REVISTA MEXICANA DE ANÁLISIS DE LA CONDUCTA
MEXICAN JOURNAL OF BEHAVIOR ANALYSIS
26, 229-249
NÚMERO 2 (SEPTIEMBRE)
NUMBER 2 (SEPTEMBER)

BEHAVIORAL PATTERNING BY NEURAL NETWORKS LACKING SENSORY INNERVATION

PATRONES CONDUCTUALES EN REDES NEURALES CARENTES DE INERVACIÓN SENSORIAL

STEVEN M. KEMP AND DAVID A. ECKERMAN[1]
UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL

ABSTRACT

A new approach is offered wherein behavior emitted by neural networks without antecedent stimuli is either shaped to produce a patterned behavioral output (Simulation 1) or is strengthened by delayed reinforcement through the mediation of response afterdischarges (Simulation 2). These networks demonstrate how Stein's In-Vitro Reinforcement (IVR) of neuronal bursts might account for various reinforcement effects at a behavioral level. The explorations presented illustrate two benefits to behavior analysis provided by biobehaviorally-based computational models of learning: accomodation of new biological information and recasting of behavioral concepts in ways compatible with this new information.

Keywords: neural networks, behavioral pattern, shaping, in-vitro reinforcement, biobehavioral learning models

RESUMEN

Se ofrece una nueva aproximación en la cual la conducta emitida por redes neurales sin estímulos antecedentes es moldeada para producir una salida de patrón conductual (Simulación 1) o es fortalecida por reforzamiento demorado a través de la mediación de postdescargas de respuesta (Simulación 2). Estas redes demuestran cómo el reforzamiento in-vitro (IVR) de explosiones neuronales propuesto por Stein puede dar cuenta de diversos efectos de reforzamiento a nivel conductual. Las exploraciones aqui

[1] Corresponding author: Steven M. Kemp, Dept. of Psychology, Davie Hall, CB# 3270 University of North Carolina, Chapel Hill, NC 27599-3270, USA. Phone: (919) 933-5447, email: steve_kemp@unc.edu, URL: http://www.unc.edu/~skemp/

presentadas ilustran dos beneficios que los models computacionales bioconductualmente basados proveen al análisis de la conducta: acomodación de información biológica nueva y modificación de conceptos conductuales en formas compatibles con esta nueva información.

Palabras clave: redes neurales, patrones conductuales, moldeamiento, reforzamiento in-vitro, modelos bioconductuales del aprendizaje

---

Neural network models can highlight new biological findings that may aid in developing an account of the biological mechanisms of behavior. In the present article, we highlight the observations by L. Stein of an operant "behavioral atom" (Stein, 1997; Stein, Xue, & Belluzzi, 1993; 1994; Stein & Belluzzi, 1989). We seek to connect the ideas of Stein and his collaborators to the phenomena of shaping a complex operant and of brief-delayed but effective operant reinforcement (action at a distance). Our goal is to see if a collection of behavioral atoms will act like an organism. In this manner, we hope our use of a neural network model will allow this new conceptualization to be entered more firmly into behavior analysis.

A second way that learning models can be useful to the experimental analysis of behavior is to encourage new descriptions of well-established functional relations that make it possible for current knowledge in neuroscience to influence how we organize our behavioral facts. In the present report we attempt to recast descriptions of the shaping of complex operants and of brief-delayed but effective operant reinforcement. We believe this recasting may open a path that will allow us to fit these phenomena into the developing behavioral neuroscience.

Stein describes the phenomenon he has studied as "in-vitro reinforcement" (IVR). In his preparation, a long-lasting change in neural activity results from the post-response infusion of the neuromodulator, dopamine. He and his colleagues work with the following procedure: A neuron (usually a pyramidal cell from the CA1 area of the hippocampus) that exhibits a characteristic occasional multi-spike burst (mediated by activity of L-type Ca2+ channels) is monitored in-vitro. Whenever a Ca2+ burst is detected, dopamine is injected around the cell via pipette. The burst rate is observed to increase. The basic notion is that initially random activity, when regularly followed by a biologically important event, will come to occur more often (reinforcement).

This increase has been demonstrated not to be the result of the effects of dopamine alone. Only the close temporal sequence of a calcium burst followed by dopamine results in this increased burst rate. Dopamine delivered at other times results in a slight decrease in burst rate. Likewise, bursts not followed by dopamine cause a decrease of the burst rate in vitro (extinction).

It is the delivery of dopamine contingent upon bursting that causes the increase in burst rate.

The conceptual basis of this IVR-based network is the R-S relation between a response and its reinforcing outcome, rather than the traditional S-R relation between an antecedent stimulus and an elicited response used in other computer models of learning. In the computer simulations we report, we model this R-S conception of learning without using any antecedent stimulus.

Furthermore, we believe that ours is the first artificial neural network to implement McCulloch & Pitts (1943) original suggestion for modeling learning in neural networks by using variable thresholds instead of the presently more common variable connection weights. In IVR, the burst rates of the pyramidal cells are presumed to be due to intrinsic properties of the cells rather than due to variable strengths of interconnections between cells. Each of our simulated neurons has a variable threshold whose level determines the burst rate. Connection strength is not modified.

We also believe it is the first network of variable-structure stochastic learning automata (Narendra & Thathachar, 1989) using a learning algorithm that is derived directly from neurophysiological evidence. Similarly to so-called Reinforcement Learning systems (Sutton & Barto, 1998), both networks presented here use only one-bit binary (reward/no reward) feedback across the entire network. Typical feedforward networks (such as backpropagation networks) receive vector feedback across different parts of the network, simplifying the "credit assignment problem" (Staddon & Zhang, 1991). Even most reinforcement learning networks require scalar feedback (indicating a quantity of reinforcement) rather than just binary feedback to operate effectively. Given the relative inefficacy of changes in amount of reinforcement provided in animal studies, the dependence on scalar feedback is not especially behaviorally plausible. With these foreshadowings, we proceed directly to the simulations

Simulation 1 -- Shaping a neural network

In the first simulation, a series of black and white images are generated by the net. Rather than viewing these images as stimuli, however, each image should be viewed as a complex response, with each binary pixel (0/1) constituting one element of the complex topography. No antecedent stimulus is supplied. In fact, the neural network, called Clavier, lacks any model of sensory systems whatsoever. As such it resembles the preparations discussed by P. Weiss in response to Lashley's canonical treatment of the problem of serial order (cited in Lashley, 1951, pp. 140-142). The goal of the simulation is to have the Clavier network emit one specific "target" response chosen

(arbitrarily) by the user. Another part of the simulation system, external to the network, evaluates the similarity of the network's output to the target and delivers contingent reinforcement in an effort to shape the network's behavior to the target.

As with all elements of any computational system, each response/image emitted by the Clavier network is made up of a string of ones and zeros. For the simulations reported here, each successive image is made up of 64 such bits. Each response can be considered to be an 8 by 8 image made up of squares where each is either black or white. After each response (image) is emitted, the external shaping module judges it against the most recent series of responses. A simple measure indicates how close this particular response (pattern) is to a "target" image. If it meets a criterion of being closer than its predecessors, it is reinforced according to a percentile reinforcement schedule (Platt, 1973; Galbicka, 1994). This is another important feature of the present simulation. It is very rare for neural networks to be trained using methods with well-established track records of success in conditioning real animals. The percentile reinforcement schedule has demonstrated its ability to shape behavior in both the laboratory and clinic, with both non-human and human organisms.

Gradually, the images move closer to the target until the neural network generates an exact copy of the target – shaping is complete. This gradual approach to the target pattern is accomplished by adjustment of the thresholds in each of the 64 units (one for each pixel in the image) that determines whether a particular response element is likely to be black or white. Each unit is a model of one of Stein & Belluzzi's pyramidal cells. Each pixel emitted on each iteration cycle indicates either a burst (one) or no burst (zero) in that interval. As each threshold shifts, each element learns to be one or the other color. The total pattern of ones and zeros models the pattern of bursting activity in some portion of the cerebral cortex, with different patterns of activity presumably producing different topographies.

In computer science terms, this sort of procedure is called a search task. The measure of difficulty of a search task is the number of distinct elements (responses). There are 2^{64} possible 64-bit long binary strings. A non-learning system using a systematic search would take an estimated mean of 9 billion billion cycles to reach the target. The expected time to reach the target for a strictly random search (the so-called British Museum algorithm, Newell, Shaw, & Simon, 1958) is double that. Typical learning systems using scalar or vector feedback complete the search in far shorter times.

The learning algorithm is implemented with a network of variable-structure stochastic learning automata (VSLAs), each of which includes an independent threshold that determines the probability of a black or white output in one position. In our simulations, an eight by eight image is duplicated in

anywhere from 25 thousand to 250 thousand cycles. Using Stein & Belluzzi's (1989) definition of cell-bursting to estimate the cycle time (at 50 ms), this would be equivalent to a range of approximately 20 minutes to 3 hours 20 minutes in real time. If the thresholds are modified using a linear (truncated) learning rule, the target is approached asymptotically. If a damped learning rule is used, the network rapidly stabilizes, emitting only the target (criterial) response shortly after shaping is completed (see below).

METHOD

Each unit in the Clavier network, called an emitter unit, is a variant on the standard McCulloch-Pitts cell with a threshold between zero and one and a random activation, also between zero and one. If the activation exceeds the threshold, a signal (one) is output by that emitter unit. (When a unit emits a one, we say that the unit has fired. When it emits a zero, we say that it did not fire.) Thus, lower thresholds mean higher mean firing rates and higher thresholds mean lower mean firing rates. Depending upon which signal (one or zero) is output by that unit and what reinforcement signal (also one or zero) is provided to the entire network by the training system, the threshold of that unit is either raised or lowered by a calculated amount.

For each unit, on each cycle, there are four possibilities. Either the unit signals or it does not. Either the network is reinforced or it is not. The learning rule for altering the threshold can thus be specified by a two-by-two table (c.f. Figure 1).

The network is trained to reach the solution using a training procedure called the percentile reinforcement schedule (Platt, 1973; Galbicka, 1994). This procedure requires that we specify the probability that a criterion response will be reinforced (p) and the number of prior responses against which the present response will be compared (m). In our case $p = .30$ and $m = 20$.

The search task in our simulation proceeded as follows: On each cycle (trial) Clavier output a string of ones and zeros that was matched to the black and white target pattern. The difference between the two pictures was interpreted as a numerical distance (inverse similarity). Our measure of (dis)similarity was the Hamming distance. Hamming distance is calculated as the number of mismatching element-pairs between the current and a target pattern. (Thus the Hamming distance for these simulations ranged from 0 -- perfect match -- to 64 -- all cells mismatching the corresponding pixels.) If this Hamming distance was closer than thirty percent of the most recent 20 distances, reinforcement was provided. Reinforcement, when given, lowered the thresholds for all units that had been active (white) on that trial. Thresholds

for units that had not been active on that trial were not altered. On non-reinforced trials the thresholds for bursting cells are raised and again, thresholds for inactive units remained unchanged.

Output Rule (y_i indicates whether unit i bursts):

$$y_i = \begin{cases} 1 \text{ iff } x_i > \tau_i \\ 0 \text{ otherwise.} \end{cases} \qquad \begin{array}{l} x_i \sim U[0,1] \\ 0 \leq \tau_i < 1 \end{array}$$

where x_i is the activation of unit i and τ_i is its threshold.

Learning Rule ($\Delta\tau_i$ is the change in the threshold for unit i):

$\Delta\tau_i =$	S^{R+}	$\sim S^{R+}$
y_i	$-\lambda$	$+\delta$
$\sim y_i$	0	0

where λ is the learning increment (set to .01)
and δ is the "decay" increment (set to .0043).

Figure 1. Algorithms for Clavier. The basic equations used to simulate In-Vitro Reinforcement (IVR). Activation is presumed to be intrinsic to the unit and is simulated with a pseudo-random variate Uniform over [0,1]. Burst rates vary inversely with value of variable threshold, t. Changes in burst rate are governed by changes in threshold. Changes in threshold only occur after a burst. Each unit emits bursts and alters its burst rate independently of other units. Different values of increments, l and d, were used in different simulations. Each unit is a Variable-structure Stochastic Learning Automaton and takes no inputs except for reinforcement

The threshold increment and decrement were set so that there was a 3 to 7 ratio for the sizes of the threshold increase (called decay rate since it represents the effect of extinction) to the size of the threshold decrease (called learning rate since it represents the effect of reinforcement). This ratio was set

to balance the overall probability of extinction (7 of 10) to probability of reinforcement (3 of 10). Subject to this ratio of extinction to reinforcement over time, the threshold of a hypothetical cell whose activity had no effect on the match of output (response) to the target would remain at a constant level (on average). The use of the Hamming distance guarantees that every cell has some effect on the overall match. Therefore, cells corresponding to white pixels on the target tend to fire more and cells corresponding to black pixels on the target tend to fire less as learning progresses.

After thresholds were adjusted, another cycle commenced, another output pattern was generated and so forth. The simulation stopped when Clavier duplicated the target pattern exactly, successfully completing the search.

RESULTS

Monte Carlo testing to criterion. Across many simulations, we have demonstrated that, using this training system, the Clavier network can be trained to duplicate any binary sequence given it. We undertook a series of Monte Carlo simulations to answer three specific questions: First, what proportion of simulations of Clavier match a randomly entered target exactly and how long does it take? Answer: Of the 1200 simulations we completed with pseudo-random patterns (with around 50% white elements), the target was found in all cases. The slowest learning required 278,901 cycles. Second, we observed that learning by Clavier is asymmetric with respect to shifting toward "white" or "black" (bit values of one or zero). Once a threshold begins to rise, the cell bursts less often and therefore is less subject to change. It is harder to shift a threshold down than up because there are fewer opportunities to move a higher threshold. This creates a bias toward going "black." We therefore tested to see if the proportion of zeros to ones in the target string would affect search time. Answer: Though all targets were reached, in general, targets with more "white" where learned more slowly than targets with more "black." This effect is demonstrated in Figure 2. Third, in reacting to our initial observations, a well known computational model builder (R. D. Luce, personal communication) suggested that we add a damping component (see appendix) that shifted the thresholds more smoothly and more readily toward the extremes (low or high) without the necessity of truncation. This damping component was a bidirectional variant on the old Bush-Mosteller (Bush & Mosteller, 1951) asymptotic learning rule. Did it affect performance? Answer: Yes. The damping function reduced the "black bias" (see Figure 2). The damped version of Clavier tended to reach all targets in a mean time

comparable to the best times of the original algorithm.

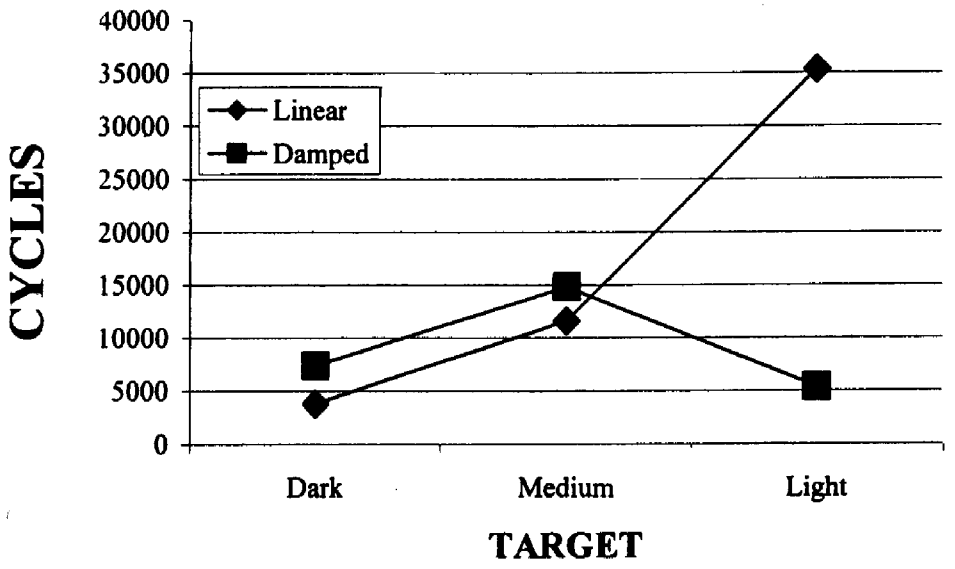

Figure 2. Search results for Clavier. Mean number of cycles to emission of first perfect match to target for Clavier/Police Artist simulations. The three values (Dark, Medium, and Light) across the X-axis represent the proportion of black pixels in the 64-pixel black and white targets. Diamond-shaped points are means from simulations using the linear (truncated) learning rule. Square points are means from simulations using the damped learning rule. Each point represents the mean of 200 simulations (20 targets x 10 pseudo-random seeds)

Steady-state responding under continuous reinforcement. The very large stochastic component of the Clavier network means that the first emission of the target response is no guarantee of further success. Given the relative stability and reliability of the damped version of Clavier (see above), that version was selected for an attempt to condition Clavier to a steady state.

The percentile reinforcement schedule was used for all sub-criterial responses. Target responses (Hamming distance of zero) were always reinforced (crf). Figure 3 shows a cumulative record of Clavier's criterial responses. Baseline (m) for the percentile reinforcement schedule was increased from 20 to 200. Learning rate was increased from .01 to .35 and decay from .0043 to .15. (Such large increments produced instability with the linear learning rule, but worked well with damped learning.) The first criterial response was emitted on the 6378th cycle, equivalent to approximately 5 minutes and 20 seconds of real time. Within a few simulated seconds, two

more criterial responses were emitted. Almost immediately thereafter, responding proceeded at a near maximal rate. (The maximal rate here is limited only by the definition of cycle time and is thus unrealistically high, particularly for a complex response. This problem was resolved in Simulation Two, below.)

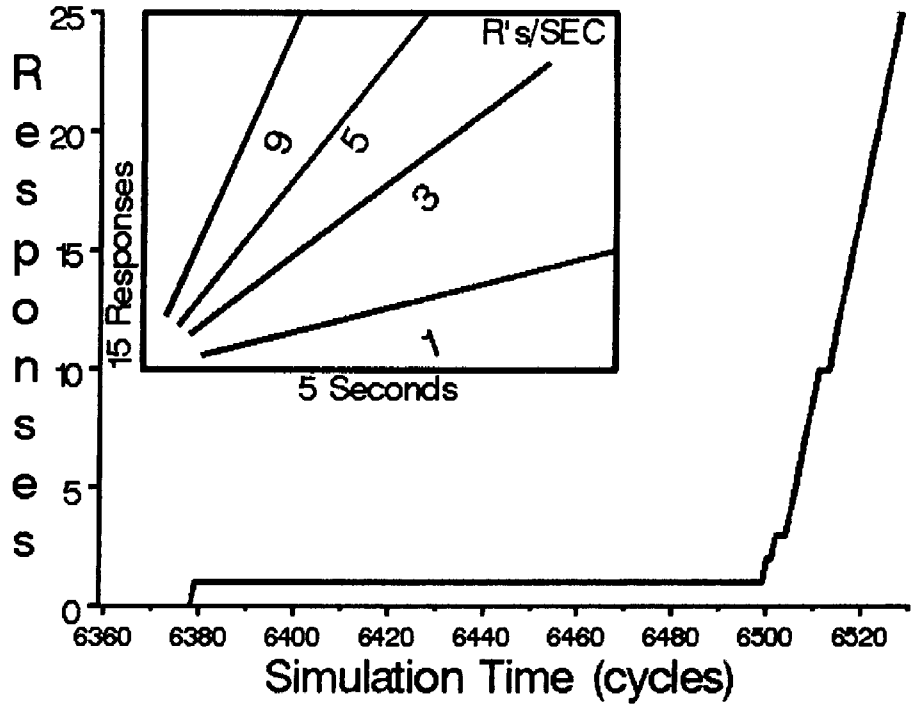

Figure 3. Clavier achieves steady-state. Cumulative record showing the cumulative number of successful target responses from the Clavier model. Note that the model successfully completes the response the first time at cycle 6378. Since cycles take a (simulated) 50 ms this means the first successful response was observed after approximately 5 minutes. Responses then deviate from the target pattern until cycle 6499. The correct responses then come rapidly as the model continues to duplicate the target response for the remainder of the simulation. The eventual rate of target response emission was approximately 10 per sec. This rate is so high since the Clavier model does not include a factor to acknowledge that the response itself takes time. This attribute is added in the second simulation (cf.)

DISCUSSION

Thus, the Clavier network has successfully linked a real training procedure from the psychological laboratory and clinic (shaping by the use of

a percentile reinforcement schedule) to a particular sort of neural plasticity found in the neurophysiological laboratory (IVR reinforcement). This accomplishes what we indicated above as the first benefit provided by neural network modeling.

With regard to the second benefit, note that the architecture of Clavier is distinct from most neural networks in that there is no input layer. Other than the reinforcement, there is no stimulus information provided to the network. In fact, we call the simulation *Police Artist*, because the task for the neural network is to draw a picture of a target it never sees. In providing such a network we believe we have demonstrated the second benefit noted above – to show what new potential may be achieved in our conceptualizing of learning phenomena if we incorporate new information coming from neuroscience. Our second simulation emphasizes this benefit even more strongly.

In addition we would like to emphasize that unlike most neural network training procedures, the present percentile reinforcement schedule has been well established as an actual training method used successfully with actual human and non-human subjects in both laboratory and classroom settings. Aside from work by Gullapalli (1990), the completed simulation is the only use of a variant of the method of successive approximations to train a network in a search task of which we are aware. Unlike the simulation presented here, Gullapalli did not use a training method validated with real organisms.

Finally, by the standards of artificial intelligence and machine learning, Clavier is a slow learner. More sophisticated search algorithms, particularly those receiving more information from the environment, can search the space of 64 bit binary strings far faster than Clavier. It must be remembered that Clavier works with only one bit of information per cycle, far less than neural networks using supervised learning (Rumelhart, McClelland, & The PDP Research Group, 1986) or reinforcement learning (Sutton & Barto, 1998). Further, most fast algorithms offer little in the way of biological plausibility and the sole biological phenomenon adumbrated as the basis for learning is inevitably the ubiquitous long-term potentiation (Bliss & Lømø, 1973; Bliss, & Colingridge, 1993).

Simulation 2 -- Action at a distance: Timing and Delay

One of the more interesting things that researchers combining Behavior Analysis and Neural Networks have been able to do is to take a new look at old problems. The best example of this is Donahoe, Palmer, & Burgos' (1997) study that looked at the age-old dichotomy between elicited and emitted responses. The authors showed that, at a biological level, individual responses might have characteristics of both elicitation and emission, categorizable as

either at a behavioral level, depending upon circumstances.

The second simulation of the present paper provides a second look at another old controversy: how to explain the effects of delayed reinforcement. We used a neural network model to look at delays of about one second or so. Within learning theory, two approaches have emerged, which we will call here "Two Factor Theory" and "Action at a Distance."

Two Factor Theory asserts that, in order for reinforcement to be effective, it must come immediately after responding. Thus, when the primary reinforcer comes after some delay, there must be one of a series of conditioned reinforcers that precedes it. Conditioned reinforcers are presumed to be produced through Classical Conditioning.

Wolfram Schulz (1997; Schultz, Dayan, & Montague, 1997) and others are doing research on a possible neurophysiological analogue to conditioned reinforcement, working with cells that release diffuse dopamine. This supports one way that the nervous system might extend the time for effective reinforcement. We are, however, not fans of Two Factor Theory. It seems too elaborate a use of higher order conditioning for us.

Action at a Distance asserts that the functional behavioral relations include the temporal relations and that, given regular and reliable relations between delay and efficacy of reinforcement, the effects observed are the effects of operant conditioning simpliciter and that no other sort of conditioning need be postulated at the behavioral level.

A characteristic of the Action at a Distance account is that, by intention, it offers no specific alternative to Two Factor Theory. Presumably, reinforcement is efficacious because some neurophysiological mechanism is susceptable to the effects of reinforcement for some period of time after a response. We try here to make that assumption a bit more concrete. We model a way the nervous system might carry out Action at a distance without a cascade of classically conditioned associations: Our proposal is that there might be a response afterdischarge of neural activity that effectively extends the response.

The Vibraphone network. The neural network of this second simulation is also based on the In-Vitro Reinforcement (IVR) of Stein & Belluzzi (1989). This second network, called Vibraphone, differs from Clavier in a number of respects. Two of the most important are a more neuroanatomically plausible structure and attention to temporal considerations in both design and implementation. In Clavier, no attempt was made to give an account of how a particular pattern of bursting would lead to a particular sort of behavior or motor movement. The architecture of Vibraphone, by contrast, makes use of recent reviews of functionally suggestive aspects of cortical neuroanatomy (Abeles, 1991; Valiant, 1994, Crick & Asanuma, 1986). Further, every event

within the Vibraphone network is modeled as occurring at a location. Any effect of one event upon another takes some specified amount of time, which is included in the model. Also, every event has a duration, which is also modeled.

The reinforced bursting of pyramidal neurons in hippocampal slices maintained in vitro is seen as a neural step or two prior to a response that might produce a reinforcer after a brief delay. Over and above the delay between overt response and reinforcer, there is the lag between the burst and the subsequent neural activity that generates the muscular movement, the time it takes for the muscular movement itself to occur, the time between the delivery of the reinforcer and the neural activity that releases the dopamine, and the time it takes for the signal from the Ventral Tegmental Area to arrive at the cortex and produce the diffusion of dopamine. If cells exhibiting IVR are, in fact, instantiations of Skinner's "behavioral atoms," then there must be a neurophysiological analogue to the delay between response and reinforcer. It is the delay between the Ca2+ burst that initiates the response and the diffusion of dopamine resulting from the reinforcer.

From a spatial perspective, delivery of the reinforcer can occasion diffusion of dopamine in the vicinity of the cells that fired to initiate this response. There is, however, a major problem that is demonstrated in a figure presented by Stein & Belluzzi: As shown in Figure 4, the In Vitro Reinforcing effects of dopamine on cell bursting are severely diminished when the diffusion of dopamine is delayed by as little as 100 ms and are effectively eliminated 200 ms. As noted above, the series of delays between pyramidal burst and dopamine diffusion, taken along with the normal response-reinforcer delay, makes for a compound delay easily an order of magnitude greater than 200 ms. That leaves a very long temporal gap (Skinner, 1984, p.722; Skinner, 1988, p.470) that must be filled in by a presumed cascade of conditioned reinforcers. Our model seeks to fill in a second or so of this gap by extending the period wherein the cortical neurons fire.

We propose the following alternative mechanism for bridging this temporal gap. Any biobehavioral model includes models of neural activity that eventually produce motor activity. S. Glenn (personal communication) advocates the notion that rather than conceive of such neural activity as the *cause* of a response, that it is better understood as a *part* of the response. This reconceptualization has many important ramifications. For our present purposes, the most important is that there may he (central) nervous activity that is concurrent with, or even subsequent to, the motor activity that may also be considered a proper part of the response. An afterdischarge of cortical activity may constitute a temporal extension of the response itself. As Lashley (1951, p.120) put it: "The fact of continued activation or after-discharge of

receptive elements and their integration during this activation justifies the assumption of a similar process during motor integration."

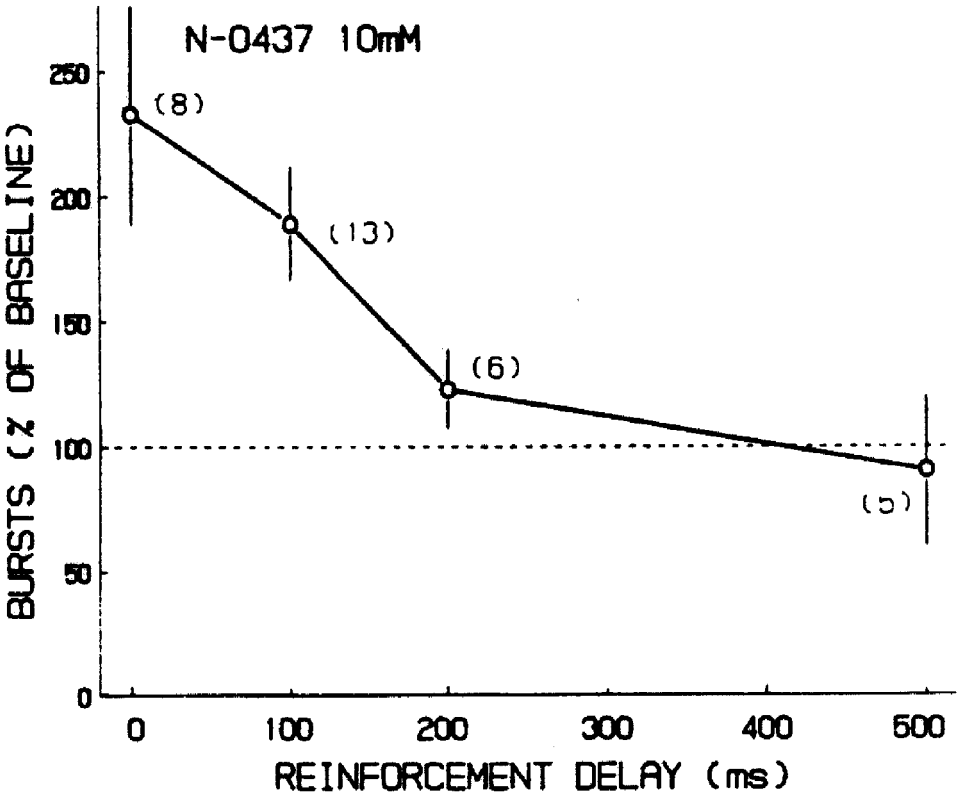

Figure 4. Delay of In-Vitro Reinforcement. Delay of reinforcement gradient in neuronal operant conditioning from Cellular investigations of behavioral reinforcement, by L. Stein & J. D. Belluzzi, 1989, *Neuroscience and Biobehavioral Review, 13*, p.75. Copyright 1989 by the Pergamon Press. Reprinted with permission. The number of bursts (y) for hippocampal pyramidal neurons measures the degree of conditioning by the diffusion of dopamine (10mM of N-0437) x ms after a burst. The number in parentheses is the number of neurons tested

The basic idea is that, should a burst occur early in the process that eventually produces a motor movement, then backward inhibitory projections (presumably from the primary motor cortex back to the secondary motor cortex) briefly dampen the bursting both of the cell that fired originally and also of cells that are near neighbors. The dynamics of the network are such that after a

period wherein activity is suppressed, upon release of that suppression, the suppressed units recover with increased activity. With the simple addition of the postulate that near neighbors of pyramidal cells are more likely to produce similar responses, the delayed bursting of neighbor cells can be timed so as to make them susceptible to diffuse dopamine at a point well after the motor activity has ceased and dopamine due to primary reinforcement is, in fact, diffusing throughout the area.

METHOD

The proposed architecture for Vibraphone is shown in Figure 5. The network consists of two layers, with activation proceeding from left to right from deeper in the cortex toward the motor systems (as indicated by the arrowheads on the connections).

Each triangular element is a unit modeling a pyramidal cell. All pyramidal units produce Calcium ($Ca2+$) bursts as well as the more usual Sodium ($Na+$) spikes. Pyramidal units in the first layer have variable thresholds similar to those of Clavier, with variable burst rates that increase with dopamine diffusion within 100ms of a burst. Cells in the second (rightmost) layer have fixed thresholds. Each pyramidal cell is paired with a smooth stellate cell (indicated by the star-shaped elements in the diagram). Smooth stellate cells such as chandelier cells and basket cells (Abeles, 1991; Valiant, 1994) in the cortex have strongly inhibitory, axo-axonic connections to nearby pyramidal cells. When these stellate cells are activated, all activity in the nearby pyramidal cells is suppressed.

The network is structured vertically as well as horizontally. Groups of cells, modeled after microcolumns in the cortex each govern a separate, competing response (labeled to the right of the second layer). The diagram shows three columns, governing three responses. In principle, of course, there could be many more. We have conducted simulations with up to four responses and twelve cells per group. Here, we report results from a network with two columns producing two responses, with six pyramidal cells in the first layer of each column.

Key to the function of the network is the firing of pyramidal cells in layer 2. There is one pyramidal cell in layer 2 for every group of cells in layer 1. Bursts (but not spikes) in layer 1 produce spikes in layer 2. (Spikes in layer 1 are modeled as background activity using) When the spike activity in the layer 2 pyramidal cell is high enough (as governed by the fixed threshold in that cell), the layer 2 pyramidal cell emits a burst. Any burst in a layer 2 pyramidal cell immediately initiates the corresponding response. In addition, backward

projections to the stellate cells in the corresponding group (same column) in layer 1 cause all pyramidal cells in the group that generated the response to be strongly inhibited. Finally, lateral projections across layer 2 to those stellate cells create strong inhibition across all of layer 2.

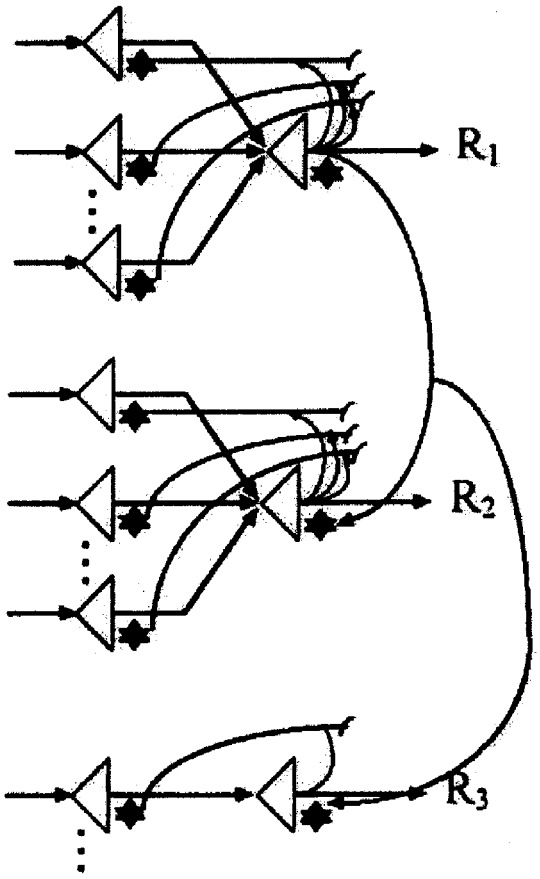

Figure 5. Vibraphone architecture. Units and connections for the Vibraphone model. Triangles are pyramidal (bursting) units. Stars are smooth stellate (inhibitory) units. Pyramidal units in Layer 1 (on left) have pseudo-random activation and variable burst thresholds as in Clavier (cf.). Pyramidal units in Layer 2 (on right) have activation as a function of bursting of connected pyramidal units from Layer 1 and fixed burst thresholds. Each stellate unit completely inhibits its neighboring pyramidal unit. Each group of units (from top to bottom) governs one competing response. Bursts (but not spikes) from pyramidal cells in Layer 2 activate corresponding response (Rx), all stellate cells in same group from Layer 1, and all stellate cells in all groups in Layer 2

This specific architecture implements the four features we believe are necessary to correct functioning of the network, namely: (1) Forward projections are excitatory and backward connections are inhibitory. (2) Lateral inhibition widens further "downstream" (closer to motor systems). In the present network, this is implemented by having forward projections focus down like a funnel from more units to fewer units. (3) Processes (refractory times and inhibitory durations) lengthen downstream.

We assume that, under ordinary conditions, these neighboring cells burst asynchronously and are all in different stages of readiness at different times. When a burst or parallel bursts at stage one generate(s) a burst at the stage two of this sequence, this blanket of inhibition is passed back and suppresses all the connected cells of stage one. By blanketing this neighborhood with inhibition, all these cells complete their refractory phase during the period of shared inhibition. Given each neuron's tendancy toward spontaneous activity, when the inhibition ceases, a synchronized volley of bursting is likely to occur in that one neighborhood, while the other neighborhoods continue in their asynchronous bursting. Should a flush of diffuse dopamine enter the entire stage one region at this point, few cells in the region will be bursting, *except* in the neighborhood of the cells that produced the response. Using the notion of IVR, the pyramidal cells in the neighborhood of the cell that originally initiated the motor movement will be differentially reinforced, since they will be bursting and their windows of susceptibility will be open. In fact, if there continues to be a situation that induces activity of these neurons and if dopamine continues to be diffused, this group of neighbor cells will continue an extended, synchronized volleying.

This afterdischarge of synchronized volleys of Ca2+ bursts effectively extends the temporal window that Stein & Beluzzi found for individual hippocampal cells for a network of interconnected cells. This (covert) temporal extension of responding via a neurally plausible mechanism offers a basis for developing a neural model for Action at a Distance: the extended, synchronized bursting of cells with similar motor function after the motor movement.

RESULTS

Computer simulations of this model, using SAS/IML® Software (1996), have just begun. We do, however, have some interim results.

Key to demonstrating the efficacy of reinforcement across brief delays is the ability to differentially reinforce a particular response. A simulation was designed with what we believe are realistic durations set for spikes, bursts, inhibitory effects, refractory effects, motoric movements, sensory detection of

reinforcement, and dopamine diffusion to the cortex. A network with two response groups was constructed and the network was trained with continuous reinforcement (crf) for one response with extinction (ext) for the other. After an initial period of conditioning, the schedules were reversed with the previously reinforced response now under extinction and vice versa. Finally, the schedules were reversed again for an A-B-A design.

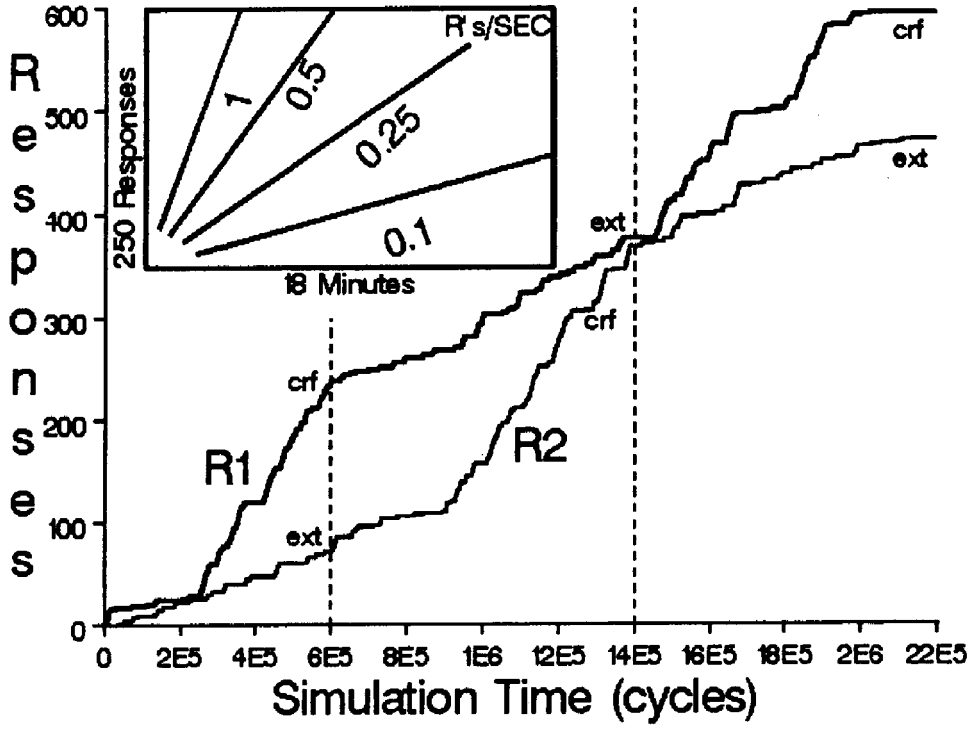

Figure 6. Response differentiation under brief-delayed reinforcement by Vibraphone network. Simultaneous cumulative records of Response 1 (upper, thicker line) and Response 2 (lower, thinner line) under alternating continuous reinforcement (crf) and extinction (ext). No discriminative stimuli supplied. Each iteration cycle (on the X-axis) models 1 ms of real time. Clear alternations of response rate occur following schedule changes (indicated by vertical dashed lines). Effects of response competition can be observed in brief bursts of responding for one operant when responding slows on other operant. Low rates for both responses at end of simulation are due to satiation effect

Over a wide variety of parameter settings and even minor changes in algorithm, the system was tuned to produce a pattern of behavior such as that

shown in Figure 6 with moderate reliability. Such consistent results indicate that it is the general approach and architecture, as illustrated in Figure 5, and not the details of the implementation nor specific parameters that are the guarantors of the successful simulation of response differentiation accomplished here.

Obviously, we would prefer to have seen stronger and more abrupt effects than those demonstrated. Response rates for this model are, however, strongly constrained by the inhibitory and refractory effects that insure that the responses are competitive and that the timing between initial burst in layer 1 and eventual motor movement is realistic. Changes in response rate are thus difficult to obtain.

As a last comment, note the leveling off of both cumulative records at the end of the session. This effect is due to all thresholds rising to the point where no amount of reinforcement can produce further responding. It is roughly analogous to satiation in that the dynamics of responding and reinforcement cause the system to drift into a state where reinforcement is inefficacious. If another model of satiation is desired, parameters can be reset to prevent this decline in reinforcer effectiveness.

DISCUSSION

The principal value of the success of the Vibraphone network in response differentiation over brief-delayed reinforcement is that the very narrow temporal window of susceptibility to dopamine reinforcement found in IVR can no longer be counted as evidence against Stein's hypothesis that the IVR mechanism may constitute the substrate of Skinner's behavioral atoms. A relatively simple network of IVR-capable cells can produce responding under control of operant reinforcement delayed in excess of the 100ms temporal window for individual neurons.

Additionally, we hope we have shed some new light on the old controversy found in the learning theory literature between Two Factor Theory and Action at a Distance. Even in contemporary treatments of delayed reinforcement such as those of Sutton & Barto (1998), Classical Conditioning is the principal, if not the sole model of how an organism produces effective sequences of behavior in serial order with primary reinforcement available only at the completion of the task. We suspect that, in real organisms, there are multiple mechanisms that create these behavior sequences. Recall that these are the kinds of sequences for which Lashley challenged us to account (1951).

If both the Classical conditionability of dopaminergic cells and the kind of response aftercharge addressed in our proposal prove able to account for

ways the nervous system spans the temporal gap of delayed reinforcement (from both ends), then it is possible that both processes together (perhaps with additional processes still to be envisioned) can account for the effectiveness of delay of reinforcement. If so, then biobehavioral approaches may have again demonstrated their ability to resolve an old controversy in learning theory by demonstrating harmony between purported alternatives. That is, often what appear to be alternatives are better thought of as complementary.

GENERAL DISCUSSION

We hope that the simulation research reported here provides an illustration of the benefits to behavior analysis uniquely obtainable by neural network research. Simulated behavior by plausibly structured networks of units exhibiting activity analogous to that found in the biological laboratory here resembles actual behavior from the operant laboratory. These results make plausible that particular biological phenomena underlie particular behavioral phenomena. This evidence is strengthened due to what P. Weiss refers to as the "rigorous limitations [placed] upon the free flight of our fancy in designing models of the nervous system" by the constraints of the ever-increasing amount of neurophysiological data available to the modeler (Lashley, 1951, p. 140). Neural networks can be used to lift flights of speculative fancy or to co-constrain these speculative fancies with a combination of biological and behavioral facts.

We have noted several times that the biological systems we have simulated here lack discriminative stimuli, or even the sensory apparatus to detect discriminative signals. While no organisms have such a structure, laboratory preparations with these characteristics have a long history (Lashley, 1951, pp. 140-142). Such preparations, like our simulation, demonstrate ongoing organized, rhythmic activity in the absence of stimulus input. In the cited passage, Weiss points out that the then contemporary models of neural systems were unrealistic because they lacked many of these properties. Even now, nearly 50 years later, the same can be said of most neural networks. (However, see Donahoe, Palmer, & Burgos, 1997, and Edelman, 1992, for networks with some of these properties.)

The clear consensus of the biobehavioral experts in 1951 was that reflexology and associationism, however dramatically elaborated, could not be reconciled with the facts of neurophysiological dynamics. Neural network simulations provide us with a tool that Lashley and his contemporaries did not have. With this tool, we can begin the reconciliation they looked forward to half a century ago.

REFERENCE NOTE

Luce, R. D. Personal communication, May, 1996.
Glenn, S. Personal communication, May, 2000.

REFERENCES

Abeles, M. (1991). *Corticonics: Neural circuits of the cerebral cortex*. Cambridge: Cambridge University Press.

Bliss, T. V. P., & Colingridge, G. L. (1993). A synaptic model of memory: Long-term potentiation in the hippocampus. *Nature, 361*, 31-39.

Bliss, T. V. P., & Lømø, T. (1973). Long-lasting potentiation of synaptic transmission in the dentate area of the anaesthetized rabbit following stimulation of the perforant path. *Journal of Physiology, 232*, 331-356.

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review, 58*, 313-323.

Crick, F., & Asanuma, C. (1986). Certain aspects of the anatomy and physiology of the cerebral cortex. In J. L. 'McClelland and D E. Rumelhart (Eds.), *Parallel Distributed Processing: Explorations in the microstructure of cognition, Volume II: Psychological and biological models* (ch. 20, pp.333-371). Cambridge, MA: Bradford Books/MIT Press.

Donahoe, J. W., Palmer, D. C., & Burgos, J. E. (1997). The S-R issue: Its status in behavior analysis and in Donahoe and Palmer's *Learning and Complex Behavior*. *Journal of the Experimental Analysis of Behavior, 67*, 193-211.

Edelman, G. M. (1992). *Bright air, brilliant fire: On the matter of the mind*. New York: Basic Books.

Galbicka, G. (1994). Shaping in the 21st Century: Moving percentile schedules into applied settings. *Journal of Applied Behavior Analysis, 27*, 739-760.

Gullapalli, V. (1990). A stochastic reinforcement learning algorithm for learning real-valued functions. *Neural Networks, 3*, 671-692.

Lashley, K. S. (1951). The problem of serial order in behavior. In L. A. Jeffress (Ed.), *Cerebral mechanisms in behavior: The Hixon Symposium* (pp. 112-146, with discussion). New York: Wiley.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115-133.

Narendra, K. S., & Thathachar, M. A. L. (1989). *Learning Automata: An introduction*. Englewood Cliffs, NJ: Prentice-Hall.

Newell, A., Shaw, J. C., & Simon, H. A. (1958). Elements of a theory of human problem solving. *Psychological Review, 65*, 151-166.

Platt, J. R. (1973). Percentile reinforcement: Paradigms for experimental analysis of response shaping. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation, Volume VII: Advances in research and theory* (pp. 271-296). New York: Academic Press.

Rumelhart, D. E., McClelland, J. L., & The PDP Research Group. (1986). *Parallel Distributed Processing: Explorations in the microstructure of cognition, Volume I: Foundations*. Cambridge, MA: Bradford Books/MIT Press.

SAS Institute. (1996). SAS/IML® [Computer software]. Cary, NC: SAS Institute, Inc.

Schultz, W. (1997). Dopamine neurons and their role in reward mechanisms. *Current Opinion in Neurobiology, 7*, 191-197.

Schultz, W., Dayan, P., & Montague, R. R. (1997). A neural substrate of prediction and reward. *Science, 275*, 1593-1599.

Skinner, B. F. (1984). Reply to Harnad. *Behavioral and Brain Sciences, 7*, 721-724.

Skinner, B. F. (1988). Reply to Harnad. In A. C. Catania, & S. Harnad (Eds.), *The Selection of Behavior: The Operant Behaviorism of B. F. Skinner: Comments and Consequences,* (pp. 468-473) . Cambridge: Cambridge University Press.

Staddon , J. E. R., & Zhang, Y. (1991). On the assignment-of-credit problem in operant learning. In M. L. Commons, S. Grossberg, & J. E. R. Staddon (Eds.), *Neural network models of conditioning and action: A volume in the quantitative analyses of behavior series* (ch. 11, pp. 279-293). Hillsdale, NJ: Lawrence Erlbaum.

Stein, L., & Belluzzi, J. D. (1989). Cellular investigations of behavioral reinforcement. *Neuroscience and Biobehavioral Reviews, 13*, 69-80.

Stein, L. (1997). Biological substrates of operant conditioning and the operant-respondent distinction. *Journal of the Experimental Analysis of Behavior, 67*, 246-253.

Stein, L., Xue, B.G., & Belluzzi, J. D. (1993). A cellular analogue of operant conditioning. *Journal of the Experimental Analysis of Behavior, 60*, 41-53.

Stein, L., Xue, B.G., & Belluzzi, J. D. (1994). In vitro reinforcement of hippocampal bursting: A search for Skinner's atoms of behavior. *Journal of the Experimental Analysis of Behavior, 61*, 155-168.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Valiant, L. G. (1994). *Circuits of the mind.* Oxford: Oxford University Press.