A Connectionist Formulation of Learning in Dynamic Decision-Making Tasks

Faison P. Gibson
Graduate School of Industrial Administration
Carnegie Mellon University
Pittsburgh, PA 15213–3890
gibson+@cmu.edu

David C. Plaut
Department of Psychology
Carnegie Mellon University, and
Center for the Neural Basis of Cognition
Pittsburgh, PA 15213–3890
plaut@cmu.edu

Abstract

A formulation of learning in dynamic decision-making tasks is developed, building on the application of control theory to the study of human performance in dynamic decision making and a connectionist approach to motor control. The formulation is implemented as a connectionist model and compared with human subjects in learning a simulated dynamic decision-making task. When the model is pretrained with the prior knowledge that subjects are hypothesized to bring to the task, the model's performance is broadly similar to that of subjects. Furthermore, individual runs of the model show variability in learning much like individual subjects. Finally, the effects of various manipulations of the task representation on model performance are used to generate predictions for future empirical work. In this way, the model provides a platform for developing hypotheses on how to facilitate learning in dynamic decision-making tasks.

Introduction

In business-related decision-making tasks, such as managing production output, decision makers make multiple recurring decisions to reach a target, and they receive feedback on the outcome of their efforts along the way. This type of *dynamic* decision-making task can be distinguished from one-time decision-making tasks, such as buying a house, by the presence of four elements (Brehmer, 1990a, 1992; Edwards, 1962): (1) The tasks require a series of decisions rather than one isolated decision; (2) The decisions are interdependent; (3) The environment changes both autonomously and as a result of decision makers' actions; (4) Decisions are goal-directed and made under time pressure, thereby reducing the decision maker's opportunities to consider and explore options. Given that dynamic decision-making tasks take place in changing environments, research to explain performance in these environments must account for the ability of the decision maker to adapt or learn while performing (Hogarth, 1981). With the emphasis on learning as a means for improving performance, the mechanism by which learning occurs becomes of central concern.

Dynamic Decision Making and Control Theory

Brehmer (1990a, 1992) uses control theory as a framework for analyzing decision makers' goal-directed behavior in dynamic decision-making environments. He hypothesizes that decision makers' ability to learn in dynamic decision-making tasks depends critically on the sophistication of their understanding or model of the environment. In particular, subjects who appear to be using less sophisticated environment models are able to learn to improve their performance only when
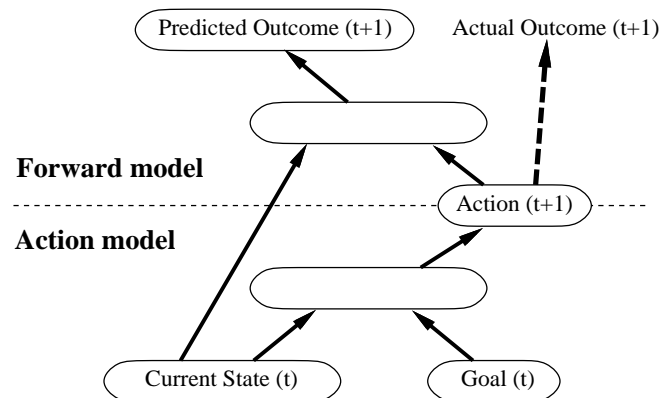

Figure 1: A connectionist framework for control tasks, based on Jordan and Rumelhart (1992). Ovals represent groups of units and arrows represent sets of connections between groups. The unlabeled groups are hidden units that learn internal representations. The dashed arrow is not part of the network, but depicts the physical process whereby actions produce outcomes.

feedback is timely and continuous (Brehmer, 1990a, in press). However, Brehmer fails to specify how decision makers form a model of the environment, how the model of the environment evolves with experience, and how decision makers use this model to learn to improve their performance when interacting with the environment.

Casting dynamic decision making in terms of control theory allows for the transfer of insights from other related domains (Hogarth, 1986). In motor learning, Jordan and Rumelhart (1992; Jordan, 1992, in press) address issues very similar to those addressed by Brehmer. The key to applying their approach to dynamic decision making is to divide the learning problem into two interdependent subproblems: (1) learning how actions affect the environment, and (2) learning what actions to take to achieve specific goals, given an understanding of (1). These two subproblems are solved simultaneously by two connectionist networks joined in series (see Figure 1).

The task of the *action* model is to take as input the current state of the environment and the specific goal to achieve, and to generate as output an action that achieves that goal. This action then leads to an outcome which can be compared with the goal to guide behavior. Unfortunately, when the outcome fails to match the goal—as it generally will until learning is complete—the environment does not provide direct feedback on how to adjust the action so as to improve the corresponding outcome's match to the goal.

Such feedback can, however, be derived from an internal model of the environment, in the form of a *forward* model. This network takes as input the current state of the environment and an action, and generates as output a predicted outcome. This predicted outcome can be compared with the actual outcome to derive an error signal. A gradient-descent procedure, such as back-propagation (Rumelhart, Hinton, & Williams, 1986), can then be used to adjust the parameters (i.e., connection weights) of the forward model to improve its ability to predict the effects of actions on the environment. Notice that learning in the forward model is dependent on the behavior of the action model because it can learn environmental outcomes only over the range of actions actually produced by the action model.

To the extent that the behavior of the forward model approximates that of the environment, it can provide the action model with feedback for learning in the following way. The actual outcome produced by the action is compared with the goal to derive a second error signal. Back-propagation can again be applied to the forward model (without changing its own parameters) to determine how changing the action would change the error. This information corresponds to the error signal that the action model requires to determine how to adjust its parameters so as to reduce the discrepancy between the goal and the actual outcome produced by its action.

Jordan and Rumelhart's (1992) framework provides an explicit formulation of the points left unclear in Brehmer's (1990a, 1992) original assertion that the environment model plays a central role in learning in dynamic decision-making tasks. In Jordan and Rumelhart's formulation, an internal or forward model of environment is formed and revised on the basis of goal-directed interaction with the environment. Furthermore, the importance of this forward model resides in its role of interpreting outcome feedback as the decision maker attempts to learn what actions to take in order to achieve given goals in an evolving context.

A Test Case: The Sugar Production Factory

In order to evaluate Jordan and Rumelhart's (1992) computational framework in the context of dynamic decision making, a version of the model depicted in Figure 1 was implemented to learn a computer-simulated dynamic decision-making task that has received significant attention in the experimental literature, the Sugar Production Factory (Berry & Broadbent, 1984, 1988; Brehmer, 1992; Stanley, Mathews, Russ, & Kotler-Cope, 1989). In one version of the task (the *original learners* condition from Experiment 1 of Stanley et al., 1989), subjects manipulate the workforce of a hypothetical sugar factory to attempt to achieve a particular goal production level. At every time step t, subjects are presented with a display screen depicting the current workforce (measured in hundreds of workers), the current production level (measured in thousands of pounds of sugar), and a graph of all past production levels. Subjects must indicate the workforce for time $t + 1$ and are limited to 12 discrete values ranging from 1 to 12 (representing hundreds of workers). Similarly, the output of the factory is bounded between 1 and 12 thousand tons in discrete steps, and is governed by the following equation (which is unknown to subjects):

$$P(t + 1) = 2W(t + 1) - P(t) + \epsilon \qquad (1)$$

where $P(t + 1)$ represents the new production at time $t + 1$ (in thousands), $W(t + 1)$ is the specified workforce at $t + 1$ (in hundreds), and ϵ is a random error term of -1, 0, or 1. Over a series of such trials within a training set, subjects repeatedly specify a new workforce and observe the resulting production level, attempting to achieve a prespecified goal production.

Stanley et al. (1989) report on the performance of eleven subjects trained on this task in three sessions taking place over three weeks. Each session was divided into twenty sets of 10 trials or time steps during which the subjects attempted to reach and maintain a goal level of 6 thousand tons of sugar production. At the start of each set of trials, initial workforce was always set at 9 hundred and initial production was allowed to vary randomly between 1 and 12 thousand. Subjects were told to try to reach the goal production exactly. However, due to the random element in the underlying system, Stanley et al. scored subject performance as correct if it ranged within ± 1 thousand tons of the goal. In addition, at the end of each set of 10 trials, subjects attempted to write down a set of instructions for yoked naive subjects to follow. The relative initial success of these yoked subjects compared with that of purely naive subjects was taken as a measure of the degree of explicit knowledge developed by the original subjects. The instruction writing also had a direct beneficial impact on the performance of the original subjects.

The Sugar Production Factory task contains all of the elements of more general dynamic decision-making environments, with the exception of time pressure. In this regard, Brehmer (1992) has observed that, although removing time pressure may lead to improved performance, the relative effects of other factors on performance are the same. Furthermore, although the task appears fairly simple, it exhibits complex behaviors that are challenging to subjects (Berry & Broadbent, 1984, 1988; Stanley et al., 1989). In particular, due to the lag term $P(t)$, two separate, interdependent inputs are required at times t and $t + 1$ to reach steady-state production. In addition, also due to the lag term, maintaining steady-state workforce at non-equilibrium values leads to oscillations in performance. Finally, the random element allows the system to change autonomously, forcing subjects to exercise adaptive control. The random element also bounds the expected percentage of trials at goal performance to between 11% (for randomly selected workforce values; Berry & Broadbent, 1984) and 83% (for a perfect model of the system; Stanley et al., 1989).

Implementation of the Model

Jordan and Rumelhart's (1992) framework, depicted in Figure 1, was instantiated in the following way to interactively learn to control the Sugar Production Factory. The goal production value was indicated as a real value on a single goal unit. The current production and the current workforce were each represented as real values on separate input units. The graph of past production values was represented as a series of real values on separate input units, one for each value. All of these inputs were scaled linearly to between 0 and 1. Finally, the hidden layers in both the forward and action models each contained 30 hidden units with sigmoidal output ranging between ± 1. The number of hidden units was established empirically based on series of simulation experiments intended

to determine the minimum hidden units required to learn a slightly more complex version of the task.

As described earlier, the network used two different error signals to train the forward and action models. The predicted outcome generated by the forward model was subtracted from the actual (scaled) production value generated by Equation 1 to produce the error signal for the forward model. The error signal for the action model was generated by subtracting the actual production generated by Equation 1 from the goal level and multiplying the difference by the scale factor.

One training trial with the model occurred as follows. The initial input values, including the goal, were placed on the input units. These then fed forward through the action model hidden layer. A single action unit took a linear weighted sum of the action hidden unit activations, and this sum served as the model's indication of the workforce for the next time period. This workforce value was used in two ways. First, conforming to the bounds stipulated in Stanley et al.'s original experiment, the value was used to determine the next period's production using Equation 1. Second, the unmodified workforce value served as input into the forward model, along with all of the inputs to the action model except the goal. These inputs fed through the forward hidden layer. A single predicted outcome unit computed a linear weighted sum of the forward hidden unit activations, and this sum served as the model's prediction of production for the next period. It is important to note that the forward and action models were trained simultaneously.

The model was trained under two conditions corresponding to different assumptions about the prior knowledge and expectations that subjects bring to the task. In the first condition, corresponding to no knowledge or expectations, the connection weights of both the forward and action models were set to random initial values sampled uniformly between ± 0.5. However, using the same task but a different training regimen, Berry and Broadbent (1984) observed that naive human subjects appear to adopt an initial "direct" strategy of moving workforce in the same direction that they want to move production. To approximate this strategy, in the second training condition, models were pretrained for two sets of ten trials on a system in which production was commensurate to size of workforce without lagged or random error terms.

For both initial conditions, the regimen of training on the Sugar Production Factory task exactly mimicked that of Stanley et al. (1989) for human subjects, as described above, except that no attempt was made to model instruction writing for yoked subjects. In the course of training, back-propagation (Rumelhart et al., 1986) was applied and the weights of both the forward and action models were updated after each trial (with a learning rate of 0.1 and no momentum). To get an accurate estimate of the abilities of the network, 200 instances (with different initial random weights prior to any pretraining) were trained in each experiment.

Comparison of Model and Human Performance

Aggregate Comparison with Stanley et al. (1989). Figure 2 shows a comparison of the average performance of the model under the two different initial conditions (with or without pretraining) and Stanley et al.'s (1989) eleven original learners. Performance is measured based on number of trials correct out of ten using the performance criterion of goal
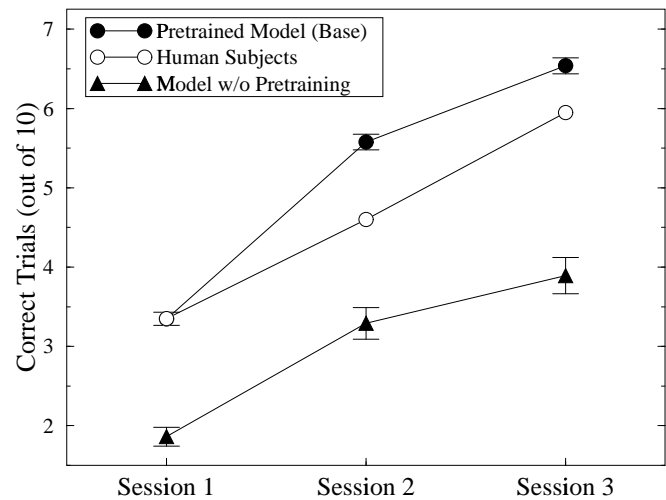

Figure 2: A comparison of the average learning performance across training sessions of Stanley et al.'s (1989) human subjects and models with and without pretraining.

production ± 1 thousand. As is clear in the figure, the performance of the randomly initialized models is far below that of human subjects. This difference is unlikely to be due to explicit knowledge, unavailable to the network, that subjects were able to acquire early on in the task: Stanley et al. (1989) found that the instructions written by the original subjects were useful to yoked naive subjects only near the end of the third training session.

By contrast, the pretrained models perform equivalently to human subjects in the first training session, and actually learn somewhat more quickly than do subjects over the subsequent two sessions. This advantage may be due to the fact that the model is not subject to forgetting during an intervening week between each training session. The findings of the current modeling work suggest that the prior knowledge and expectations that subjects bring to the task are critical in accounting for their ability to learn the task as effectively as they do. Accordingly, the remainder of the paper presents data only from models with pretraining.

Why should pretraining, particularly on a system that differs in important aspects from the Sugar Production Factory, improve performance in learning to perform in the task? Pretraining provides the model with a coherent set of initial parameter estimates describing system performance. Although the initial model parameters do not describe the true system well, the model is systematic in applying them in attempting to control the system. By contrast, models with no pretraining do not have the benefit of a coherent (albeit incorrect) set of parameter estimates describing system performance when starting the Sugar Production Factory task. Thus, their initial attempts to control the system do not show the same systematicity and their learning does not have the advantage of adjusting an already coherent set of parameters.

Single-Subject Comparison across Training Sets. In addition to aggregate data, Stanley et al. (1989) provide two examples of individual subject performance for each set of trials, over the full 60 sets. Figure 3 shows a comparison between the learning performance of one such subject and that
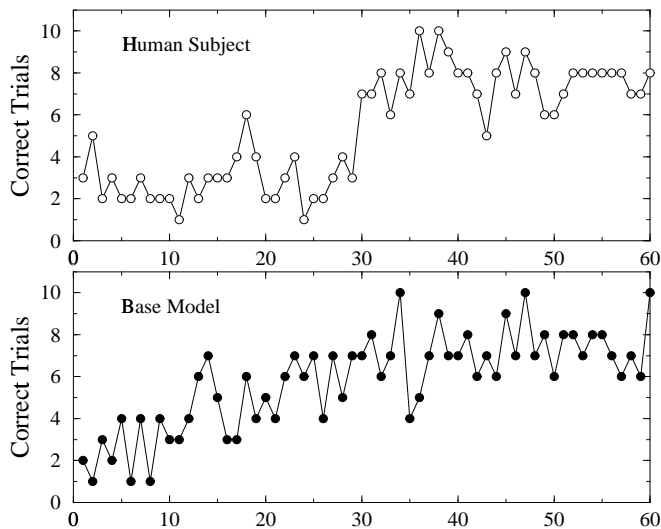
Figure 3: The performance over 60 training sets of a single human subject (Stanley et al., 1989, Subject 10) and a single model.

Figure 4: The sugar production generated by the actions of a single model across the ten trials within training sets 1, 20, and 60.

of an example (pretrained) model over the course of the 60 training sets.

Although there is substantial variability over the course of training, the subject appears to show a breakpoint around training set 30, when the improvement in performance is much more dramatic than at any prior or subsequent time. There is no apparent breakpoint for the model (and other models are broadly similar). One possibility is that the subject (but not the model) acquired an explicit insight into the behavior of the underlying system at the time of the breakpoint. To test this possibility, Stanley et al. (1989) analyzed the performance of the original subjects for breakpoints. They hypothesized that instructions that these subjects wrote for their naive yoked partners immediately after these breakpoints would have a significant positive impact on the naive yoked partners' performance, thereby indicating a link between explicit understanding of the system and performance. However, this hypothesis was not confirmed; instructions written just after breakpoints were no more effective than those written just prior to breakpoints in guiding yoked subjects. Thus, it appears that the breakpoints do not represent a measurable increase in subjects' verbalizeable knowledge about controlling the task. Furthermore, not all subjects exhibited clear breakpoints in learning. Nonetheless, the contrast between subject and model performance suggests that human learning may be more subject to rapid transitions than model learning (but see McClelland, 1994; McClelland & Jenkins, 1990, for examples of staged learning in connectionist networks).

Within-Set Performance for the Model Learner. As mentioned earlier, Berry and Broadbent (1984) found that, at the beginning of training, subjects attempt to increase workforce to increase production and vice versa. Furthermore, they noted that subjects' performance initially shows large oscillations and that these oscillations decrease as the subjects gain experience with the system. Figure 4 shows three detailed sets of trials in which the example model whose overall performance is depicted in Figure 3 attempts to control the system. Like human subjects, the model starts with highly
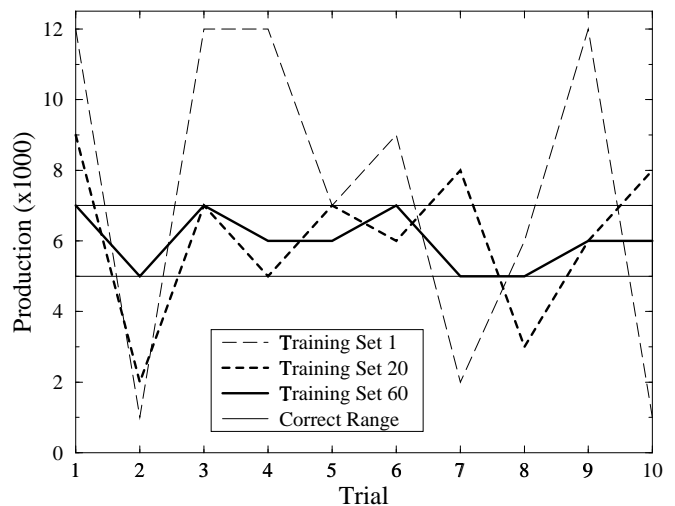
oscillatory performance and reduces those oscillations as it becomes more adept at controlling the system.

The initial over- and under-correction is a hallmark of the model's systematic application of its pretrained conceptualization of the system. Attempting to bring about a change in production by a commensurate change in the workforce has the effect of increasing oscillation in production at non-equilibrium values. As training progresses, the model slowly revises its internal model of the system, as represented in its parameter estimates. By training set 60, the model has overcome its tendency to over- and under-correct.

Forward and Action Model Learning Over Time. The importance of the network's internal model of the system can be clarified by separately examining the time course of learning in the forward and action models. Figure 5 shows the total error across training sets for the forward and action models, averaged over 40 models. Two observations are relevant. First, the difference in squared error between the forward and action models decreases over time. Second, in the early stages of learning, the error for the forward model drops more steeply than that of the action model. These two observations illustrate how the improvement in the model's understanding of the environment precedes and guides its increasing effectiveness in taking action.

Summary of Comparisons with Human Performance. The performance of the model presented here is broadly similar to that of human subjects, given the available data. Although models without pretraining performed more poorly than human subjects, pretrained models outperformed subjects in performance measures aggregated over sets. Pretraining gives the model an edge because it has a coherent model of a system instantiated in its parameters that it applies and revises systematically. Consistent with this observation is the model's conformance to Berry and Broadbent's (1984) observation that human subjects tend to reduce oscillations in performance as they became more experienced. However, unlike the model, some human subjects show breakpoints in their learning performance across training sets, although these do
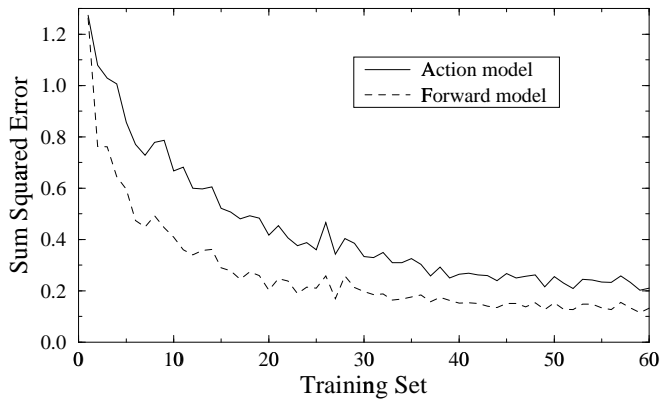
Figure 5: Reduction in error of the forward and action models over training sets.
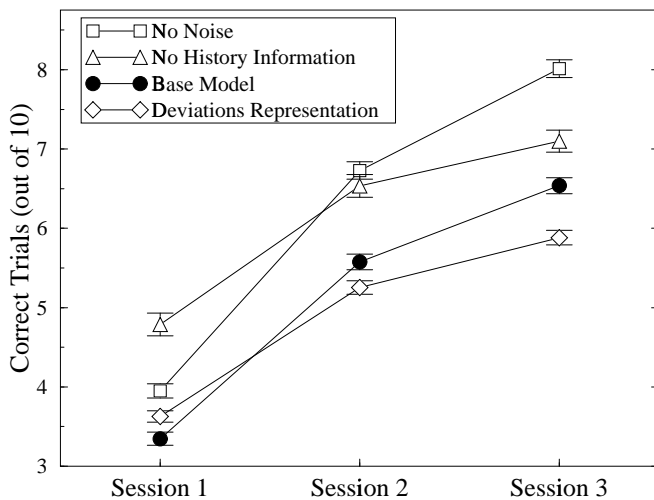

Figure 6: Effects of various manipulations of the task environment on model performance.

not appear to be due to an increase in explicit knowledge about the underlying system. Finally, Figure 5 characterizes the evolution of model performance in terms of improvements in the forward model guiding improvements in the action model for this task.

Effects of Manipulating Task Environment

An important benefit of developing an explicit computational formulation of dynamic decision making is that it provides a platform for evaluating factors that influence the effectiveness of learning in such tasks. In general, many of the relevant factors have not been studied extensively in the existing empirical literature. Nonetheless, we can use the implemented model to generate predictions of how various manipulations will affect the performance of subjects. As a first step, we performed a number of simulation experiments to evaluate how model performance depends on certain aspects of the task representation. The results from these experiments are presented in Figure 6. Empirical studies to evaluate these predictions are currently being planned.

Representing Values as Deviations.   In the first experiment, the model is used to predict how two different representations

of task quantities might affect human performance. In the *deviations* representation, the goal and new workforce which the model sets are represented as deviations (differences) from the current production and current workforce, respectively. In addition, the production history is also presented as deviations from the goal.

As can be seen in Figure 6, model performance in the deviations condition starts out slightly better than base in the first session and slowly diverges over the next two sessions until it is almost a full point below in the third session. The reason for the divergence in performance appears to be as follows. The size of the error term relative to the action the model is trying to modify is larger in the deviations condition than in the base condition. At the beginning of learning, models in both conditions are trying to produce relatively large modifications in workforce (i.e., size of error term is large for both conditions), so the difference in conditions is not apparent. However, later in learning, the modifications that both models are trying to produce in the workforce levels they are learning to set become finer. It is here that the difference in size of the error term relative to the action to be modified becomes significant and affects learning performance.

Similar effects of feedback magnitude have been found in human learning. In a repeated prediction task, Hogarth, McKenzie, Gibbs, and Marquis (1991) found that subject performance was influenced by the absolute scale of the feedback they received. In particular, subjects receiving feedback with low-magnitude variance tended to undercorrect, whereas those receiving feedback with high-magnitude variance tended to overcorrect.

Reducing the Number of Presented Relations.   The second experiment involved manipulating the number of variable relationships with which the model is presented. In particular, a model was trained without presenting the history graph of past production values. As mentioned earlier, this graph is represented in the base model as additional input values. As such, the representation has the effect of providing the model with a greater number of possible relationships between variables to sort through as it attempts to control the system. However, as learning progresses and the model learns which relationships are relevant to performance, the difference in performance between the base model and the one trained with no history graph lessens.

Eliminating Random Variability.   In the final experiment, performance on learning the original system was compared with learning an equivalent deterministic system (i.e., without the random component ϵ in Equation 1). In the original system, the model attempts to adapt to the random element. By definition, this random element cannot be learned, so, as would be expected in Figure 6, performance for the model in the original system is degraded relative to the deterministic system. Additionally, the model's attempts to adapt to the random element appear to be responsible for slowing the rate of learning in the original system. By showing a decrease in the long term rate of adaptation due to the learning mechanism itself, this result conforms with and extends Brehmer's (1990b) observation that random elements in system performance present a limit to human adaptation.

Conclusion

This paper presents a connectionist model that builds on the previous application of control theory to psychological studies of dynamic decision making and a connectionist formulation of motor control. The model provides a broad approximation to existing data on human learning performance in the Sugar Production Factory, an example dynamic decision-making task. In addition, the model makes a number of untested predictions for future empirical work.

This model's approach may be contrasted with alternatives that rely on explicit hypothesis testing or sequences of training trials to initiate learning. Explicit hypothesis testing would imply that improved verbal knowledge of the task would co-occur with improved performance. However, the results of Stanley et al. (1989) indicate that improved verbal knowledge occurs well after improved performance.

Two sets of authors present theories that require sequences of attempts at controlling the system to initiate learning. First, Mitchell and Thrun (1993) present a learner implemented as a neural network that attempts to pick the best action based on its existing model of the environment. This model is updated based on its assessed accuracy in predicting the outcome of a sequence of trials once that sequence has occurred. Second, Stanley et al. (1989) conjecture that performance in the Sugar Production Factory depends on the learner's ability to make analogies between the current situation and prior (successful) sequences of examples. Thus, in this scheme, knowledge can be said to increase every time a successful sequence is encountered and retained. The model proposed here differs fundamentally from these two approaches in that it is able to use information from both successful and unsuccessful single control trials to alter its parameters (connection weights) to reduce the error in its performance. In particular, this property of the model is critical in producing a relatively rapid decrease in production oscillations as training progresses. If implemented to perform the Sugar Production Factory task, it seems unlikely that either Mitchell and Thrun's or Stanley et al.'s approach would produce similarly rapid decreases in oscillations.

Clearly, the model presented here has several limitations. It does not account for meta-strategies such as planning how to learn in the task. It also does not account for how verbalizeable knowledge is acquired during learning. Finally, it does not account for how relevant information presented across multiple time steps might be integrated while learning to perform in dynamic decision-making tasks. Empirical validation of the predictions made so far and this last limitation are the focus of ongoing research. Even with its limitation, the model constitutes one of the first explicit computational formulations of how subjects develop and use an internal model of the environment in learning to perform dynamic decision-making tasks.

Acknowledgments

References

Berry, D. C., & Broadbent, D. E. (1984). On the relationship between task performance and associated verbalizable knowledge. *Quarterly Journal of Experimental Psychology*, *36A*, 209–231.

Berry, D. C., & Broadbent, D. E. (1988). Interactive tasks and the implicit-explicit distinction. *British Journal of Psychology*, *79*, 251–272.

Brehmer, B. (1990a). Strategies in real-time, dynamic decision making. In R. Hogarth (Ed.), *Insights from decision making*. Chicago: University of Chicago Press.

Brehmer, B. (1990b). Variable errors set a limit to adaptation. *Ergonomics*, *33*, 1231–1239.

Brehmer, B. (1992). Dynamic decision making: Human control of complex systems. *Acta Psychologica*, *81*, 211–241.

Brehmer, B. (in press). Feedback delays in complex dynamic decision tasks. In P. Frensch, & J. Funke (Eds.), *Complex problem solving: The European perspective*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Edwards, W. (1962). Dynamic decision theory and probabilistic information processing. *Human Factors*, *4*, 59–73.

Hogarth, R. M. (1981). Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. *Psychological Bulletin*, *90*, 197–217.

Hogarth, R. M. (1986). Generalization in decision research: The role of formal models. *IEEE Transactions on Systems, Man, and Cybernetics*, *16*, 439–449.

Hogarth, R. M., McKenzie, R. M., Gibbs, B. J., & Marquis, M. A. (1991). Learning from feedback: Exactingness and incentives. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(4), 734–752.

Jordan, M. I. (1992). Constrained supervised learning. *Journal of Mathematical Psychology*, *36*, 396–425.

Jordan, M. I. (in press). Computational aspects of motor control and motor learning. In H. Heuer, & S. Keele (Eds.), *Handbook of perception and action: Motor skills*. New York: Academic Press.

Jordan, M. I., & Rumelhart, D. E. (1992). Forward models: Supervised learning with a distal teacher. *Cognitive Science*, *16*(3), 307–354.

McClelland, J. L. (1994). The interaction of nature and nurture in development: A parallel distributed processing perspective. In P. Bertelson, P. Eelen, & G. d'Ydewalle (Eds.), *International perspectives on psychological science, Volume 1: Leading themes* (pp. 57–88). Hillsdale, NJ: Lawrence Erlbaum Associates.

McClelland, J. L., & Jenkins, E. (1990). Nature, nurture, and connections: Implications of connectionist models for cognitive development. In K. VanLehn (Ed.), *Architectures for intelligence* (pp. 41–73). Hillsdale, NJ: Lawrence Erlbaum Associates.

Mitchell, T. M., & Thrun, S. B. (1993). Explanation-based neural network learning for robot control. In S. J. Hanson, J. D. Cowan, & C. L. Giles (Eds.), *Advances in neural information processing systems 5* (pp. 287–294). San Mateo, CA: Morgan Kaufmann.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, *323*(9), 533–536.

Stanley, W. B., Mathews, R. C., Russ, R. R., & Kotler-Cope, S. (1989). Insight without awareness: On the interaction of verbalization, instruction, and practice in a simulated process control task. *Quarterly Journal of Experimental Psychology*, *41A*(3), 553–577.