random system

This creature differs from the ordinary random automaton in the following way: when, for any state a and input b another state a' is randomly choosen as the state to which the automaton goes from a under b, the probability that a = a' is greater than 1/n; more precisely, it is = i (for "inertia") which is fixed in advance for the whole construction. In other respects, the system

is still symmetrical; hence, for each a'' not = a, the probability that a' = a'' is = (1-i)/(n-1). This latter fact means that the inert automaton is not biased to learn to prefer any pre-selected state - although it is certainly biased to prefer not to change its state.

We now assign outputs randomly and independently to the states of the inert automaton and ask: what is the probability that an RND device will make it learn to perform a certain response which has an "inbuilt" probability (i.e., a probability of prior assignment to each state) = q? This question is not quite as easy to answer as the previous ones. It is however clear that for fixed n and q, the probability is a positively monotonous function of i: a rise in i means a greater probability for basins of size 1, and the probability that all states in a basin of size s or the state in a basin of size 1 have (has) the "correct" output is larger than the probability that all states in a basin of size s+1 have the desired output. By the same token, the probability is maximal for i = 1; then it is simply = 1 - $(1-q)^n$. (Rather interestingly, for q = 1/n the latter expression grows to a limit = 1/e, which means that even the inert automaton does not learn to "approach" one state very efficiently!) - On the other hand, for each i a lower bound for the probability of some basin with the desired property is given by the probability of some such basin with size 1, which is = 1 - $(1-iq)^n$. For fixed i and q, this expression evidently tends towards 1 as a limit when n grows. For modestly large i and n, the lower bound will be a good approximation of the sought-for probability, for the obvious reasons that many basins in such automata will be 1-basins and that such basins have the best chances of being well-performing. As an example (cf. below), with n = 16, q = 0.25 and i = 0.5625 the lower and upper bounds are 0.9115 and 0.9900 respectively.

I have not been able to derive any simple formula for the exact probabilities. Instead, I have used a computer algorithm in which for each possible partition of the state space the probability is calculated that all elements of the partition except one are basins, at least one of which contains only states with the desired output, and that the remaining element of the partition does not contain any basin. All such probabilities are then summated. For n = 16, q = 0.25 and i = 0.5625 the total probability turns out to be = 0.9123.

From the general idea in the argument, and from the numerical example given, it can be seen that an inert automaton has a considerable chance to learn by the RND device to perform steadily a response which has a much smaller "inbuilt" probability - in contrast to what is the case with the automaton without inertia, which "only" learns with the same probability as that of the prior assignment.

The hierarchical model with an arousal system

The present model, considered as an explanation of operant conditioning by negative reinforcement, has two serious drawbacks. First, it makes use of a mysterious "RND" device which has cer-

tainly not been explicated in abstract causal terms; second, it only works under the presupposition that the automaton is given a constant input sequence - what we want to explain is, however, how animals can "store" a correct response R to a certain input b although they certainly experience other, "disturbing" inputs during the intervals between the occurrences of b.

Both these remaining problems have been solved by nature by means of the hierarchically organized, arousable nervous systems of higher animals. These systems have, among other features, the following properties:

1) The "lower" parts are relatively stimulus-bound, i.e. they react in a rather stereotyped way and do not learn much. They are also more reactive to external stimuli than the "higher" parts.

2) In comparison to the lower parts, the higher ones are, as a rule, shielded from external stimuli. On the other hand, it seems that their reactions to external stimuli are more flexible.

3) The output of the organism is a result of convergence between impulses from lower and from higher parts. In view of 1) and 2) this entails that differences between responses on similar stimuli are mostly due to a changed state of the higher systems. (This is often expressed, somewhat misleadingly, as the higher centers "inhibiting" the action of the lower ones.)

4) The information flow "upwards" (to higher centers) is sometimes dramatically augmented, especially when the organism faces new and/or dangereous stimuli. This "activation" is at least partly accomplished by structurally identifiable systems.

Let us now arrange some peripherals for an inert random automaton which will make it similar in essence to the hierarchical nervous system.

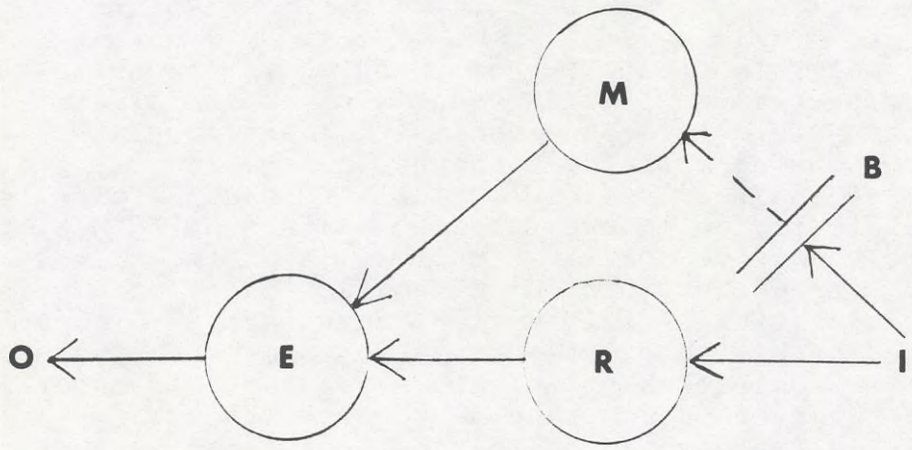

Figure 1. The hierarchical arousable automaton.

Here, M (for Memory) is an inert automaton with input. However, as a rule it is shielded from the environmental input I by the barrier B; hence, in effect it is usually an automaton with constant input. When the barrier is lifted, M realizes its po-

tentialities for one single moment and works like an automaton with all environmental stimuli as inputs. The system R (for Reflex) is not a random automaton, but a random transducer of environmental information. (Like M, it has evolved, phylogenetically, from random automata, but that is another story.) For simplicity, put state of R = input. The system E (Effector) is another specialized system, namely a random integrator of information from M and R: its output is a random function from the two states of M and R to an output set. The output of E is the output of the whole organism.

Now, arrange a contingency which consists in the information barrier B being temporarily lifted whenever the whole organism is given a certain stimulus b and does not produce a certain output o. During periods when the barrier is not lifted, M behaves in a rather stable way because it has, in effect, a constant input. Hence, there is an appreciable probability - cf. above - that it will contain at least some basin, the outputs of which together with the state of R produce o as the "total" response on b. If M is in such a basin, the organism will always perform correctly, and the barrier will not be lifted even when b occurs. If it does not perform correctly - and hence is in some state outside all "well-performing" basins - the information barrier will be lifted, and M will for a moment behave as an automaton with diversified input. Depending on the degree to which the environment varies in its perceptible details between "trials", and on the degree i of inertness, this lifting of the barrier will approximate more or less to the RND device outlined above. Note that it will work more slowly for higher values of i, and that it will not work at all for i = 1! Presupposed that i < 1, M may well leave its "bad-performing" basin and enter a "good-performing" one. After this quasi-random step, M will again be rather stable until next "trial"; and so on until full stability - i.e. a well-performing basin - is (probably) reached.

The really important thing to notice is of course the fact, that the hierarchical arrangement with an information barrier makes it possible for the organism to learn a relatively stable response to one stimulus although it is also exposed to the other ones and reacts differentially to them. This is simply not feasible if the output reflects an internal state which, in turn, is sensitive to all stimuli all the time. This, in its turn, is but another wording of the fact that remembering, which means transmission of information about the past, always implies diminished transmission of information about the present.

As an illustration of the above argument, I construed hierarchical, inert pseudo-random systems ("rats") on a personal computer and represented them as being located in a certain position on the screen. Every position on the screen had been randomly assigned an "environmental" number (one set of numbers for each "rat"). The rats had four different outputs, each corresponding to a step in one of the four possible directions of movement. As "background" stimuli, the "higher centers" of the rats were given a constant sequence of one selected input. The "reflex part" of the rats reacted to a much richer input; in the

first experimental series reported below the value of the input was for each occasion pseudo-randomly selected from a certain number of possibilities; in the second series it simply consisted in the number pre-assigned to the rats's present position. Think of the two input conditions as pseudo-random tones and local characteristics of the environment, respectively. The output of the rats was determined as described in connection with Figure 1, i.e. in effect as a random function of (full) input and the state of the "higher center". From time to time, the rats were subjected to a discriminative ("warning") stimulus signalling a hungry bird; they then had to go to the right in order not to be shocked - i.e., in order not to expose their "higher centers" to the full range of inputs in the next moment.

In Table 1, the figures denote the percentage of correct responses on the consecutive trials. Note that without learning the expected value for all trials would be 25%; also, that in the experiments the number of correct responses does not reflect only those "rats" which have learnt to criterion (full stability of correct response) but also those who from time to time perform correctly. In both series, the number of "rats" was = 1000, and each "rat" had to confront 20 "birds". The number of states in each "rat brain" was = 16, the inertness = 0.5625 and the number of possible inputs = 21 + warning + background = 23.

Table 1

Percentage of correct responses on the k:th trial

Trial no	1	2	3	4	5	6	7	8	9	10...18	19	20
1st series	28	37	41	47	52	56	60	63	65	68...78	81	80
2d series	15	24	29	35	40	40	40	40	42	46...48	48	48

The difference between the results can be partly explained by the fact that the input condition in the second series was very much less random-like than that of the first series. This is connected with the circumstance that the rats of the second series were often confined to small areas on the screen. Both series of results should of course be compared with the theoretical result for a pure RND device operating on a system with the present parameter values. If my argument is correct, the differences between this result (0.91, see above!) and the experimental observations are mainly to be explained by the imperfection of the approximations to the RND device.

Concluding comments

It should finally be pointed out that a combination of the previous theory of habituation and classical conditioning (Malm-

gren 1984) and the present model might turn out to be fruitful. Actual nervous systems almost certainly use dishabituation as one important signal for the lifting of information barriers on different levels in the hierarchy. The mechanism proposed in Malmgren 1984 for converting "number of state changes" to "amount of gross response" should then be supposed to be amplified through the arousal system. If this is the case, and the fundamentals of the present model is correct, the interesting consequence ensues that classical conditioning will always be (positively) operantly reinforced, since a "correct" conditioned response means that there is less state change between CS and UCS than if the "incorrect" response is present. However, the theoretical complexities which arise from such a combination of models are left for future investigations.

I have profited much from criticism by Björn Haglund. - Calculations and simulations were performed on an Apple IIc computer. Details are supplied on request. Author's present address: Dept. of Philosophy, University of Göteborg, S-412 98 Göteborg, Sweden.

References

1. Carlile, M.J. (Ed.) (1975). Primitive Sensory and Communication Systems. Academic Press, London etc.

2. Mackintosh, N.J. (1983). Conditioning and Associative Learning. Clarendon, Oxford; Oxford U.P., New York.

3. Malmgren, H. (1980). Om sannolikheten för inlärning i slumpvis sammansatta deterministiska system. (On the probability of learning in randomly composed deterministic automata. In Swedish.) Philosophical Communications, Green Series, 6. University of Göteborg.

4. Malmgren, H. (1984). Habituation and associative learning in random mixtures of deterministic automata. Göteborg Psychological Reports 14:2.

5. Skinner, B.F. (1981). Selection by consequences. Science, 213, 501-4.