MANAGEMENT SCIENCE
Vol. 50, No. 10, October 2004, pp. 1366–1378
ISSN 0025-1909 | EISSN 1526-5501 | 04 | 5010 | 1366

INFORMS

DOI 10.1287/mnsc.1040.0271

From T-Mazes to Labyrinths: Learning from Model-Based Feedback

Jerker Denrell
Graduate School of Business, Stanford University, Stanford, California 94305, denrell@gsb.stanford.edu

Christina Fang
Stern School of Business, New York University, New York, New York 10012,
and National University of Singapore, Singapore, cfang@stern.nyu.edu

Daniel A. Levinthal
Department of Management and Economics, The Wharton School, University of Pennsylvania,
Philadelphia, Pennsylvania 19104, levinthal@wharton.upenn.edu

Many organizational actions need not have any immediate or direct payoff consequence but set the stage for subsequent actions that bring the organization toward some actual payoff. Learning in such settings poses the challenge of credit assignment (Minsky 1961), that is, how to assign credit for the overall outcome of a sequence of actions to each of the antecedent actions. To explore the process of learning in such contexts, we create a formal model in which the actors develop a mental model of the value of stage-setting actions as a complex problem-solving task is repeated. Partial knowledge, either of particular states in the problem space or inefficient and circuitous routines through the space, is shown to be quite valuable. Because of the interdependence of intelligent action when a sequence of actions must be identified, however, organizational knowledge is relatively fragile. As a consequence, while turnover may stimulate search and have largely benign implications in less interdependent task settings, it is very destructive of the organization's near-term performance when the learning problem requires a complementarity among the actors' knowledge.

Key words: organizational learning; credit assignment; organizational routines; task interdependency; reinforcement learning
History: Accepted by Linda Argote, organization performance, strategy, and design; received December 4, 2001. This paper was with the authors 9 months for 2 revisions.

1. Introduction

Organization literature has explored both the power and the limits of experiential learning. Although learning stemming from prior experience can be associated with performance improvement (Yelle 1979, Argote 1999), it can also pose difficult challenges of inference. The link between current actions and observed outcomes may be conflated by noise (Lant 1994) and interaction with other learners (Lounamaa and March 1987, Levinthal 1997). The link between actions and outcomes can also be obscured by temporal interdependencies, such as competency traps (Levinthal and March 1981, Levitt and March 1988) and delays in feedback (Sterman 1989a, Sastry 1997).

Less explored is the challenge of learning in settings in which outcomes can only be observed after a series of actions have been performed. Many organizational actions do not result in any immediate or direct payoff consequence but set the stage for subsequent actions that bring the organization toward some actual payoff. Consider, for instance, most of the activities in organizational routines. As emphasized by Nelson and Winter (1982) and Cohen and Bacdayan (1994), routines consist of patterned sequences of activities involving multiple individuals. Most of the activities do not have any immediate payoff consequences but only trigger the activities of other individuals.

This delay in realizing feedback complicates search behavior because only at the end of a long sequence of actions is there some discernable payoff. More generally, this challenge is known as the credit assignment problem (Samuel 1959, Holland et al. 1986, Axelrod and Cohen 1999, Levinthal 2000)—how should one assign the credit arising from the overall sequence of actions to each of the antecedent actions?

Existing models of organizational learning typically model learning as a behavioral change in response to immediate performance feedback. In settings in which outcome feedback is not immediately available, these models suggest that learning about the value of alternative actions in such contexts would be very difficult as the basis for reinforcement learning would be absent. Nevertheless, organizations do seem to be able to develop mental maps that make it possible to impute the value of intermediary actions. Several authors (Sterman 1989b, Brehmer 1995, Gibson et al. 1997) have argued that an understanding of how such mental models develop is essential for understanding dynamic decision-making tasks. Capturing such effects, however, requires a greater emphasis on the cognitive aspects of learning (Glynn et al. 1994, Walsh 1995).

We approach this modeling challenge by extending the standard model of reinforcement learning featured in the organizational learning literature (cf. Lave and March 1975) to let actors' existing mental model of the world be a basis for reinforcement. As a result, actions are reinforced not only when they result in payoff immediately but also when they lead to states that are "believed" to be valuable as stepping stones toward some ultimate goal. In particular, we model the challenge of credit assignment by building on an analytical structure known as temporal differencing, which has existed for some time in the computer science literature (Samuel 1959) and has received recent renewed interest (Holland et al. 1986, Bertsekas and Tsitsiklis 1996, Sutton and Barto 1998).

We examine a task structure in which the outcome payoff requires that a sequence of actions be completed. Agents develop a mental model of the value of intermediate or stage-setting actions by gradually incorporating their experience both across and within problem-solving efforts. We show that the use of credit assignment results in the actors' mental maps quickly assigning positive value to states close to the solution. We also find that organizations can enhance their rate of performance improvement by assigning credit more aggressively to a longer sequence of antecedent actions. However, by doing so, organizations risk making spurious associations between the ultimate outcome and prior actions. As a result, a variant of the exploration/exploitation trade-off (March 1991) emerges.

Our analysis highlights several features of organizational learning that have not been emphasized in previous research. In particular, it illustrates the performance benefits of even partial knowledge as well as the high costs of disrupting such partial knowledge in order to search for further performance improvements. Partial information, whether of an isolated node of knowledge in the problem space or a circuitous and relatively inefficient routine, may be of tremendous value in guiding searches and avoiding long, unproductive, random walks. Thus, the existence of *some* routine, even limited and inefficient, can greatly enhance the speed of discovery and performance improvements.

The opposite side of the same coin is the high immediate cost of ignoring such knowledge to search for further improvement. In context with sequential interdependency (Thompson 1967), the value of knowledge of some intermediate stage-setting action on the part of one individual is highly contingent on the knowledge and beliefs of others with whom that individual interacts. The expertise of one actor with respect to what constitutes appropriate action is more or less useful if it triggers behavior of an adjacent actor who, in turn, also has a useful point of view as to what constitutes appropriate action. This feature of organizational learning casts a very different light on the role of turnover in organizational search processes. Contrary to the benign role of turnover in March's (1991) model, we demonstrate that turnover of personnel is likely to result in performance decline because it disrupts the knowledge of organizational actors. We also find that the turnover of personnel more proximate to the solution is particularly problematic.

2. Complications of Experiential Learning and the Credit Assignment Problem

While learning processes can lead to improvement over time (Yelle 1979, Argote 1999), explicit models of learning have revealed several important limitations of organizational learning (Huber 1991, Lant 1994, Miner and Mezias 1996). Based on the problem contexts explored and the complications of learning investigated, three strands of work can be distinguished.

First, there is a set of models that investigates the complications of experiential learning in noisy, ambiguous, and changing environments (Lant 1994). In this tradition, organizational learning is conceptualized as an incremental, myopic process in which actions that appear successful relative to an adaptive aspiration level for performance are repeated, and actions that appear unsuccessful are changed (or the propensity to engage in them is reduced) (Cyert and March 1963, Levinthal and March 1981, Lant 1992, Greve 1998). As a result of noisy signals of performance and path dependency introduced by competence multipliers, experiential learning can produce superstitious learning (Lave and March 1975, Levinthal and March 1981, Levitt and March 1988) and converge to inferior alternatives (Levinthal and March 1981; Herriott et al. 1985; Lant and Mezias 1990, 1992; Mezias and Glynn 1993; Sastry 1997; Denrell and March 2001; Repenning and Sterman 2002).

A second strand of work examines the complications of experiential learning introduced by interdependency and complexity as well as mutual adaptation within organizations. Organizational learning typically has been conceived of as a process of local search, modeled using hill-climbing algorithms. Studies have examined the properties of such organizational search processes in rugged landscapes (Levinthal 1997, McKelvey 1999, Rivkin 2000) as well as in situations of mutual learning by several interacting subunits (Lounamaa and March 1987, Carley 1992, Lin and Carley 1997, Chang and Harrington 1998, Rivkin and Siggelkow 2003). Due to the multiplicity of local optima in such contexts, experiential learning is not guaranteed to converge to a global optimum. Rather, learning in such contexts is a path-dependent process in which organizations facing a common environment are likely to end up at different local optima (Levinthal 1997, Gavetti and Levinthal 2000, Rivkin 2000).

Finally, there is a strand of work examining the complications of experiential learning processes in situations with temporal delays and nonlinearities (Sterman 1989a, Lomi et al. 1997, Sastry 1997, Sterman 2000). The prototypical problem investigated in this literature can be illustrated by the classic "beer game" (Forrester 1961, Sterman 1989a),[1] in which actions have immediate as well as complicated delayed effects. Due to the potentially misleading character of immediate performance feedback in such contexts (Sterman 1989a, b; Sastry 1997), myopic learning processes based on immediate performance feedback typically lead to suboptimal behavior (Sastry 1997) and costly oscillations (Sterman 1989a, b; Lomi et al. 1997).

This prior work, however, has generally examined contexts in which immediate outcome feedback is available, although possibly misleading. In this sense, these settings are analogous to that of the classical T-maze choice problem in which an actor chooses which of two branches to go down and receives, with some probability, the reward associated with that choice. However, in many situations, actions are not followed by immediate feedback. Rather, outcome feedback may only be available after a sequence of actions has been performed. The problem of choosing a sequence of actions is more like navigating in a labyrinth in which an action takes one to another decision context rather than to some ultimate end state.

To illustrate the difficulties of learning in the absence of information about immediate outcomes, consider the task featured in the experiments by Cohen and Bacdayan (1994) on the development of routines. The goal in the experiment was to move a specific card into a target area through an exchange of cards between two players. Not all exchanges were permitted, however. As a result, to achieve the goal, subjects sometimes had to take actions that would move the card away from the target area. Learning from experience in such situations is challenging, because even when the goal has been achieved, it is seldom obvious whether specific moves were good or bad. In particular, evaluating individual moves often requires recognition of their long-term implications. The long-term implications depend on subsequent moves, however. Thus, developing a sophisticated understanding of the values of different moves is difficult even with repeated experience.

Such learning challenges are known in the artificial intelligence literature as the credit assignment problem (Minsky 1961, Axelrod and Cohen 1999). How do we assign credit for the overall outcome of a learning system to each of its individual actions, possibly taken several steps before the outcomes could be observed? Intuitively, the credit assignment problem seems highly complex. In the context of the above experiment, for example, it would seem that we would have to examine numerous paths to the solution and keep track of all moves to evaluate just a single move. Given the vast combinatorial possibilities of moves, such a process would require an enormous number of trials.

Despite this apparent difficulty, Samuel (1959, 1967), in his pioneering work on credit assignment, demonstrated in an important early application to the game of checkers: Credit assignment can be done incrementally, on the basis of immediate experience (see also Holland 1998, Ch. 4). The intuition behind Samuel's model is that interim predictions based on a player's own estimates can inform action as well as direct feedback from the environment. A player is modeled as making predictions along the way and constantly adjusting them based on the available information he or she collects en route. In time, the player's estimates may become a closer approximation to the true mapping from actions to eventual outcomes.

This important contribution of Samuel, developed further by Sutton and Barto (1981, 1998), Holland et al. (1986), and Bertsekas and Tsitsiklis (1996), poses the idea that the actor's own mental model of the environment can be used to provide interim feedback regarding the value of actions in lieu of feedback from the environment. In particular, actions that do not result in immediate payoff are nonetheless reinforced if they lead to states that, according to the actor's current mental model, are believed to be valuable. In this sense, the agent's mental model can provide feedback to guide behavior even though immediate outcome feedback is absent. Below we outline a model

[1] In the "beer game," participants are faced with the task of maximizing profits in a multistage supply chain. Feedback is available, although possibly misleading, because exogenous changes in demand are confounded by delays caused by inventory buildup and depletion.

based on this structure and examine its implications for organizational learning.

3. Modeling Credit Assignment

Learning Algorithm

Learning in the absence of immediate feedback requires that actors in the organization develop beliefs not only about the immediate payoffs of actions but also about the value of actions as potential stepping stones for reaching valuable states. We represent these beliefs in a very simple, stylized manner by an action-value function $Q(s, a)$ (Watkins 1989). Such a function represents the beliefs of actors in the organization about the immediate reward for the organization of taking action a in state s, as well as its stage-setting implications for the organization—does it lead to a more or less promising state from which greater reward can be earned in the future?[2] Abstracting from any incentive problems, we model each actor as taking actions that are perceived to have the highest value for the organization (i.e., that yield the maximum value of $Q(s, a)$ for a given state s). However, this optimization covers the actor's mental model of the world, not the true payoff structure. As Camerer (1997) has noted, the behavioral inaccuracy of rational choice models may have less to do with the inappropriateness of the algorithm—that is, choose the best alternative—than with the assumption that actors apply that algorithm to the actual representation. In that spirit, we assume that based on their Q function, actors choose their best action, but we make no presumption that their Q function corresponds to the actual payoffs.

Indeed, the critical question for our analysis is how these belief structures emerge over time. Following Samuel (1959), we assume that actors update their Q function incrementally, making use of information provided by deviations from predictions based on the actor's current belief structure. The particular structure that we explore is that of temporal differencing (Kaelbling 1993, Sutton and Barto 1998). To illustrate the algorithm, suppose that an actor in the organization has carried out action a in state s. Taking this action, the organization arrives at a new state s'. This new state may or may not provide some immediate reward, which we will term R. Independent of the presence of any immediate reward, this new state s' is now a launching point for subsequent action taken by the same or possibly a new actor. After landing in state s', the actor then examines what is known about the best state-action pair $Q(s', a')$ available in that state. This estimate of the best state-action pair $Q(s', a')$, together with any possible instantaneous payoff, constitutes the input to the agent's revised prediction of the value of the state-action pair that he or she just carried out.

The exact updating of the old value function is based on the difference between the belief about the value of the new state to which the prior state-action pair led and the prior beliefs about that state-action pair, subject to a learning-rate parameter α and a discount factor γ. This specification is drawn directly from Sutton and Barto (1998):

$$Q(s, a) \leftarrow (1-\alpha)Q(s, a) + \alpha\{R + \gamma Q(s', a')\}. \quad (1)$$

The parameter γ weights the importance of the value of future returns from actions taken starting from s' and in that sense acts like a discount rate. However, it is important to note that γ is not discounting some tangible payoff but is influencing the degree to which the prior state-action pair, $Q(s, a)$, gets "credit" for the position (i.e., state s') that it has identified.

As long as γ is positive, the interim prediction $Q(s', a')$ serves as a substitute for, or at least a supplement to, the immediate payoff R in the updating algorithm above. Even if the immediate payoff is zero, the value of the prior state-action pair $Q(s, a)$ can still be augmented so long as $Q(s', a')$ is nonzero. However, if γ is set to zero, there is no updating based on the interim feedback provided by the mental model; rather, in such cases, updating only occurs when outcome performance feedback is obtained. Therefore, with $\gamma = 0$, the updating algorithm in (1) reduces to a standard model of reinforcement learning, with updating based only on immediate outcome feedback.

In this sense, our model is a natural extension of standard reinforcement learning models. The reinforcement becomes the actors' mental model of value rather than an actual payoff. Following this approach, an agent positively updates his or her belief about the value of performing an action that led to a state from which the agent or colleague believes that it is easier for the organization to solve the problem. As a result, even in the absence of immediate performance feedback, it is possible to develop beliefs about antecedent actions. The basic structure of reinforcement learning is preserved, while the basis of reinforcement is extended from immediate outcome feedback to include feedback about the value of realized states based on the actors' mental models of the value of these states.

To illustrate how such model-based feedback learning would operate in an organization, consider how an organization would learn an effective sequence of

[2] This is analogous to the logic of dynamic programming (Bellman 1957), in which the value of an action is the immediate payoff plus the maximum payoff from subsequent actions given that resulting state. A critical difference between the two processes is that dynamic programming pushes through the whole "tree" of possible action to the end states. In many applications, such a process is not computationally feasible, given the enormous branching of possible actions and states, let alone behaviorally realistic.

actions for producing a complex product consisting of several parts. An important determinant of productivity in such production processes is the order in which different parts are produced and their timing. Unfortunate choices of order and timing will cause delays and large inventories and thereby reduce performance. Initially, the actors involved in this production process may have little information about the value of alternative actions at different stages. Thus, initial efforts have to be based on guesswork.

Due to the large number of possible permutations, however, it will be difficult to try out even a small fraction of all possible sequences to learn how they work. Nevertheless, actors can learn to improve their predictions about how valuable different actions are at different stages. For example, the actor responsible for the last stage of production may notice that the end result is superior if the product arrives with some parts still to be assembled and may communicate this information to the actor responsible for the previous stage. Based on this information, the actor responsible for this stage may be better able to identify how various designs that are presented to him or her will contribute to the end result. Such information, in turn, provides the basis for the next individual to evaluate the products presented to him or her. Through this process, the actors may gradually build up a more complete understanding of the value of alternative actions, without the need for communication among all individuals or for an exhaustive search of all permutations.

Task Structure

To focus on the essentials of the credit assignment problem, we model a task structure in which there is a positive reward in only a single state and the reward is zero for all other states. Such a task environment represents a demanding case for learning because it provides only one instance of immediate performance outcome feedback to guide the search process, with no rewards for all state-action pairs except the solution state. A visual representation of the payoff surface is a flat landscape with a single spike, corresponding to a unique combination of policy choices (Bruderer and Singh 1996). However, state-action pairs with no immediate payoff are not valueless, as there exists a sequence of actions starting from every state that eventually leads to the solution. Hence, the key to intelligence is learning the value of states as stepping stones to the eventual goal. Learning to recognize such positional value poses a significant challenge in a world of no immediate feedback; yet it is precisely this type of task environment that highlights the importance of credit assignment.[3]

We represent a state s as an N-element binary string, in which each element can take on the value of 0 or 1. In subsequent analysis, we set N equal to 10 elements. As a result, there are 2^{10} possible configurations, or 1,024 possible states. Furthermore, there are $N+1$ possible actions that can be taken from any given state, because an agent can choose to stay where he or she is by keeping the original configuration intact, as well as shifting one of the N parameter values. Each element of the 10-element string can be seen as corresponding to one possible action. The agent can change, for instance, the fifth element from 0 to 1 or vice versa. As such, the state-action space in our problem context can be described as a table with 1,024 rows and 11 columns, resulting in 11,264 cells. Only when all the elements in a state match those of the solution state can a solution be found and a positive reward earned. In subsequent analysis, we set the solution to be in the location in which all elements take on the value 1. Given the random starting position of the agents, the location of the solution is arbitrary.

In this setting, the only real payoff occurs when the organization reaches the goal state. All other states only have an imputed value; that is, their value is based on beliefs about whether reaching these states facilitates the realization of the desired goal state. It is possible, however, to distinguish between the perceived value of states and their "objective" value if the agent were to follow an optimal policy. In particular, if the discount factor is 0.9, a state two steps from the goal state has an "objective" value, in this sense, of $(0.9)^2$ times the payoff associated with the goal state.

The organization's task is to identify a path from its random starting position in the landscape to the solution state. In each time period, the agent may move from one state-action pair to one of its N possible one-step neighbors, or the agent may remain at the current location. However, in contrast to other models of local search (cf. Levinthal 1997), the value the agent places on these neighboring points may also reflect their value as stepping stones to the ultimate solution. Thus, while the examination of possible actions is local, the value the agent places on these actions may reflect their global properties.

While for our analysis it is useful to label the points in this N-dimensional space in a manner such that there is a well-defined ordering among points as to their proximity to the solution, decision-making agents do not have access to a labeling scheme

[3] In future work, it would be interesting to explore problem settings in which all states offer both real payoffs and positional value. Search in a rugged landscape (Kauffman 1993, Levinthal 1997) would provide such a setting. Among the complications such an analysis presents are issues of discounting payoffs and stopping rules for the search process.

or problem representation with such a structure. In particular, actors cannot know if a choice that shifts a value from a 0 for a given dimension of the problem to a 1 moves the organization closer to the solution.

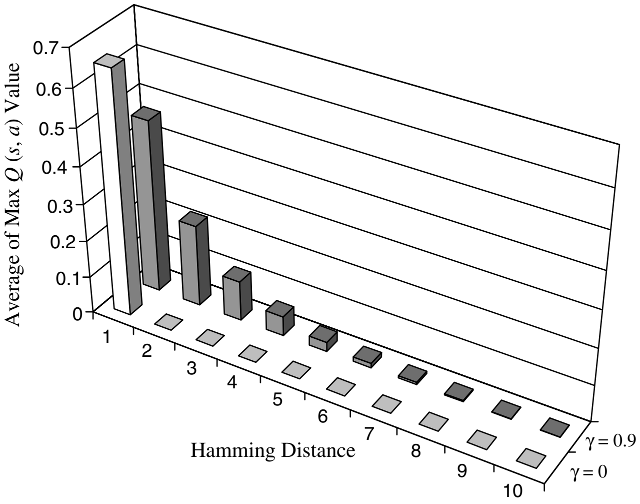
4. Analysis and Results

Consider an organization trying to find a sequence of moves that leads to a solution in the above task structure. First, the agents have some initial guess of what the state-action space looks like. In our analysis, we assume that the actors have a flat initial belief of zero for all state-action pairs (i.e., $Q(s, a) = 0$ for all a and s). The search process is then started at a randomly assigned state s. The actor responsible for this state evaluates the values of all available actions from this state and identifies the action corresponding to the highest value of $Q(s, a)$. If there is more than one action that has this value, then the agent chooses randomly among them. As a result, in the initial period, with all actions valued at zero, the action will be randomly chosen.

Once the solution is found, the organization is restarted in a new, randomly assigned state in a second round of search. However, the actors do not begin fully anew but maintain their updated $Q(s, a)$ functions as a guide for action. As a result, in this subsequent round of search, positive updating of beliefs will occur not only when the solution is found but also whenever an actor reaches a positively valued $Q(s', a')$ identified in an earlier search effort. In this way, with each successive round of searches, more and more valuable state-action pairs are discovered and positively updated. Gradually, the $Q(s, a)$ function better approximates the positional values of the vast set of state-action pairs.

The results in the following analysis are based on the average behavior over 1,000 independent histories of search, where each "history" commences with a new set of beliefs and comprises 100 episodes of problem solving. The learning rate α is set at 0.2, while the value of γ is varied to illustrate the role of credit assignment.[4]

Figure 1 Staircase-Like Mental Models (as of the End of Episode 100)



Emergence of Mental Models

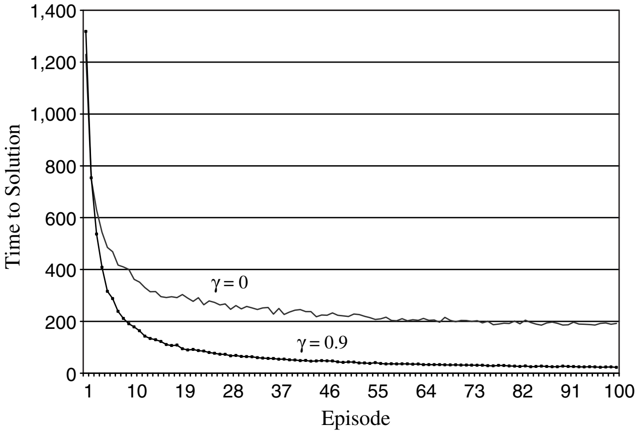
In our analysis, beliefs—instead of being assumed intelligent from the start—are seen as emerging gradually out of a sea of ignorance. Initial beliefs are set to zero across all possible actions for any given state. Over time, however, actors increasingly learn to discriminate among state-action pairs of different positional values. In particular, agents learn to discriminate between the values of states more or less proximate to the solution. States closer to the solution are likely to lead to the solution quickly, thus they are more valuable than states farther away. In Figure 1, we plot the average of the maximum $Q(s, a)$ value of states at various hamming distances away from the solution.[5] The objective ordering of states is indeed reflected in the value placed by actors on the actions in these states.

However, in the absence of credit assignment (when $\gamma = 0$), as seen in Figure 1, the resulting mental model does not resemble a staircase at all. Only states with hamming distance of 1 have positive $Q(s, a)$ values, while the valuation of every other state remains at 0. In this sense, a standard reinforcement model in this problem context is only capable of providing guidance for states in the immediate neighborhood of the goal state. That is because in the absence of credit assignment, updating will only be based on the immediate reward, which is zero for any action-state pair two steps or more from the solution. However, in cases where $\gamma > 0$, even if the immediate payoff is 0, the value of the prior state-action pair $Q(s, a)$ can still be augmented as long as the positional value of the new state, $Q(s', a')$, is positive.

[4] The magnitude of either α or γ does not affect the results as long as those values are positive. Given the specification of priors that all values of $Q(s, a)$ are initially set at zero, any positive updating of the value of a given action in a particular state will result in that action being chosen upon subsequent visits to that same state. Thus, the magnitude of α or γ would matter only if choice were not based on the maximum $Q(s, a)$ but on some other, perhaps less "greedy," choice algorithm. Such modifications could provide an interesting extension, but the current structure focuses on the core issue of the contrast of reinforcement learning based solely on external reward versus reinforcement learning supplemented by reinforcement based on perceived valuation of states.

[5] Hamming distance is a measure of distance in an N-dimensional space. It is simply the number of elements in a string that are different from one another. For instance, the state 0110001111 is a hamming distance of 3 from 1000001111.

Figure 2 Learning Curves



To investigate the performance implication of the different learning strategies, we compute the average number of periods needed to reach the solution for 1,000 agents. Time to solution is a good measure of performance given the task structure, as it captures the idea that more intelligent beliefs should result in a more efficient search, thereby reducing the amount of time needed to find the solution. We examine the problem-solving task 100 times, where there is a random starting point for each iteration, or episode, of the search process. In Figure 2, we plot time to solution as a function of episode, when the number of episodes approximates the amount of experience an organization has with the problem context.

As illustrated in Figure 2, for both γ values of 0 and 0.9, performance, as measured by time to solution, is an exponential function of cumulative experience. As agents get to perform more of the same repetitive task, they can find the solution more and more quickly. However, Figure 2 also demonstrates that the performance implications of simple reinforcement learning and credit assignment are very different. At the end of 100 episodes, the time to solution for agents using credit assignment ($\gamma = 0.9$) is down to around 22 periods, whereas the time to solution associated with agents engaged in simple reinforcement learning ($\gamma = 0$) hovers around 192—a difference in performance of nearly a factor of 10. Thus, the better mental model produced by credit assignment indeed translates into better performance.

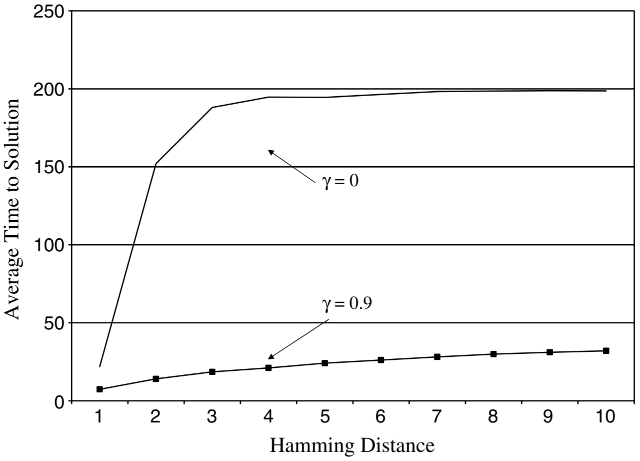
Furthermore, the value of credit assignment is not limited to situations in which a single problem-solving task is repeated numerous times. In fact, only a few repetitions are needed to produce substantial benefits in terms of improved performance. To illustrate this, consider the contrast in the decline in time to solution during Episodes 5–20 for the case with and without credit assignment. As demonstrated in Figure 2, credit assignment leads to a substantial decline in the time to solution during this interval. However, without credit assignment, there is only moderate improvement.

To further illustrate the power of credit assignment, we carry out another experiment, where, as before, the search process is initiated randomly in the first 99 episodes. However, in the last episode, we start the search at a fixed hamming distance from the solution. This eliminates one source of differential performance: Some search efforts may require few periods not because of better mental models but purely as a result of a more favorable starting point closer to the solution. In Figure 3, we plot the average time to solution as a function of the hamming distance at which we start the search in the last episode.

The link between the mental model and performance is immediately clear. In the case in which $\gamma = 0$, which corresponds to a standard reinforcement learning model, if we start the search at a hamming distance of 2, a region just outside the organization's region of information, the average time to solution shoots up dramatically. Only states located one hamming distance away are updated positively, thus the region of information is quite limited. In the vast majority of states, the $Q(s, a)$ values remain what they were initially, and any resulting action can only be uninformed and therefore random.

Contrast this with the case of a positive γ, in which credit assignment is at work. Here, a staircase-like mental model quickly emerges and serves as a guide for intelligent action over a much broader portion of the problem space. The presence of nonzero $Q(s, a)$ provides information as to the value of different state-action pairs. In particular, higher $Q(s, a)$ effectively provides a gradient in the search process. As such, even if the search starts at the maximum hamming distance, actors need not be completely lost. The search is quickly made purposive and takes an average of 32 periods, whereas in the case of no credit

Figure 3 Performance as a Function of Starting Hamming Distance (Based on Mental Models as of the End of Episode 100)



assignment the time to solution is nearly 200 periods. With credit assignment, as the hamming distance at which the search is started increases, there is only a linear increase in the amount of time to solution.

Finally, it should be noted that these plots of average performance mask considerable variation in performance across "histories," or runs of the model. Our analysis shows that most of this variation is due to initial unguided random search. Once an organization finds a routine, however, the routine tends to be efficient, in the sense that each step brings the organization closer to the solution state. In particular, we find that 92% of all informed moves reduce hamming distance. The performance improvement we observed above is the result of gradually extending such routines to states farther away from the solution. Even after 100 episodes, however, some initial periods of random search cannot be avoided. Thus, in the task structure we examine, learning results in a few clues rather than comprehensive causal knowledge. Such clues, however, dramatically improve average performance.

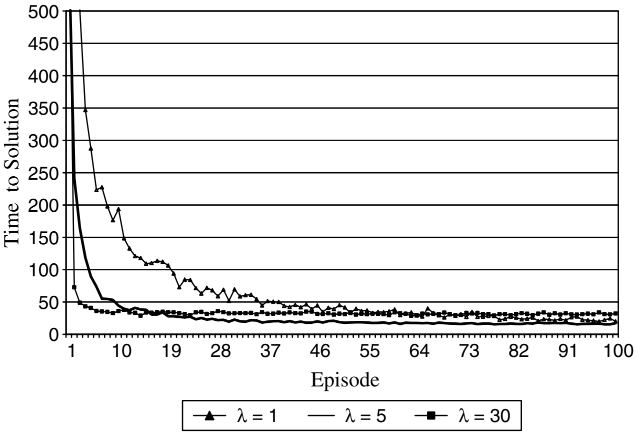
Varying Extensiveness of Credit Assignment

The analysis to this point would seem to suggest that a more aggressive use of prior experience could further enhance learning. After all, after a solution path is identified, why wouldn't one want to replicate that path and codify that pattern action sequence as a routine? However, the average time to solution for the first episode is more than 1,000 periods. This is not a journey that one would want to replicate, because most of the actions taken did not direct the agent toward the solution. Consistent with the notion of superstitious learning (Levitt and March 1988), there is an association between certain actions and the ultimate successful outcome—but an association that in most cases is by chance and need not be causal. In contrast, a structure of one-step credit assignment will tend to give credit only in cases where the action is in fact helpful to the identification of the solution, as only states next to the solution state and states next to these states are given credit.

Nevertheless, the modest degree of backward credit assignment in the case of one-step temporal differencing may underutilize experience. It is probably the case that the action that was two steps away from the reward-generating action is also a useful step and should be rewarded. More generally, we observe a tension between over- and underutilization of experience. Extensive utilization of experience leads to superstitious beliefs; at the same time, underutilization of experience reduces the speed with which intelligence evolves.

The effects of more extensive credit assignment can be investigated using a more general version of our basic model structure. Consider a parameter λ that specifies the number of preceding state-action pairs to which credit is assigned. The basic updating mechanism specified in Expression (1) is maintained, but the updating process is applied to the proceeding λ state-action pairs. As such, values are cascaded back not just to the antecedent action, but to a connected path of length λ. Figure 4 contrasts performance with varying λ values of 1 (the setting in the prior analyses), 5, and 30.

Figure 4 Learning Curves for Different λ Values



As Figure 4 shows, if credit is assigned 30 steps back, the average time to the solution quickly falls after a few episodes of search. However, the time to solution then levels out. More modest credit assignment, where credit is only assigned one step back, eventually produces a lower average time to solution. Thus, if the learner is patient, less extensive credit assignment is to be preferred. On the other hand, if the learner is impatient and quickly wants to learn some way to the solution, it is preferable to assign credit to many preceding actions. Such learning, however, will be mostly superstitious. In fact, if credit is assigned 30 actions back, the organization is making a routine of what was largely a random walk to the solution.

This result provides another illustration of the classical trade-off between fast and slow learning (Levinthal and March 1981, Lant 1994). More specifically, in the context of our model, the results illustrate the trade-off between the efficiency of learned routines and the extensiveness of routines. By assigning credit many moves back, the organization is developing a point of view regarding appropriate actions at many states, including states distant from the solution state. As a result, it is more likely that such an organization will encounter a state that is part of some routine. However, the routines will be less efficient, in the sense that a larger proportion of moves will be hamming distance increasing. While the above results show that less extensive credit assignment, i.e., slower learning, will eventually result in higher performance, they also illustrate the early advantages of developing extensive routines, even if they are inefficient and to some degree superstitious.

In short, when organizations face the challenge of learning an effective sequence of actions, i.e., a routine, recognition of the positional value of *some* states without immediate reward is crucial for high performance. Once an actor finds him- or herself at such a state, an established routine exists for traveling to the solution state. As demonstrated above, even a small amount of knowledge or an inefficient routine may provide large performance advantages by cutting down the search space and thus reduce an otherwise lengthy random search process.

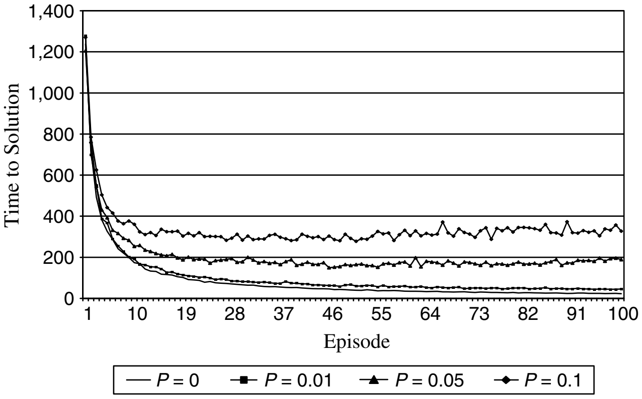
Effect of Turnover

The above analysis of credit assignment in learning and the benefits of following some routine suggests an important but underemphasized characteristic of learning in an organizational context. Whenever multiple actors in organizations face the challenge of learning a sequence of actions—a routine—then the value of information is embedded in the organizational context. An action that takes one closer to the solution may not be helpful if the action leads to a node in the problem space for which the organization has no knowledge. As a result, there is a complementarity to knowledge. The expertise of one actor of what constitutes appropriate action is more or less useful if it triggers behavior of an adjacent actor who, in turn, also has a useful point of view as to what constitutes appropriate action. This social quality of knowledge has important implications for ideas about the benefits of combining new and old knowledge.

In particular, consider the value of turnover as a mechanism to sustain search in an adaptive organization. Empirical work on the impact of turnover on learning processes has provided mixed results. Some work, such as Argote et al. (1995), suggests that turnover is disruptive to learning processes; in contrast, other work, particularly work on innovation processes (Katz 1982), finds that a moderate level of turnover can enhance organizational performance. March's (1991) model of exploration and exploitation finds an important role for turnover as a sustained source of variety that leads to a persistent level of search and avoids the liability of premature lock-in to a particular belief.

A careful examination of the task environment studied in these different research efforts may help explain their divergent findings. In March's (1991) model of the development of an organizational code that is a result of learning from actors in the organization, performance is specified as an additive function of the distance between the code and reality for each dimension of the code. As a result, insight on the part of one actor may, in part, be compensatory for ignorance of another actor. Thus, the task environment is what Thompson (1967) terms pooled interdependence. In contrast, as emphasized above, where organizations must learn a sequence of actions, there will be sequential interdependence. The intelligence of an action then depends not only on whether the action represents movement towards the solution, but also on whether the individual who now bears responsibility for subsequent action is informed as to what might constitute useful behavior. As a result, bringing in uninformed individuals is more likely to be detrimental. The detrimental effects of turnover in our model are illustrated in Figure 5, which plots the performance consequences of different degrees of turnover.[6] Turnover is here defined as setting $Q(s, a)$ to zero for those actors who are replaced. Thus, the new individuals are assumed to have no beliefs as to what constitutes appropriate action. As in March (1991), there is a fixed probability of turnover in each episode. The results regarding the impact on turnover on learning, however, contrast with March (1991, Figure 4), with zero turnover dominating a positive level of turnover. This result also holds if we assume that turnover brings in new individuals with different knowledge, that is, if $Q(s, a)$ is set to some new random value rather than to zero.[7]

Figure 5 Effect of Turnover as a Function of the Probability of Turnover



Note. $\lambda = 1$; each state experiences turnover according to probability P in each period.

[6] Carley (1992) also examines the effect of turnover on organizational learning. While the primary focus of Carley (1992) is on the effect of hierarchical versus team structures, her results on the impact of turnover are consistent with our own. Turnover is shown to have little impact on organizational effectiveness in the setting of a decomposable task structure; however, in a nondecomposable task environment, turnover is shown to be disruptive to organizational learning.

[7] In this case, the turnover beliefs are determined the following way: Draw a random number to see which action gets the

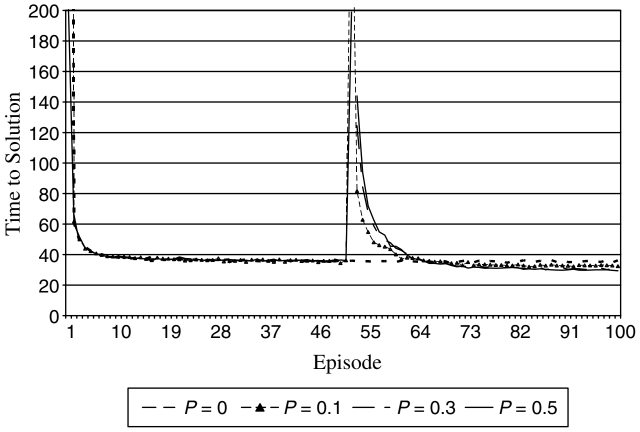
Again, what explains the dramatic difference between our results and March's (1991) results about the effects of turnover is the sequential character of the task we examine. Knowledge about how to execute a given routine will only be useful if all actors follow the routine. In this sense, a routine is only as strong as its weakest link. Because of this interdependency of the actors executing a routine, the immediate costs of deviating in one element of the organizational routine will be very high. Once outside the familiar territory of a routine, actors may have to engage in random search. Organizational processes such as Intel's "copy exact" program (Winter and Szulanski 2000) reflect this potential fragility of organizational routines.

Contrast this with the effects of turnover in the task structure of March's (1991) model. In that setting, organizational performance is an additive function of the accuracy of each element of the organizational code. As a result, changes in one element of the organizational code, caused by turnover of a member of the organization, will only have incremental consequences for organizational performance. Thus, the immediate costs of deviating in one element from the established code because of turnover are not substantial.

Consistent with our study, the Argote et al. (1995) experimental work on group problem solving shows that where groups need to solve a sequential task, turnover proves dysfunctional. In contrast, in settings where the task is to identify innovative or creative solutions to a problem, as a group, turnover may provide a useful source of variety. This later finding is observed in the innovation literature (cf. Katz 1982) and reflects the nature of March's (1991) simulation findings as well.[8]

The immediate effect of turnover in our model is that the time to solution increases substantially. However, in line with the results of March (1991; Figure 4), turnover can also eventually increase performance, even in our problem structure. Turnover can increase the rate of exploration and thereby improve performance in settings in which exploration is particularly valuable. In March's (1991) model, such a setting occurs when the organization is likely to get prematurely stuck at a suboptimal solution because of a high socialization rate. In an analogous fashion, the positive effects of turnover in our model occur when an organization can get prematurely stuck at an inefficient routine because of extensive credit assignment (i.e., a high value of λ) and, as a result, generate superstitious learning. In this situation, turnover can ultimately lead to improved performance, because it allows the organization to deviate from its routine and explore alternatives. This effect is illustrated in Figure 6, which shows the effect of turnover when $\lambda = 50$ and turnover only occurs in Episode 50.[9] As illustrated, a positive level of turnover eventually leads to a lower time to the solution than with zero turnover. Recall from Figure 5, however, that this is not the case when $\lambda = 1$, corresponding to the case of slow learning or a low socialization rate in March (1991) model. In this case, zero turnover always dominates a positive level of turnover.

Figure 6 **Effect of Turnover as a Function of the Probability of Turnover**



Note. $\lambda = 50$; each state experiences turnover only at Episode 50, subject to the probability P; note that in results not displayed due to scale limits, at the point of turnover the time to solution reaches over 1100 for $P = 0.5$, over 800 for $P = 0.3$, and over 300 for $P = 0.1$.

However, as seen in Figure 6, the immediate costs of turnover are quite high while the eventual benefits are relatively modest, even when λ is as high as 50. In fact, in the case of turnover probability of 0.1, time to solution in several episodes after turnover reaches more than 300 periods. This represents an immediate 10-fold increase in time to solution as compared to the case with no turnover. In the long run, however, this modest degree of turnover of 0.1 produces slightly better performance, with the difference amounting to only a few periods. As a result, in this task setting an organization has to be very patient and have large reserves to be able to survive and eventually benefit from turnover.

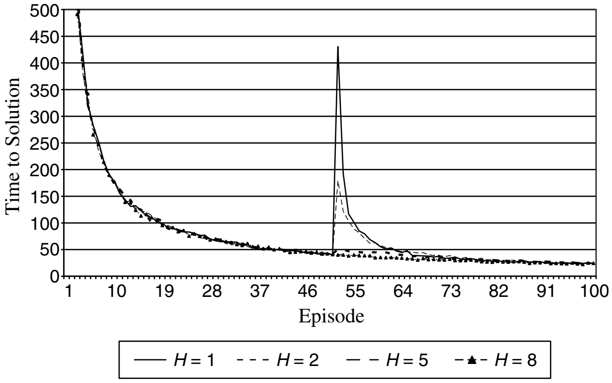
Our analysis also suggests that the costs of turnover will depend crucially on where turnover occurs. In contrast to models where organizational performance is an additive function of individual actions and

maximum value; then draw another random number between 0 and 10 to decide the magnitude of the assigned value. In this way, only one of the $N+1$ possible actions for any given state is assigned a belief. However, given the "greedy" choice algorithm, it is necessary to specify only the action with the highest value.

[8] It is also consistent with a secondary finding of Argote et al. (1995), who find that the negative effect of turnover is mitigated in complex task environments where there are opportunities for innovation with respect to production processes and not merely the refinement of an operating routine.

[9] Turnover is here defined as setting $Q(s, a)$ to zero for those actors who are replaced.

Figure 7 Effect of Turnover at Different Hamming Distances



Note. $\lambda = 1$; all states at specific hamming distances experience turnover at Episode 50.

turnover is equally damaging wherever it occurs (March 1991), the effect of turnover in our task structure depends on how close to the solution it occurs. In particular, as seen in Figure 7, the immediate costs of turnover will be greater if it occurs at a state close to the solution.

In this task structure, knowledge about appropriate moves at states close to the solution is much more valuable than knowledge about appropriate moves at states far away from the solution. Consider what would happen if individuals responsible for the states one step away from the solution were replaced with individuals lacking any knowledge about the location of the solution state. Initially, the behavior of such new members would be random. Such random action raises the possibility that the organization will be thrown into unfamiliar territory, where searching will continue to be random until some prior routine is identified, thereby substantially increasing the time to the solution. While loss of knowledge at states farther away from the solution state will also increase the time to the solution, the loss will not be as significant. This is, in part, because such states are less likely to be informed, but more importantly because such states are less likely to be visited on the way to the solution. The argument suggests that an organization's ability to recognize when it is close to a solution is quite valuable. For example, an organization that understands the preferences of consumers can better evaluate suggestions for new products at the point of commercialization and can in turn be expected to have much better performance in product development than a firm with insight about what basic technical approaches might be useful but little understanding of the subsequent path to commercialization.

Overall, these results suggest that the effect of turnover will differ depending upon the task structure facing the organization. In settings in which deviations by one individual will not have a large influence on organizational performance, such as when decisions are made by a group of individuals, the immediate costs of turnover may be small but the eventual benefits large. In other settings, such as that modeled here, when effective organizational performance requires adherence to a routine, the immediate costs of turnover will be substantial. In addition, when organizations need to learn an effective sequence of actions, knowledge about the positional value of some states close to the solution is critical, and as a result, turnover of actors in these states is particularly damaging.

5. Conclusion

Many organizational actions are characterized by sequential interdependency (Thompson 1967). As a result, many such actions do not have any immediate or direct payoff consequence but set the stage for subsequent actions that bring the organization toward some actual payoff. This paper has examined the possibilities and limitations of organizational learning in such contexts. In particular, we have modeled organizational learning in the absence of immediate payoff feedback; we have done this as successive attempts at credit assignment in which the process of reinforcement learning is extended to include actors' mental model as a basis for reinforcement. The standard model of reinforcement learning in the literature, in which only outcome feedback is viewed as a basis for reinforcement, emerges as a special case of this more general model. Credit assignment makes possible the development of a more intelligent belief structure over time. A valid mental map tends to emerge quickly for states close to the solution; however, assigning credit to more distant states requires many more trials and is likely to provide a less accurate sense of value. Nevertheless, even a fragmentary mental model may provide substantial performance improvements.

Our analysis highlights several features of organizational learning that have not been emphasized in previous research. First, it illustrates the performance improvement of even partial knowledge as well as the high costs of ignoring such partial knowledge to search for improvements. Knowledge may be partial in the sense that a path, or routine, has been identified, even if this routine is far from the most efficient patterned action sequence possible. Knowledge may also be partial in that the organization may have information about appropriate action at an isolated node in the problem space, even though it may have little knowledge outside of that particular node. Partial information, whether of an isolated node of knowledge in the problem space or a circuitous and relatively inefficient routine, is of tremendous value in guiding search and avoiding long random walks.

Simon's (1969) example of partial clues in the context of cracking a defective safe is illustrative. Searching for the code of a safe that has 10 dials each with 100 possible settings is virtually impossible, as there are 100^{10} possible configurations. However, if the safe is defective, such that a click can be heard whenever any one dial happens to correspond to the correct setting, then the total number of settings is greatly reduced to only 500. In a similar way, the existence of *some* routine, although limited and even inefficient, can greatly reduce the time spent in unproductive random searching.

Second, the fragility of learning in the context of a highly interdependent task environment casts a very different light on the role of turnover on organizational search processes. The benign role of turnover in March's (1991) analysis of search can be understood as a result of the task structure with only pooled interdependence in which actors' judgments are aggregated to form an "organizational" point of view. This sort of process of pooling of opinions seems reflective of problems of innovation and creativity and is consistent with empirical research in this domain that suggests a positive value of turnover (Katz 1982). However, in empirical work examining the development of routine behavior (Argote et al. 1995), turnover is shown to have a dysfunctional implication, as suggested by our simulation analysis. Socially interdependent knowledge is much more fragile and likely to be disrupted by turnover of personnel. We also find that turnover more proximate to the solution is particularly problematic.

There are, however, at least three important caveats to our work. First, we have restricted our attention to a stationary environment. A full analysis of shifting problem landscapes would require an examination of the generalization of the imputed value of state-action pairs in one environmental setting to another. Second, the learning problem we have studied involves a game against nature, not a game against other like-minded opponents. As such, we leave out potentially interesting strategic interactions as well.

Last, with respect to the task structure, we have created a world in which only positional advantage matters, not any immediate payoffs. Thus, our analyses have not been able to address how organizations could learn in problem contexts when immediate feedback is available but possibly misleading, such as when competency traps are present (Lave and March 1975, Levinthal and March 1981). Such a problem context is different in important ways that make direct application of our model to this context nontrivial. One way to think about such a problem context would be to examine a problem structure that contains immediate payoffs throughout the surface but that has an imperfect correlation among these payoffs. As a result, moving in the direction of a higher immediate payoff need not take one to even higher payoffs; the surface might have multiple peaks and, in Kauffman's (1993) terms, may be a rugged landscape.

It is possible that credit assignment in such a context, if the same task were repeated, might allow agents to develop cognitive representations of the value surface that effectively allows them to "bridge valleys" in such rugged landscapes and thus avoid local optima. As a result, the mental model could potentially free the search process from the topography of the multipeak surface and reduce the tendency of local search in such settings to result in competency traps of modest local peaks. However, given the scope of the current modeling effort and the complications that such analysis poses regarding the discounting of payoffs and appropriate stopping rules for the search process, we have not tried to incorporate such an analysis into the present paper.

Despite these limitations, the current effort begins to engage a largely neglected topic in models of organizational learning. Organizations' beliefs and predictions about the world can be an important basis for reinforcement learning, as well as actual outcomes. Indeed, in complex learning tasks with few and infrequent external cues of performance, such internal (or model-based) bases of reinforcement are critical. In this manner, models of behavioral learning and models of cognition can be effectively joined to provide us with a fuller conception of learning processes.

Acknowledgments

The authors gratefully acknowledge research support from the Sol Snider Center for Entrepreneurship at the Wharton School and the Institute of International Business at Stockholm School of Economics. The authors are grateful for comments on a prior draft by Mark Fichman, Teck Ho, Stefan Jonsson, Nicolaj Siggelkow, Sidney Winter, participants at seminars at the Academy of Management Conference (Washington DC, 2001), Stockholm School of Economics, UCLA Lake Arrowhead Conference, two anonymous reviewers, the associate editor, and the departmental editor, Linda Argote.

References

Argote, L. 1999. *Organizational Learning: Creating, Retaining and Transferring Knowledge*. Kluwer Academic Publishers, Boston, MA.

Argote, L., C. Insko, N. Yovetich, A. Romero. 1995. Group learning curves: The effects of turnover and task complexity on group performance. *J. Appl. Social Psych.* **25** 512–529.

Axelrod, R., M. D. Cohen. 1999. *Harnessing Complexity: Organizational Implications of a Scientific Frontier*. The Free Press, New York.

Bellman, R. 1957. *Dynamic Programming*. Princeton University Press, Princeton, NJ.

Bertsekas, D. P., J. Tsitsiklis. 1996. *Neuro-Dynamic Programming*. Athena Scientific, Belmont, MA.

Block, Z., I. MacMillan. 1985. Milestones for successful venture planning. *Harvard Bus. Rev.* **63**(5) 184–190.

Brehmer, B. 1995. Feedback delays in complex dynamic decision tasks. P. Frensch, J. Funke, eds. *Complex Problem Solving, The European Perspective*. Erlbaum, Hillsdale, NJ, 103–130.

Bruderer, E., J. Singh. 1996. Organizational evolution, learning and selection: A genetic algorithm based model. *Acad. Management J.* **39**(5) 1322–1329.

Camerer, C. 1997. Progress in behavioral game theory. *J. Econom. Perspect.* **11**(4) 167–188.

Carley, K. 1992. Organizational learning and personnel turnover. *Organ. Sci.* **3**(1) 20–46.

Chang, M.-H., J. E. Harrington. 1998. Centralization vs. decentralization in a muli-unit organization: A computational model of a retail chain as a multi-agent adaptive system. *Management Sci.* **46**(11) 1427–1440.

Cohen, M., P. Bacdayan. 1994. Organizational routines are stored as procedural memory. *Organ. Sci.* **5** 554–568.

Cyert, R., J. March. 1963. *A Behavioral Theory of The Firm*. Prentice Hall, Englewood Cliffs, NJ.

Denrell, J., J. March. 2001. Adaptation as information restriction: The hot stove effect. *Organ. Sci.* **12**(5) 523–538.

Forrester, J. W. 1961. *Industrial Dynamics*. Productivity Press, Portland, OR.

Gavetti, G., D. Levinthal. 2000. Looking forward and looking backward: Cognitive and experiential search. *Admin. Sci. Quart.* **45**(1) 113–137.

Gibson, F., M. Fichman, D. C. Plaut. 1997. Learning in dynamic decision tasks: Computational model and empirical evidence. *Organ. Behavior Human Decision Processes* **71**(1) 1–35.

Glynn, M. A., T. K. Lant, F. J. Milliken. 1994. Mapping learning processes in organizations. *Adv. Managerial Cognition Organ. Inform. Processing* **5** 43–93.

Greve, H. R. 1998. Performance, aspirations, and risky organizational change. *Admin. Sci. Quart.* **43**(1) 58–86.

Herriott, S. R., D. Levinthal, J. G. March. 1985. Learning from experience in organizations. *Amer. Econom. Rev. Papers Proc. Ninety-Seventh Annual Meeting* **75**(2) 298–302.

Holland, J. 1998. *Emergence: From Chaos to Order*. Oxford University Press, Oxford, U.K.

Holland, J., K. Holyoak, R. Nisbett, R. Thagard. 1986. *Induction: Processes of Inference, Learning and Discovery*. MIT Press, Cambridge, MA.

Huber, G. P. 1991. Organizational learning: The contributing processes and the literatures. *Organ. Sci.* **2**(1) 88–115.

Kaelbling, L. 1993. *Learning in Embedded Systems*. MIT Press, Cambridge, MA.

Katz, R. 1982. The effects of group longevity on project communication and performance. *Admin. Sci. Quart.* **27**(1) 81–105.

Kauffman, S. 1993. *The Origins of Order*. Oxford University Press, New York.

Lant, T. K. 1992. Aspiration level updating: An empirical exploration. *Management Sci.* **38** 623–644.

Lant, T. K. 1994. Computer simulations of organizations as experiential learning systems: Implications for organization theory. K. Carley, M. Prietula, eds. *Computational Organization Theory*. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.

Lant, T. K., S. J. Mezias. 1990. Managing discontinuous change: A simulation study of organizational learning and entrepreneurship. *Strategic Management J.* **11**(4) 147–179.

Lant, T. K., S. J. Mezias. 1992. An organizational learning model of convergence and reorientation. *Organ. Sci.* **3**(1) 47–71.

Lave, C., J. March. 1975. *An Introduction to Models in the Social Sciences*. Harper and Row, New York.

Levinthal, D. 1997. Adaptation on rugged landscapes. *Management Sci.* **43**(7) 934–950.

Levinthal, D. 2000. Organizational capabilities in complex worlds. G. Dosi, R. Nelson, S. Winter, eds. *The Nature and Dynamics of Organizational Capabilities*. Oxford University Press, New York.

Levinthal, D., J. March. 1981. A model of adaptive organizational search. *J. Econom. Behavior Organ.* **2** 307–333.

Levitt, B., J. March. 1988. Organization learning. *Ann. Rev. Sociology* **14** 319–340.

Lin, Z., K. M. Carley. 1997. Organizational response: The cost performance tradeoff. *Management Sci.* **43**(2) 217–234.

Lomi, A., E. R. Larsen, A. Ginsberg. 1997. Adaptive learning in organizations: A system dynamics-based explorations. *J. Management* **23**(4) 561–582.

Lounamaa, P., J. March. 1987. Adaptive coordination of a learning team. *Management Sci.* **33**(1) 107–123.

Lyles, M. A., C. R. Schwenk. 1992. Top management, strategy, and organizational knowledge structures. *J. Management Stud.* **29**(2) 155–174.

March, J. G. 1991. Exploration and exploitation in organizational learning. *Organ. Sci.* **2** 71–87.

McKelvey, B. 1999. Avoiding complexity catastrophe in coevolutionary pockets: Strategies for rugged landscape. *Organ. Sci.* **10**(3) 294–321.

Mezias, S. J., M. A. Glynn. 1993. Three faces of corporate renewal: Institution, revolution, and evolution. *Strategic Management J.* **14**(2) 77–101.

Miner, A. S., S. J. Mezias. 1996. Ugly duckling no more: Pasts and futures of organizational learning research. *Organ. Sci.* **7**(1) 88–99.

Minsky, M. L. 1961. Steps towards artificial intelligence. *Proc. Inst. Radio Engineers* **49** 8–30.

Nelson, R., S. G. Winter. 1982. *An Evolutionary Theory of Economic Change.* Harvard University Press, Cambridge, MA.

Repenning, N. P., J. D. Sterman. 2002. Capability traps and self-confirming attribution errors in the dynamics of product development. *Admin. Sci. Quart.* **47**(2) 265–295.

Rivkin, J. W. 2000. Imitation of complex strategies. *Management Sci.* **46**(6) 824–844.

Rivkin, J. W., N. Siggelkow. 2003. Balancing search and stability: Interdependencies among elements of organizational design. *Management Sci.* **49** 290–311.

Samuel, A. 1959. Some studies in machine learning using the game of checkers. *IBM J. Res. Development* **31** 211–229.

Samuel, A. 1967. Some studies in machine learning using the game of checkers II—Recent progress. *IBM J. Res. Development* **11** 601–617.

Sastry, A. 1997. Problems and paradoxes in a model of punctuated organizational change. *Admin. Sci. Quart.* **42**(2) 237–275.

Simon, H. 1969. *The Sciences of the Artificial*. MIT Press, Cambridge, MA.

Sterman, J. 1989a. Modeling managerial behavior: Misperceptions of feedback in a dynamic decision making experiment. *Management Sci.* **35**(3) 321–339.

Sterman, J. 1989b. Misperceptions of feedback in dynamic decision making. *Organ. Behavior Human Decision Processes* **43**(3) 301–328.

Sterman, J. 2000. *Business Dynamics Systems Thinking for a Complex World*. Irwin/McGraw-Hill, New York.

Sutton, R., A. Barto. 1981. Toward a modern theory of adaptive networks: Expectation and prediction. *Psych. Rev.* **88** 135–170.

Sutton, R., A. Barto. 1998. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.

Thompson, J. 1967. *Organizations in Action*. McGraw-Hill, New York.

Walsh, J. 1995. Managerial and organizational cognition: Notes from a trip down memory lane. *Organ. Sci.* **6**(3) 280–321.

Watkins, C. 1989. Learning from delayed rewards. Ph.D. thesis, Kings' College, Cambridge, U.K.

Yelle, L. E. 1979. The learning curve: Historical review and comprehensive survey. *Decision Sci.* **10**(2) 302–328.