

Effects of feedback delay on learning

Hazhir Rahmandad,^{a*} Nelson Repenning^b and John Sterman^b

Hazhir Rahmandad is assistant professor of Industrial and Systems Engineering at Virginia Tech. He received his Ph.D. from MIT System Dynamics group and his undergraduate in Industrial Engineering from the Sharif University of Technology. Hazhir's research interests include organizational learning, dynamics of product development capability, comparing different simulation methods, and model estimation methodologies.

Nelson P. Repenning is an Associate Professor of Management Science and Organization Studies at the MIT Sloan School of Management. His work focuses on understanding the factors that contribute to the successful implementation, execution, and improvement of business processes. Professor's Repenning has received several awards for his work, including best paper recognition from both the *California Management Review* and the *Journal of Product Innovation Management*. In 2003

Abstract

Understanding barriers to organizational learning is central to understanding firm performance. We investigate the role of time delays between taking an action and observing the results in impeding learning. These delays, ubiquitous in real-world settings, are relevant to tradeoffs between long term and short term. We build four learning heuristics, with different levels of complexity and rationality, and analyze their performance in a simple resource allocation task. All reliably converge to the optimal solution when there are no/short delays, and when those delays are correctly assessed. However, learning is slowed significantly when decision makers err in assessing the length of the delay. In many cases, the decision maker never finds the optimal solution, wandering in the action space or converging to a suboptimal allocation. Results are robust to the organization's level of rationality. The proposed heuristics can be applied to a range of problems for modeling learning from experience in the presence of delays. Copyright © 2009 John Wiley & Sons, Ltd.

Supporting information may be found in the online version of this article.

Syst. Dyn. Rev. **25**, 309–338, (2009)

Introduction

Learning figures prominently in many perspectives on organizations (Cyert and March, 1963). Learning underlies performance improvements in organizations including firms (Argote and Epple, 1990), and is viewed by many researchers as an important source of competitive advantage (DeGeus, 1988; Stata, 1989; Senge, 1990). Organizational routines and capabilities fundamental to firm performance are learned through an adaptive process of local search and exploration (Nelson and Winter, 1982). Therefore, it is critical to understand what mechanisms promote and, possibly, hinder learning in organizations.

The literature stresses a number of impediments to learning (Sterman, 1994). Noise and ambiguity in the payoff that an organization receives increase the chance of superstitious learning (Argyris and Schön, 1978; Levitt and March, 1988), where the organization draws erroneous conclusions from its experience. Stochasticity in results can also prevent selection of a useful strategy in which an unlucky first experience discourages future experimentation (Denrell and March, 2001).

^a Industrial and Systems Engineering Department, Virginia Tech, Northern Virginia Center, 7054 Haycock Road, Falls Church, VA 22043, U.S.A.

^b Sloan School of Management, MIT, E53–335, 30 Wadsworth Ave., Cambridge, MA 02142, U.S.A.

* Correspondence to: Hazhir Rahmandad. E-mail: hazhir@vt.edu

Received February 2008; Accepted April 2009

he received the International System Dynamics Society's Jay Wright Forrester award, which recognizes the best work in the field in the previous five years. His current interests include safety in high hazard production environments and the connection between efficient internal operations and effective strategic positions.

John D. Sterman is the Jay W. Forrester Professor of Management at the MIT Sloan School of Management and Director of the MIT System Dynamics Group.

Repetition and practice build competence, but can also impede exploration of untried and possibly better options through habit, inertia, and, paradoxically, through improvement itself. As experience with a particular set of routines and behaviors improves performance, the (opportunity) cost of trying other options rises, closing the door to experimentation with other, potentially superior methods (Levinthal and March, 1981; Herriott *et al.*, 1985). Such competency traps (Levitt and March, 1988) also discourage exploration of untried possibilities in the face of a changing environment, thereby turning core capabilities into core rigidities (Leonard-Barton, 1992).

The complexity of the payoff landscape on which organizations adapt poses another challenge to organizational learning. The complementarities among different elements of a firm's strategy (Milgrom and Roberts, 1990) result in a complex payoff landscape in which improvements in performance often come from changing multiple aspects of the strategy together. In other words, after some local adaptation to find an internally consistent strategy, incremental changes in one aspect of the strategy are often not conducive to performance gains, so that the firm rests on a local peak of a rugged payoff landscape (Busemeyer *et al.*, 1986; Levinthal, 1997).

Temporal challenges to learning such as delays between actions and payoffs have received little attention in organizational learning research. In contrast, as early as the 1920s the importance of the temporal contiguity of stimulus and response has been recognized by psychologists studying conditioning (Warren, 1921). However, despite empirical evidence of the adverse effects of delays on individual learning (Sengupta *et al.*, 1999; Gibson, 2000) and some case-specific evidence at the organizational level (Repenning and Sterman, 2002), with a few exceptions (Denrell *et al.*, 2004; Fang and Levinthal, 2008; Rahmandad, 2008), the effects of delays on organizational learning have not been formally studied. In fact, few formal models of organizational learning capture such delays explicitly.

Denrell and colleagues (2004) build on the literature of learning models in artificial intelligence to specify a Q-learning (Watkins, 1989) model of a learning task in a multidimensional space with a single state with non-zero payoff. The Q-learning algorithm is based on updating the long-term value (Q-function) of taking different actions from different points on the state space based on the immediate reward of action-state combinations and the value of the resulting state. Their analysis informs situations in which several actions must be taken before any payoff is observed but does not capture the tradeoffs between short-term and long-term alternatives. Rahmandad (2008) extends their framework to a simple task where one of three actions can be chosen at every period and investigates the impact of delays on the evolution of cognitive maps. Fang and Levinthal (2008) use this framework to show that exploration has unique benefits associated with signaling the value of intermediate steps in multi-stage tasks. These studies elaborate on the central problem of credit assignment, how to build associations between actions and payoffs in the presence of delays,

and highlight the importance of organizing experience in efficient cognitive maps. However, they use an information-intensive learning algorithm that requires thousands of action–payoff cycles to discover the optimum even in a simple task. Moreover, the tasks used in these studies are stylized and much simpler than typical organizational phenomena.

We expand these results by examining learning in a resource allocation setting that resembles some important real-world challenges facing organizations, specifically, allocating resources among activities whose contribution to performance can manifest with different time delays. For example, plant managers must allocate worker time between production (which pays off today) and process improvement (which pays off in the future); firms must allocate budget between sales effort (which pays off relatively quickly) and product development (which pays off later). We study how much performance and learning weaken as a result of delays, how sensitive the results are to the rationality of the decision maker, and describe the mechanisms through which delays impede learning. A formal model enables us to pin down the effects of time delays on learning by removing other confounding factors that potentially influence learning, i.e., feedback noise, multiple peaks and competency traps, high discount rates, and interpersonal dynamics. We draw on the current literature on learning in organizational behavior, psychology, game theory, and attribution theory to specify and analyze four different learning procedures with different levels of rationality, information-processing requirements, and cognitive search ability.

Our contributions are twofold. First, by running controlled computational experiments, we gain insight into the mechanisms through which delays impede learning. By examining different heuristics, we can distinguish the behavioral patterns that are common across different heuristics and therefore derive more robust conclusions. Moreover, using different heuristics allows us to differentiate between the effects of decision maker rationality and problems arising from delayed feedback. Second, we introduce a framework to model organizational learning in the presence of delays in continuous time and space. Previous models of learning in organizations do not consider delays, or require unrealistic cognitive power and amounts of data to learn about delays in very simple tasks (e.g., Rahmandad, 2008). This latter contribution is of special interest to modelers of dynamic organizational phenomena, including the system dynamics community. Feedback, its misperception, and learning traps figure prominently in this literature (e.g., Sterman, 1994). This study provides formal models of the learning process that can be used in a variety of models of organizational dynamics.

Our results confirm the hypothesis that delays complicate the attribution of causality between action and payoff and therefore can hinder learning (Einhorn and Hogarth, 1985; Levinthal and March, 1993). Furthermore, our analysis shows that performance can still be significantly and persistently suboptimal in very easy learning tasks, specifically, even when the payoff landscape is

unchanging, smooth, and has a unique optimum. Moreover, the organization can come to believe that suboptimal performance is really the best it can do, converging to a poor allocation and becoming unwilling to change it. Finally, the results are robust to different learning heuristics spanning a wide range of rationality and computational intensity. The difficulty of learning in the presence of delays appears to be rooted in the misperception of the delays between actions and outcomes, rather than the specific learning procedures we examine.

Beyond the general problem of credit assignment, three main processes drive the results. First, dynamic systems often exhibit tradeoffs between the short- and long-term effects of actions (Forrester, 1969, Ch. 6; Repenning and Sterman, 2001; Repenning and Sterman, 2002). Such “worse-before-better” dynamics can impede learning when decision makers do not properly account for the delays. Second, delays mean the information the decision maker might use to assess the gradient of the payoff landscape is dated and potentially misleading. Finally, a low return on experimentation in the face of delays results in the convergence of the organization to suboptimal policies.

Next we describe the organizational task and learning heuristics in detail. The “Results and analysis” section presents a base case demonstrating that all four learning procedures can discover the optimum allocation in the absence of action–payoff delays. We then analyze the performance of the four learning procedures in the presence of action–payoff delays, followed by tests to examine the robustness of these results under different parameter settings. We close with a discussion of the implications, limitations, and possible extensions.

Modeling learning with delayed feedback

Resource allocation problems provide an appropriate context for studying the effect of time delays on learning. First, many important situations, at multiple levels of analysis, involve allocating resources among different activities with different delays between allocation and results. Examples include a factory allocating resources among production, maintenance, and process improvement, an individual allocating her time among work, education and leisure, and a government allocating budget among defense, social expenditures, roads and schools.

Moreover, these situations often involve tradeoffs between short- and long-term results. For example, the factory can boost output in the short run by cutting maintenance, but in the long run output falls as breakdowns increase; attending school to develop new skills lowers disposable income and leisure today, but may increase income and leisure later. Other examples include learning and process improvement (Repenning and Sterman, 2002), investment in proactive maintenance (Repenning and Sterman, 2001), and environmental degradation by human activity (Meadows *et al.*, 1972). These tradeoffs suggest

that individuals, organizations, and societies can often fail to learn from experience to improve their performance in these allocation and decision-making tasks.

Previous studies motivated the formulation of our model. Repenning and Sterman (2002) found that managers learned erroneous lessons from their experience with the workforce as a result of different time delays in the ways different types of worker activity influence the system's performance. Specifically, managers seeking to meet production targets have two basic options: (1) increase process productivity and yield through better maintenance and investment in improvement activity; or (2) pressure the workforce to "work harder" through overtime, speeding production, taking fewer breaks, and, most importantly, by cutting back on the time devoted to maintenance and process improvement. Though "working smarter" provides a greater payoff than working harder, many organizations find themselves stuck in a "capability trap" of working harder, resulting in reduced maintenance and improvement activity, lower productivity, greater pressure to hit targets and thus even less time for improvement and maintenance (Repenning and Sterman, 2002).

Paralleling this setting, our model represents an organization engaged in a continuous-time resource allocation task. The organization must allocate a fixed resource among different activities. The payoff generated by these decisions can depend on the lagged allocation of resources, and there may be different delays between the allocation of resources to each activity and the impact on the payoff. The organization receives outcome feedback about the payoff from past resource allocations, and attempts to learn from this information how to adjust resource allocations to improve performance. As a concrete example, consider a manufacturing firm allocating its resources (e.g., employee time) among three activities: production, maintenance, and process improvement. These activities influence production, but with different delays. Time spent on production yields results almost immediately. There is a longer delay between a change in maintenance activity and machine uptime (and hence production). Finally, it takes even longer for process improvement activity to affect output.

The organization gains experience by observing the results of past decisions and seeks to increase production based on this experience. It may have some understanding of the complicated processes controlling production, captured in the mental models of individuals and in organizational routines, and it may be aware of the existence of different delays between each activity and observed production. Consequently, when evaluating the effectiveness of its past decisions, the organization takes these delays into consideration (e.g., it does not expect last week's process improvement effort to enhance production today). However, the mental models of the production process may be imperfect, and there may be discrepancies between the length of the delays managers perceive and the true delays.

Learning task

We assume the payoff, $P(t)$, to be a constant-returns-to-scale,¹ Cobb–Douglas function of the effective allocations, which yields a smooth landscape with a single peak, allowing us to eliminate learning problems that arise in more complicated landscapes:

$$P(t) = \prod_j \hat{A}_j(t)^{\alpha_j}, \sum_j \alpha_j = 1 \quad (1)$$

Effective allocations, $\hat{A}_j(t)$, capture time delays between actions and outcomes. They lag behind the current allocation of resources, A_j , by T_j periods:

$$\hat{A}_j(t) = A_j(t - T_j) \quad (2)$$

The organization continuously allocates a fraction of total resources (e.g., the available person-hours) to activity j of m possible activities at time t , $F_j(t)$ where

$$\sum_j F_j(t) = 1 \quad \text{for } j: 1, \dots, m \quad (3)$$

In our simulations we assume $m = 3$ activities, so the organization has two degrees of freedom. Three activities keep the analysis simple while requiring the resolution of several formulation challenges that one faces in modeling learning where more than a single degree of freedom exists.² Total resources, $R(t)$, are constant, $R(t) = R$,³ so amount of resource allocated to activity j at time t , $A_j(t)$, is

$$A_j(t) = F_j(t) * R \quad (4)$$

In our manufacturing firm example, R is the total number of employees, $F_j(t)$ is the fraction of people working on activity j (j : producing, maintenance, process improvement) and $A_j(t)$ is the number of people working on activity j . The delays in realizing the impact of these activities on production (the payoff) are ordered approximately as

$$0 \approx T_{\text{production}} < T_{\text{maintenance}} < T_{\text{improvement}}$$

The perception of delays

The organization knows its own action and payoff history and uses this information to develop better allocation strategies. We assume that the organization accounts for the existence of the delays between allocations and payoffs. While the perceived and actual payoff, $P(t)$, are often different in real organizations, to give the learning heuristics the most favorable circumstances we assume

that measurement and perception are fast and unbiased. The organization accounts for the delays between allocations and payoff based on its beliefs about the length of the payoff generation delays, \bar{T}_j . It attributes the current observed payoff, $P(t)$, to allocations made \bar{T}_j periods ago, so the action associated with the current payoff, $\bar{A}_j(t)$, is

$$\bar{A}_j(t) = A_j(t - \bar{T}_j) \quad (5)$$

With this set-up, the task facing the organization is to determine the best allocation strategy, $F_j(t)$, given that it knows both its past payoff and action histories, $\bar{A}_j(s)$ and $P(s)$, $s \in [0, t]$.

Learning procedures

Modeling the process through which individuals and organizations translate past experience into current action requires confronting several outstanding questions concerning the limits of human rationality. Substantial scholarly ink has been spilled debating the degree of, and limits to, human rationality (see Lord and Maher, 1990, and Walsh, 1995, for reviews). Rather than taking a strong position on a specific model of rationality, we investigate the effect of time delays on decision-making performance as the level of rationality is varied. Specifically, to study the effects of delays on learning, we employ four different learning models that vary from myopic to highly sophisticated to explore the sensitivity of the results to different assumptions about how organizations learn. The inputs to all of these heuristics are the perceived payoff and action histories, $P(t)$ and $\bar{A}_j(t)$; the outputs of the heuristics are the allocation decisions $F_j(t)$. We first explain the different learning heuristics and then discuss the decision-making process for selecting $F_j(t)$ s.

The learning heuristics differ in their level of rationality, information-processing requirements, and assumed prior knowledge about the shape of the payoff landscape. Here, rationality indicates an organization's ability to make the best use of the available information by trying explicitly to optimize its allocation decisions. Information-processing capability indicates an organization's capacity to keep track of and use information about past allocations and payoffs. Prior knowledge about the shape of the payoff landscape determines the ability to use offline cognitive search (Gavetti and Levinthal, 2000) to find better policies.

We label the four learning heuristics *Reinforcement*, *Myopic Search*, *Correlation*, and *Regression*.⁴ Each learning heuristic begins with the decision maker's mental representation of the relative importance of each activity, the "Activity Value", $V_j(t)$. The Activity Values determine the allocation of resources to each activity (Eqs 12–14 explain how the allocation of resources is determined based on Activity Values). Each learning heuristic consists of a different set of assumptions concerning how the Activity Values are updated. Below we

discuss the four learning heuristics in more detail; the complete formulation of all the learning heuristics can be found in the supplement.

Reinforcement learning

In this method, the value (or attractiveness) of each activity is determined by (a function of) the cumulative payoff achieved so far by using that alternative. Attractiveness then influences the probability of choosing each alternative in the future. Reinforcement learning has a rich tradition in psychology, game theory, and machine learning (Erev and Roth, 1998; Sutton and Barto, 1998). It has been used in a variety of applications, from training animals to explaining the results of learning in games and designing machines to play backgammon (Sutton and Barto, 1998).

In the reinforcement learning model, each perceived payoff, $P(t)$, is associated with the allocations believed to be responsible for that payoff, and the value of each activity, $V_i(t)$, is increased based on its contribution to the perceived payoff. The increase in value depends on the perceived payoff itself, so a small payoff indicates a small increase in the values of different activities, while a large payoff increases the value much more. Therefore, large payoffs shift the relative weight of different Activity Values towards the allocations responsible for those better payoffs.

$$\frac{d}{dt} V_i(t) = P(t)^\gamma * \bar{A}_i(t) - V_i(t) / \tau_f \quad (6)$$

Our implementation of reinforcement learning includes a forgetting process, captured by the second term on the right-hand side of Eq. 6, representing the organization's discounting of older information. Discounting old information helps the heuristic adjust the Activity Values better. Both γ , the reinforcement power, and τ_f , the reinforcement forgetting time, are parameters specific to this heuristic. Reinforcement power indicates how strongly we feed back the payoff as reinforcement to adjust the Activity Values and therefore determines the speed of convergence to better policies. The reinforcement forgetting time is the time constant for depreciating the old payoff reinforcements.

The *Reinforcement* method is a low-information, low-rationality procedure: decision makers are assumed to continue to do what has worked well in the past, adjusting only slowly to new information. They do not attempt to extrapolate from these beliefs about activity value to the shape of the payoff landscape or to use gradient information to move towards higher-payoff allocations.

Myopic search

In this method the organization explores regions of the decision space close to the current allocation (at random). If a nearby allocation results in a better payoff than the aspiration (based on past performance), the action values are adjusted towards the explored allocation. If the exploration results in a worse payoff, the action values remain unchanged. This procedure is a variant of

Levinthal and March's (1981) adaptive search formulation. It is also similar to the process underlying many stochastic optimization techniques in which, unaware of the shape of the payoff landscape, the heuristic explores the landscape and then moves to better regions when they are found.

While optimization algorithms will often instantly adopt a newly discovered, better solution, in real-world settings more time is required to implement a new policy. It takes time to collect and report new information, process that information, and make changes. New technical and administrative innovations diffuse gradually throughout organizations. To capture the inertia in organizational mental models and routines, we assume the activity value, $V_j(t)$, adjusts gradually towards the value-set suggested by the last allocation, if the payoff improved in the last step:

$$\frac{d}{dt} V_j(t) = (V_j^*(t) - V_j(t)) / \tau_v \quad V_j^*(t) = \bar{A}_j / \sum \bar{A}_k \quad (7)$$

Activity Values remain unchanged if the payoff did not change significantly or decreased in the last step.

Here τ_v is the Value Adjustment Time Constant. The supporting information provides the full formulation (Table S2). The myopic method is a low-rationality method with medium information-processing requirements. It compares the previous payoff with the result of a local search and does not attempt to compare multiple experiments; it also does not use information about the payoffs in the neighborhood to make any inferences about the shape of the payoff landscape.

Correlation

The correlation method, which builds on principles from attribution theory, is significantly more sophisticated than the reinforcement and myopic search approaches. In our setting, learning can be viewed as the process through which organizational decision makers attribute different payoffs to different allocations and how these attributions are adjusted as new information about payoffs and allocations becomes available. Several researchers have proposed different models to explain how people make attributions. Lipe (1991) reviews these models and concludes that all major attribution theories are based on the use of counterfactual information. It is difficult, however, to obtain counterfactual information (information about contingencies that were not realized), so she proposes the use of covariation data as a proxy. The correlation heuristic is based on the hypothesis that people use the observed covariation of different activities with the payoff to make inferences about action–payoff causality.

In the correlation learning model, the correlations between the perceived payoff and the actions associated with those payoffs allow the organization to judge whether performance would improve or deteriorate if the activity increased. A positive (negative) correlation between recent values of $\bar{A}_j(t)$ and $P(t)$ suggests that more (less) of activity j will improve the payoff. Based on

these inferences the organization adjusts the Activity Values, $V_i(t)$, so that positively correlated activities are increased and negatively correlated ones are decreased:

$$\frac{d}{dt} V_i(t) = V_i(t) \cdot f(\text{Action_Payoff_Correlation}_i(t)) / \lambda \quad f(0) = 1, f'(x) > 0 \quad (8)$$

The formulation details for the correlation heuristic are found in the supporting information (Table S3). At the optimal allocation, the gradient of the payoff with allocation will be zero and so will the correlation between activities and payoff. Therefore the change in Activity Values will become zero, and the organization thus settles on the optimum policy.

The correlation method is a moderate-information, moderate-rationality approach: more data are needed than required by the myopic or reinforcement methods to estimate the correlations among activities and the payoff. Moreover, these correlations are used to make inferences about the local gradient so the organization can move uphill from the current allocations to allocations believed to yield higher payoffs, even if these allocations have not yet been tried.

Regression

The final model, the regression method, is a relatively sophisticated learning heuristic with significant information-processing requirements. We assume that the organization knows the correct shape of the payoff landscape and uses a correctly specified regression model to estimate the parameters of the payoff function.

By observing the payoffs and the activities corresponding to those payoffs, the organization receives the information needed to estimate the parameters of the payoff function. To do so, after every few periods it runs a regression using all the data from the beginning of the learning task. From these estimates the optimal allocations are readily calculated. Raghu *et al.* (2003) use a similar learning heuristic to model how simulated fund managers learn about the effectiveness of different funding strategies. For the assumed constant returns to scale Cobb–Douglas function, the regression is

$$\log(P(t)) = \log(\alpha_0) + \sum_j \alpha_j \cdot \log(\bar{A}_j(t)) + e(t) \quad (9)$$

The estimates of α_j , α_j^* are evaluated every “Evaluation Period”, E . Based on these estimates, the optimal Activity Values, $V_j^{*}(t)$, are given by

$$V_j^{*}(t) = \max\left(\alpha_j^* / \sum_j \alpha_j^*, \varepsilon\right) \quad (10)$$

Here ε is a small number that ensures the values remain above zero. The organization then adjusts the action values towards the indicated optimal

Table 1.
Sophistication and
rationality of different
learning heuristics

Dimension heuristic	Rationality	Information-processing capacity	Prior knowledge of payoff landscape
Reinforcement	Low	Low	Low
Myopic Search	Low	Medium	Low
Correlation	Medium	Medium	Low
Regression	High	High	High

values (see Eq. 7). See Table S4 in the supporting information for equation details.

The regression heuristic is a high-rationality, high-information heuristic. Although in feedback-rich settings, mental models are far from perfect and calculation of an optimal decision based on understanding of mechanisms is often cognitively infeasible (Sterman 1989a, 1989b), the regression heuristic offers a case of high rationality to test the robustness of our results. Table 1 summarizes the characteristics of the different heuristics on the three dimensions of rationality, information-processing capacity, and prior knowledge of payoff landscape.

Exploration and exploitation

Balancing exploration and exploitation is a crucial issue in learning (March, 1991; Sutton and Barto, 1998). On the one hand, to discover possible improvements and learn about the shape of the payoff landscape the organization should explore the decision space by trying some new allocation policies. On the other hand, deviations from current practice will often lower performance, and pure exploration leads to random allocation decisions with no improvement. Exploration is needed to learn about the payoff landscape and purposeful exploitation is required to make good use of what has been discovered. We use random changes in resource allocations to capture the exploration/exploitation balance in our learning models. The “Activity Values” represent the accumulation of experience and learning by the organization. In pure exploitation, the set of Activity Values, $V_j(t)$, determines the allocation decision. Specifically, the fraction of resources to be allocated to activity j would be

$$F'_j(t) = V_j(t) / \sum_j V_j(t) \quad (11)$$

The degree to which the organization follows this policy shows its tendency to exploit the experience it has gained so far. Deviations from this policy represent experiments to explore the payoff landscape. We add a random disturbance to the Activity Values to generate Operational Activity Values, $\hat{V}_j(t)$, which are the basis for the allocation decisions.

$$\hat{V}_j(t) = V_j(t) + \hat{N}_j(t) \quad (12)$$

$$\hat{N}_j(t) = N_j(t) * \sum_j V_j(t) \quad (13)$$

$$F_j(t) = \hat{V}_j(t) / \sum_j \hat{V}_j(t) \quad (14)$$

$N_j(t)$ is a first-order autocorrelated noise term specific to each activity with mean 0 and truncated to keep the operational values positive. In real settings, the organization experiments with a policy for some time before moving to another. Such persistence is physically required (organizations cannot instantly change resource allocations) and provides organizations with a large enough sample of data about each allocation to decide if a policy is beneficial or not. The “activity noise correlation time”, δ , captures the degree of autocorrelation in the noise stream $N_j(t)$. High values of δ represent organizations that only slowly change the regions of allocation space they are exploring; a low value represents organizations that move quickly from one allocation to another in the neighborhood of their current policy. To ensure that the exploration disturbances are of comparable magnitude across the different learning heuristics, we scale the noise stream, $N_j(t)$, by the sum of the action values for each heuristic (Eq. 13) to yield the actual exploration term, $\hat{N}_j(t)$.

The standard deviation of the noise term, which determines how far explorations deviate from the current resource allocation, depends on where the organization finds itself on the payoff landscape. If its recent explorations have shown no improvement in the payoff, it concludes that it is near the peak of the payoff landscape and therefore extensive exploration is not required (alternatively, it concludes that the return on exploration is low and reduces experimentation accordingly). If, however, recent exploration has resulted in the discovery of significant improvement, it concludes that the return on exploration is high and hence it should keep on exploring, so $\text{var}(N_j(t))$ remains large:

$$\begin{aligned} \text{var}(N_j(t)) &= g(\text{Recent Payoff Improvement}(t)), \\ g(0) &= \text{Minimum Exploration Variance}, g'(x) > 0 \end{aligned} \quad (15)$$

In this formulation, if the current payoff is higher than recent payoffs, Recent Payoff Improvement will be increased; if not, it will decay towards 0. Details of exploration equations can be found in Table S5 of the supporting information.

Results and analysis

In this section we investigate the behavior of the learning model under different conditions. We first explore the behavior of the model and its learning

capabilities when there are no delays. The no-delay case helps with comparing the capabilities of the different learning heuristics and provides a base against which we can compare the behavior of the model under other conditions. Next we introduce delays between activities and payoff and analyze the ability of the different learning heuristics to find the optimal payoff, for a wide range of parameters.

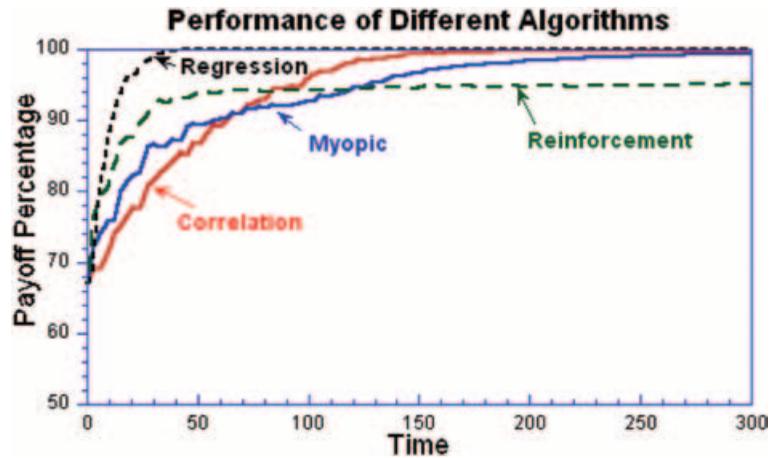
Without loss of generality, the total resource, R , is 100 units. The payoff function exponents are set to $1/2$, $1/3$, and $1/6$ for activities one, two, and three, respectively. These values ensure that all activities make nontrivial, but different, contributions to the payoff. The optimal allocation fractions are easily shown to be $1/2$, $1/3$, and $1/6$, and optimal payoff is 36.37. Each simulation begins with a random initial resource allocation. The simulation horizon is set to 300 periods: long enough to give the organization opportunity to learn, while keeping the simulation time reasonable. Consistent with the manufacturing firm example, we set the time step for the simulation to one month and choose the time constants to be consistent with typical manufacturing contexts. The 300-period horizon thus corresponds to 25 years: ample time to examine how much the organization learns.

We report two aspects of learning performance: (1) proximity to the optimal solution; and (2) time to convergence, given that convergence occurs. The percentage of the optimal payoff achieved by the organization at the end of simulation is reported as Achieved Payoff Percentage. Monte Carlo simulations with different random noise seeds for the exploration term (Eq. 13) and for the randomly chosen initial resource allocation give statistically reliable results for each scenario analyzed (see the supporting information for additional notes on the definition of metrics). To facilitate comparison of the four learning heuristics, we use the same noise seed across the four models. Differences across models in a given run are due, therefore, only to the differences among the four learning procedures and not to the realizations of the random variables.

Base case

The base case represents an easy learning task in which there are no delays between actions and the payoff. The organization is also aware of this fact and therefore has a perfect understanding of the delay structure. Figure 1 shows the trajectory of the payoff for each learning heuristic, averaged over 100 simulation runs. The vertical axis shows the percentage of the optimal payoff achieved by each learning heuristic. As expected, when there are no delays all four learning heuristics converge to the neighborhood of the optimal resource allocation. For comparison, the average payoff achieved by a random resource allocation strategy is 68 percent of optimal (since the 100 simulations for each learning heuristic start from randomly chosen allocations, the expected value of the initial payoff for all four heuristics is 68 percent).

Fig. 1. Percentage of payoff relative to optimal in the base case, averaged over 100 runs



An important aspect of learning is how fast the organization converges to the allocation policy it perceives to be optimal.

Table 2 reports, for the base case, the payoffs, the fraction of simulations that have converged within 200 months, and the average convergence time (defined as the time at which the standard deviation of performance falls below 1 percent of its historical average, for those that converged). The regression, correlation, and myopic heuristics find the optimum almost perfectly, and reinforcement gets quite close. The majority of simulations converge within 300 periods. The average convergence times range from a low of 38 periods for the regression heuristic to a high of 78 periods for the myopic model.

The impact of delays

Most models of organizational learning share the assumption that there are no delays between taking an action and perceiving its full consequences. Under

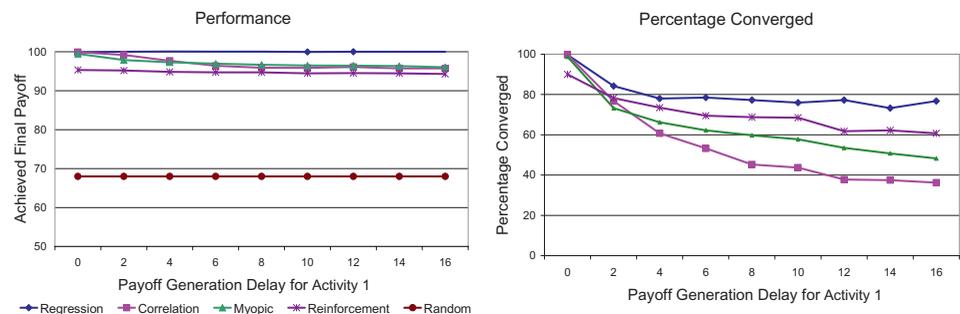
Table 2. Achieved Payoff, Convergence Time and Percentage Converged for the base case. Statistics from 400 simulations

Learning heuristic Variable/estimates	Regression		Correlation		Myopic		Reinforcement	
	μ	σ	μ	σ	μ	σ	μ	σ
Achieved Payoff percentage (at $t = 300$)	99.9	0.08	99.83	1.02	99.86	1.13	95.76	4.00
Convergence Time	38.04	11.92	78.51	40.53	78.47	49.82	45.60	19.21
Percentage Converged	99.75		99.75		99.5		91.5	

this conventional assumption, all our learning heuristics reliably find the neighborhood of the optimum allocation and converge to it. To begin investigating the consequences of relaxing this assumption, we introduce a delay in the impact of one of the activities, leaving the delays for the other two activities at zero. We simulate the model for nine different values of the “Payoff Generation Delay” for Activity 1, T_1 , ranging from 0 to 16 periods. Returning to the factory management example, the long delay in the impact of Activity 1 is analogous to process improvement activities that take a long time to bear fruit. Because Activity 1 is the most influential in determining the payoff, these settings are expected to highlight the effect of delays more clearly. Delays as high as 16 periods are realistic in magnitude (1.33 years in our manufacturing context—quite short relative to the actual time required for process improvement programs to bear fruit (Sterman *et al.*, 1997; Repenning and Sterman, 2002).) These delays are also comparatively short relative to the dynamics of exploration and learning. For example, in the base case, it takes between 38 (for regression) and 78 (for myopic) periods for different learning models to converge when there are no delays. (Table S6 in the supporting information provides an alternative measure of the speed of internal dynamics.)

We analyze two settings for the perceived payoff generation delay. First, we examine the results when the delays are correctly perceived and accounted for ($T_j = \bar{T}_j$). Next we analyze the effect of misperception of delays by setting the perceived payoff generation delay for all activities, including Activity 1, at zero. This setting corresponds to an organization that believes all activities affect the payoff immediately. Figure 2 reports the behavior of the four learning heuristics under the first setting, where delays are correctly perceived. Under each delay time the Average Payoff Percentage achieved by the organization at the end of 300 periods and the Percentage Converged are reported. In

Fig. 2. Payoff Percentage (left) and Percentage Converged (right) with different time delays for Activity 1. The delay in the impact of Activity 1 on performance is correctly perceived. There is no delay in the impact of Activities 2 and 3 on the payoff. Results are averages over 400 simulations



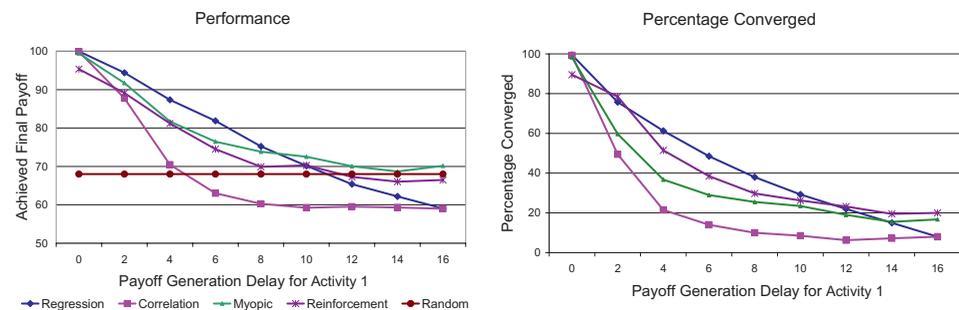
the “Average Payoff Percentage” graph, the line denoted “Pure Random” represents the case where the organization selects its initial allocation policy randomly and sticks to that, without trying to learn from its experience. Results are averaged over 400 simulations for each delay setting.

The performance of each of the four learning heuristics remains good. The regression model finds the optimum allocation policy in every delay setting if it has a correct perception of the delays involved. The average performance of all the heuristics remains over 94 percent of optimal. The convergence percentages decline as the organization continues to explore the neighborhood of the optimal allocation, but never settles down, in a substantial number of simulations. Figure 3 shows the performance of the four learning heuristics when the delay between Activity 1 and realizing its outcome is incorrectly perceived. The performance of each learning heuristic is adversely influenced by the misperception of a delay. There is also a clear drop in the convergence rates as a consequence of such delays.

In summary, the following patterns can be discerned from the results. First, learning is significantly hampered if the perceived delays do not match the actual delays. The pattern is consistent across the different learning procedures. As the misperception of delays grows, all four learning procedures yield mean performances worse than or equal to the random allocation policy. However, correctly perceived delays have only a minor impact on performance.

Second, if delays are misperceived, the convergence percentages decrease significantly, suggesting that the heuristics usually keep on exploring as a result of the delayed performance feedback they receive. Convergence is higher when the delays are perceived correctly. Inspection of individual simulations reveals that under misperceived delays some simulations converge to subop-

Fig. 3. Average Payoff Percentage (left) and Percentage Converged (right) with different time delays for first activity, when perceived delays are all zero. In this case absolute perception error equals the delay in each case. Results are averaged over 400 simulations after 300 periods. Random allocation performance is reported as a benchmark (left)



timal policies in all the heuristics. This means that the organization can conclude that it has found the optimal payoff and stop exploring other regions of the payoff landscape, even though that allocation is in fact suboptimal and there is significant unrealized potential for improvement.

Finally, the different learning heuristics show the same qualitative patterns. They also show some differences in their performance. When delays are correctly perceived the more rational algorithms do better, but when the length of the time delays is misperceived the more rational learning heuristics do just as poorly as the more myopic heuristics. In short, a common failure mode persists independent of assumptions about the learning capabilities of the organization: when there is a mismatch between the true delay and the delay perceived by the organization, learning is slower and usually does not result in improved performance. The simulated organization frequently concludes that an inferior policy is the best it can achieve and settles on equilibrium allocations that yield payoffs significantly lower than the performance of a completely random strategy.

Robustness of results

Each learning model involves parameters whose values are highly uncertain. To understand the more general impacts of delays we conducted sensitivity analysis over all the important parameters of each learning procedure. In this section we investigate the robustness of the results presented above. We focus on the following questions:

- To what extent do the delay effects depend on the parameter values of the model?
- Is there any asymmetry in the effect of misperceived delays on performance? Is underestimating the length of a delay more or less damaging to performance than overestimating it?
- Is there any nonlinear effect of delays on learning? Is there a limit to the negative effects of delays on performance? Do such saturation effects exist, and, if so, at what delay levels do they become important?
- What mechanisms contribute to observed performance?

We conducted a Monte Carlo analysis, varying each of the important model parameters randomly from a uniform distribution over a wide range (a factor of four around the base case; see the complete listing and ranges in the supporting information, Table S7). We carried out 3000 simulations, using random initial allocations in each.

We investigate the results of this sensitivity analysis using regressions with three dependent variables: Achieved Payoff Percentage, Distance Traveled, and Probability of Convergence. Distance Traveled measures how far the heuristic has traveled in the action space, and is a measure of the cost of search

and exploration. We also report some illustrative graphs in the online appendix (supporting information, Figure S1).

For each of these variables and for each of the learning heuristics, we run a regression over the independent parameters in that heuristic (see the supporting information for complete listing). The regressors also include the Perception Error, which is the difference between the Payoff Generation Delay and its perceived value; Absolute Perception Error; and Squared Perception Error. Having both Absolute Perception Error and Perception Error allows us to examine possible asymmetries, for example, whether the impact of overestimating a delay differs from that of an equal underestimation. The second-order term allows us to detect nonlinearities such as saturation effects. To save space and focus on the main results, we only report the regression estimates for the main independent variables (Delay, Perception Error, Absolute Perception Error, and Squared Perception Error), even though all the model parameters are included on the right-hand side of the reported regressions. OLS is used for the Achieved Payoff Percentage and Distance Traveled; logistic regression is used for the Convergence Probability. Tables 3–5 show the regression results; summary statistics are available in the supporting information (Table S8). All the models are significant at $p < 0.001$.

The following results can be inferred from these tables:

- Increasing the absolute difference between the real delay and the perceived delay always decreases the achieved payoff (Table 3, row 4). The coefficient for this effect is large and highly significant, indicating the persistence of the effect across different parameters and learning heuristics. Absolute perception error also usually increases the cost of learning (Distance Traveled), even though the effect is not as pronounced as that of the achieved payoff (compare Table 3 and Table 4, rows 4).

Table 3. OLS regression for achieved payoff percentage

Variable/heuristic	Regression		Correlation		Myopic		Reinforcement	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1 Adj. R^2 /model DF	0.269	12	0.172	13	0.074	11	0.075	12
2 Intercept (Est., SD, significance)	91.82***	1.97	91.54***	3.16	88.26***	2.43	87.11***	2.32
3 Payoff Generation Delay (a1)	0.027	0.068	-0.29**	0.11	-0.20*	0.087	-0.112	0.081
4 Absolute Perception Error	-1.86***	0.20	-4.72***	0.31	-2.59***	0.26	-2.40***	0.24
5 Perception Error	-0.70***	0.049	-0.42***	0.074	0.22**	0.062	0.034	0.057
6 (Perception Error) ²	0.025*	0.012	0.21***	0.019	0.13***	0.016	0.099***	0.015

Asterisks indicate significance at * < 0.05, ** < 0.01, and *** < 0.0001 levels.

Table 4. OLS regression for distance traveled

Variable/heuristic	Regression		Correlation		Myopic		Reinforcement	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
1 Adj. R^2 /model DF	0.695	12	0.662	13	0.686	11	0.682	12
2 Intercept (Est., SD, significance)	18.48***	0.822	20.30***	0.871	20.93***	0.852	19.81***	0.896
3 Payoff Generation Delay (a1)	0.099**	0.029	0.055	0.029	0.15***	0.031	0.12**	0.031
4 Absolute Perception Error	0.47***	0.085	0.136	0.085	0.23*	0.091	0.25**	0.093
5 Perception Error	0.016	0.020	0.019	0.020	-0.079**	0.022	-0.026	0.022
6 (Perception Error) ²	-0.014**	0.005	-0.006	0.005	-0.013*	0.006	-0.011	0.006

Asterisks indicate significance at * < 0.05, ** < 0.01, and *** < 0.0001 levels.

- Delay alone appears to have much less influence on payoff. The regression model is insensitive to correctly perceived delays and the rest of the models show only a small, though statistically significant, decline in performance (Table 3, row 3).
- Three of the models show asymmetric behavior in the presence of misperceived delays. For Regression and Correlation, underestimating the delays is more detrimental than overestimating them (Table 3, row 5) while myopic is better off underestimating the delays rather than overestimating them. In all models the absolute error has a much larger effect than perception error itself (Table 3, comparing rows 4 and 5).
- There is a significant nonlinear effect of misperceived delays on performance (Table 3, row 6). This effect is in the expected direction, with higher perception errors associated with smaller incremental effects on learning. The effect saturates (the second-order effect neutralizes the first-order effect) when the delay perception error is around 14 (except for the regression heuristic, which has a much smaller saturation effect).
- Convergence is influenced both by delay and by error in perception. Higher delays, even if perceived correctly, make it more difficult for the heuristics to converge (Table 5).

These results are in general aligned with the analysis of delays offered in the last section and highlight the robustness of the results. Failure to account for delays between resource allocations and their impact causes suboptimal performance, slows learning, and raises the cost of learning.

How delays impede learning

The core of the challenge of learning in complex environments is the credit assignment problem (Samuel, 1957; Minsky, 1961); that is, the challenge of

Variable/heuristic	Regression			Correlation			Myopic			Reinforcement						
	594 Est.	707 SD	707 Sig.	401 Odds r.	401 Est.	367 S.D.	367 Sig.	492 Odds r.	492 Est.	535 S.D.	535 Sig.	528 Odds r.	528 Est.	600 S.D.	600 Sig.	600 Odds r.
2 Chi-sq./no. converged	-0.690	0.429	0.108	0.952	-0.122	0.017	<0.0001	0.885	-0.074	0.016	<0.0001	0.929	-0.070	0.0157	<0.0001	0.932
3 Estimate/SD/Odds r./sig.	-0.049	0.015	0.001	0.931	-0.150	0.051	0.003	0.861	-0.122	0.049	0.012	0.885	-0.143	0.0487	0.0033	0.867
4 Payoff Generation Delay (a1)	-0.071	0.047	0.130	0.956	-0.009	0.012	0.448	0.991	0.040	0.011	0.000	1.041	0.014	0.0113	0.207	1.014
5 Absolute Perception Error	-0.045	0.011	<0.0001	0.997	0.006	0.003	0.051	1.006	0.006	0.003	0.055	1.006	0.00417	0.00301	0.166	0.004
6 Perception Error	-0.003	0.003	0.348	0.997	0.006	0.003	0.051	1.006	0.006	0.003	0.055	1.006	0.00417	0.00301	0.166	0.004
7 (Perception Error) ²																

Table 5. Logistic regression for convergence probability

making appropriate connections between actions and payoffs. The introduction of delays, and particularly misperceptions of their duration, causes error in the association of actions and payoffs and results in the organization's inability to learn the true relationships between actions and outcome and hence improve performance through experience. It is important, however, to specify precisely how the misperception of delays degrades learning and performance.

One important contributing process is the tradeoff between short-term and long-term results in resource allocation. Increasing investment in the activity with longer delays necessarily causes a short-term decline in performance, even if the payoff improves in the long term. Given fixed resources, larger investment in the long-term activity entails a reduction of investment in the short-term activity and therefore hurts the payoff immediately. If the organization does not appreciate the existence of the delay in the payoff to the long-term activity, or underestimates its duration, this "worse-before-better" dynamic can cause the organization to learn the wrong lessons from its experience, specifically, concluding that the long-term alternative is not beneficial because it hurts performance.

For a concrete example, consider a firm facing a production shortfall. Pressuring employees to work harder generates a short-run increase in output, as front-line workers reallocate time from improvement to production. The resulting decline in productivity and equipment uptime, however, comes only after a delay. It appears to be difficult for many managers to recognize and account for such delays, so they conclude that pressuring their people to work harder is the right thing to do, even if it is in fact harmful. Over time such attributions become part of the organization's culture and have a lasting impact on firm strategy. Repenning and Sterman (2002) show how, over time, managers develop strongly held beliefs that pressuring workers for more output is the best policy, that the workers are intrinsically lazy and require continuous supervision, and that process improvement is ineffective or impossible in their organization—even when these beliefs are false. Such convergence to suboptimal strategies as a result of delays in payoff constitutes an avenue through which the temporal complexity of the payoff landscape can create heterogeneity in firm performance (Levinthal, 2000).

Another impediment to learning concerns the timeliness of data about the payoff landscape. Gradient-based methods, like the correlation heuristic used here, use action and payoff information to estimate the slope of the payoff landscape and then move uphill to higher performance. However, in the presence of delays, even when their duration is correctly perceived, information about the slope informs the agent about a previous location on the landscape, which may not characterize the gradient at the organization's current location. The result is often overshoot and oscillation in the organization's resource allocations. Note that the potential for oscillation arises even when the organization knows the duration of the delays and accounts for them correctly in

assessing the action–outcome gradient. Delays (technically, phase lag) in the negative feedbacks through which the organization adjusts its activities to improve performance amplify the potential for oscillation. The longer the delays, or the higher the gain of the negative feedback (that is, the stronger the organization's response to information about the slope of the payoff landscape), the less stable the system will be. In many of the simulations, the system is sufficiently unstable that performance does not converge to the optimum; rather, the organization continually oscillates around the peak in the payoff landscape.

A third dynamic concerns premature convergence to a suboptimal allocation. Explorations (random changes in resource allocations) are designed to prevent such problems by (1) generating information about the direction to move in order to improve performance and (2) avoiding becoming stuck on a local peak or plateau in performance space. Exploration, however, is costly, because it entails deliberate deviations from the decisions currently thought to be superior. As in many real settings, the extent to which the simulated organization is willing to explore depends on the perceived return to exploration (Eq. 15). When performance is changing quickly, the organization concludes that it is not close to the optimal resource allocation and hence that the return to continued exploration is high; the magnitude of explorations is large. When performance is not changing significantly, the organization concludes that exploration will not yield much improvement and so is not worth the cost; the magnitude of explorations falls. Given the smooth, single-peaked payoff function we assume, exploration always generates information about the direction in which the organization should move to improve performance—provided the delays are correctly perceived. When delays are misperceived, however, the outcome feedback the organization receives does not provide a consistent guide to improve performance. Some of the exploratory moves the organization makes lower performance because the information it receives about the connection between actions and outcomes is erroneous. The organization experiences a lower return to exploration because performance does not improve as quickly as it would if the delays were correctly understood. The organization therefore cuts back on exploration prematurely, causing convergence to a suboptimal strategy. In essence, the misperception of the delays means the organization cannot reliably find ways to improve performance, which induces the belief that exploration to discover better allocations has low payoff, which then confirms the belief that no improvement is possible. This self-reinforcing learning error explains the considerable fraction of simulations that converge to low-performing strategies in the misperceived delay condition. The strength of this self-reinforcing trap depends strongly on the organization's persistence in the exploration of a new policy. The “activity noise correlation time”, δ , determines the persistence of the organization in exploring a new direction. In the presence of misperceived delays, high values of δ can allow the organization to explore a new direction long enough to see

the benefits of the long-term activities, and therefore to learn a more useful lesson from the experience. To illustrate, consider the manufacturing plant seeking to find the right balance between production and improvement activity. In such a context, the introduction of a quality improvement program constitutes an exploration: a deliberate deviation from current practice implemented in the hope that it will improve performance. If the organization has low persistence in its explorations (low δ), then it will try the new quality program for only a short while before evaluating the results. Because the short-run impact of increased investment in process improvement is a reduction in performance, the organization will likely conclude that the quality improvement program does not work (or does not work in this organization) and abandon it. Such dynamics are well documented, and lead to the “flavor of the month” of organizational fads well known in the quality movement (Oliva *et al.*, 1998; Keating *et al.*, 1999; Repenning and Sterman, 2002). A more persistent organization (longer δ) would have a better chance of discovering the true, long-run benefits of improvement and be less likely to abandon a promising program.

Too much persistence, however, can be counterproductive. Persisting in an exploration that is truly a poor idea can lead the organization to regions in which performance is so low that recovery is difficult. For example, the myopic heuristic does not benefit from high δ values because, lacking gradient information to indicate the direction in which to move to higher ground, the persistence of explorations often takes the organization to the low-payoff borders of the payoff landscape, which hurts performance and slows learning. The regression, correlation, and reinforcement heuristics do not suffer from this problem to the same extent because they receive information about the direction in which to move to higher performance, and thus persistence (longer δ) helps them overall by providing more reliable information about the benefits of different regions.

Other algorithm-specific mechanisms further influence the observed results. For example, the reinforcement algorithm converges to an allocation policy once the activity value stock ($V_j(t)$) has grown very large due to past reinforcement (Eq. 6) and therefore small exploratory moves can no more shift the relative value of different activities significantly. Consequently we observe relatively higher convergence levels for the reinforcement algorithm. In the reinforcement learning literature this problem is usually handled by increasing the rate of “forgetting” in the dynamics of the Activity Values (reducing the parameter τ_j in Eq. 6).

Similarly, past experience accumulates over time in the regression model. The growing database used by the regression model makes it harder for later data points to alter the estimated location of the performance peak, and therefore the regression heuristic is also prone to higher convergence fractions, including convergence to suboptimal allocations (see Figure 3). Discarding older data would help with this problem, but would also increase the standard errors of the estimates of the optimal allocations. Finally, part of the poor

performance of the regression algorithm in the presence of misperceived delays arises from the measurement error that is introduced in the independent variables ($\bar{A}_j(t)$) by the misperception of delays. Such measurement biases lead to biased parameter estimates and therefore suboptimal allocation (Greene, 2000).

Discussion

The results suggest that learning is slow and ineffective when organizations misperceive the delays between taking actions and realizing their full consequences. One might argue that this is hardly surprising because the organization's model of the underlying task is misspecified. A truly rational organization would not only seek to learn about better allocations, but would also seek to test its assumptions about the true temporal relationships between actions and their impacts; if initial beliefs about the delay structure were wrong, experience should reveal the problem and lead to a correctly specified model.

Such second-order learning can be captured through defining distinct discrete states for different potential combinations of past actions, and updating the value of those states based on immediate payoff, as well as the value of states that are stepping stones to the current condition. For example, consider a discrete time model of an organization with only two actions: invest and harvest.⁵ Investing in one period, followed by a harvest, leads to positive payoff, while other combinations of past and present action are ineffective. The organization can keep track of what actions it has taken in the previous periods, and update their value, i.e., by assigning an increasing value to invest-harvest sequences. It will then learn the value of taking the action "invest", after the "harvest", because "invest" becomes a stepping stone for realizing invest-harvest combination later. Such a framework can be applied to building robust algorithms for larger problems and longer delays (as long as the maximum delay length is bounded and known), from playing checkers to training robots (Sutton and Barto, 1998). We do not include the analysis of such second-order learning here; further details can be found elsewhere (Fang and Levinthal, 2008; Rahmandad, 2008), and in the online appendix. Nevertheless, such learning is likely to be difficult for real organizations.

First, estimating delay length and distributions is difficult and requires substantial data; in some domains, such as the delay in the capital investment process, a principal concern in macroeconomic theories of investment, growth, and business cycles, it took decades for consensus about the length and distribution of the delay to emerge (see Serman, 2000, Ch. 11 and references therein). The results of an experimental study by Gibson (2000) are illuminating. In a simple task (one degree of freedom) with delays of one or two periods, Gibson shows that individuals had a hard time learning over 120 trials. A

simulation model designed to learn about delays and fitted to human behavior learns the delay structure only after ten times more training data.

Second, the experimental research suggests people have great difficulty recognizing and accounting for delays, even when information about their existence, length, and content is available and salient (Sterman, 1989a, 1989b, 1994). In our setting there are no direct cues to indicate the length, or even the presence, of delays.

Third, the outcome feedback the decision maker receives may be attributed to changes in the payoff landscape and noise in the environment, rather than to a problem with the initial assumptions about the delay structure. The decline in convergence and learning speed should be the main cues from which the organization could deduce that its assumptions about the delay structure are erroneous. However, in the real world, where the payoff is not solely determined by the organization's actions, several other explanatory factors, from competitors' actions to luck, are likely to be invoked to explain low performance.

Our results are robust to a variety of assumptions concerning rationality. The effects of time delays persist over a wide range of model complexity and rationality in the different learning heuristics we examined. Of course, we do not claim that no learning algorithm can be designed to account for delays. In fact, a rich literature in machine learning and optimal control has developed algorithms for learning, many of which are effective in particular environments (e.g., see Bertsekas, 2007). However, many of these algorithms are more sophisticated than the heuristics people use; in particular, they do not suffer from the many systematic errors and biases documented in human judgment and decision making (Kahneman *et al.*, 1982). Errors in beliefs about the lengths of delays substantially hinder learning even in an environment highly favorable to learning: in our experiments the payoff landscape is smooth, has a single peak, and is unchanging; furthermore, we assume no noise in action–payoff relationships. In the real world, payoff landscapes are typically rugged, possess many local peaks, and change over time; action–payoff relationships include many sources of stochastic variation. The results suggest that, for the range of learning heuristics and opportunities for experimentation realistically available to organizations, delays between actions and payoffs can become a significant barrier to learning.

Our research also contributes to the system dynamics literature by introducing four different learning models that can be applied to different problems. Historically, researchers in system dynamics often model learning through formulations that represent learning as the gradual convergence to some underlying value. For example, it is common to assume that the true value of a variable, whether customer orders or the marginal productivity of labor, is not known, but discovered gradually, and this adjustment is often modeled via information delays. The presumption is that the decision makers will reliably learn the true value of the cue, almost always with a fixed adjustment time. In

practice, however, the underlying values are not always known and may change as part of the learning process. This study provides alternative formulations that capture the learning process itself. Because we formulate the learning heuristics for the general case of continuous time, these formulations can easily be used in any model to represent individual and organizational learning from experience. Documented models are provided online to assist with such applications.

This study focused on resource allocation as the task of interest and assumed a fixed budget constraint. In reality, however, the total pool of resources available to organizations is itself dynamic. How might the results change if the simulated organization did not face a fixed resource constraint? A fixed budget constraint may strengthen the negative impact of delays on learning because it forces a strong tradeoff between short-term and long-term results. If managers could invest freely in each of the activities, they could simplify the learning problem by doing controlled experiments, increasing investment in one activity while holding constant the investment in others. Such a setting might enhance learning and moderate the suboptimal allocations and slow learning we observe. In reality, of course, organizations do not have unlimited resources. Budgets depend on organizational performance and success in the marketplace, which in turn depend on the organization's ability to learn, introducing another positive feedback that could strengthen, rather than weaken, the negative impact of delays on learning we observe with static resources. Exploration of this issue is left for future research.

Some of the processes leading to poor performance in the presence of delays, such as credit assignment, outdated gradient information, and rigidity arising from the accumulation of past experience, transfer to other tasks as well. Therefore we may expect the impact of delays on learning to apply to other tasks, though perhaps somewhat more weakly. Further, our study deliberately omits considerations such as process noise, information reporting delays and measurement error, all of which are likely to make learning more difficult, particularly in the presence of delays. More research is needed to address alternative ways delays impact upon learning in different types of tasks.

Extensions of this research can take several directions and address the shortcomings of the current study. First, there is room to develop more realistic learning models that capture the real-world challenges facing organizations. Possible extensions include more realistic payoff landscapes (e.g., modeling the plant explicitly) and going beyond simple stimulus–response models (e.g., including the formation of cognitive maps). There may be important interactions between the challenges associated with learning on rugged landscapes and learning in the presence of delays. It is possible to investigate the effect of delays on learning in the presence of measurement and perception delays for the payoff, which we assumed to be zero. The results should be tested against empirical cases of learning failure. Finally, it is important to look at the feasibility of second-order learning, i.e., learning about the delay structure.

Our research also has important practical implications. The results highlight the importance of employing time horizons substantially longer than the delays embedded in current strategies for evaluating performance. This is important both in designing incentive structures inside the organization as well as in the evaluation of firm performance by the capital markets. For example, Hendricks and Singhal (2001) show that stock prices do not fully incorporate the positive effects of total quality management programs on firm performance at the time the information about the quality initiative becomes available to the public.

The results of our analysis suggest that learning can be significantly hampered when the delays between actions and payoffs are not correctly taken into account. In fact, even with relatively short delays, a wide range of learning heuristics fail to surpass the performance of a random allocation policy. The results are robust over a wide range of parameters and for a wide range of variation in the sophistication of the information processing and rationality in the learning heuristics. Explicitly including the time dimension in action–payoff relationships is critical in learning research and empirical work to understand learning failures and performance variability among firms.

Notes

1. A constant returns to scale payoff function is analytically helpful as the optimum allocation policy does not change when delays are introduced (contrary to increasing returns functions), facilitating comparison across delay conditions.
2. When only one degree of freedom exists (two activities), any algorithm is bound to cross the optimal allocation when shifting from full allocation to activity one to full allocation to activity two. This characteristic masks several challenges in learning when more activities are present.
3. In practice, the budget constraint (resources available to the organization) is variable, and partly a function of organizational performance. Introducing endogenous resources, however, confounds the learning dynamics central to this paper with resource dynamics. We therefore leave this extension to future research.
4. These heuristics only cover a subset of algorithms for modeling learning in the presence of delays. A host of more complex, information-intensive algorithms have been developed in machine learning and control theory. (For reviews see Bertsekas and Tsitsiklis, 1996; Sutton and Barto, 1998; Bertsekas, 2007). We also experimented with one of these algorithms, Q-learning, but its informational and computational requirements are too high to be realistic for the type of organizational setting we consider. Details of the results with Q-learning are available in the supporting information.

5. This concept is much simpler than the continuous-time, continuous-action space model analyzed here.

Supporting information

Supporting information may be found in the online version of this article.

Acknowledgements

We thank Faison Gibson, Jerker Denrell, the Associate Editor and two anonymous reviewers of *System Dynamics Review*, and seminar participants at MIT, the 2003 NAACSOS conference, the 2003 AOM conference, and the 2002 ISDC for helpful comments.

References

- Argote L, Epple D. 1990. Learning-curves in manufacturing. *Science* **247**(4945): 920–924.
- Argyris C, Schön DA. 1978. *Organizational Learning: A Theory of Action Perspective*. Addison-Wesley: Reading, MA.
- Bertsekas DP. 2007. *Dynamic Programming and Optimal Control*. Athena Scientific: Belmont, MA.
- Bertsekas DP, Tsitsiklis JN. 1996. *Neuro-dynamic Programming*. Athena Scientific: Belmont, MA.
- Busemeyer JR, Swenson KN, Lazarte A. 1986. An adaptive approach to resource-allocation. *Organizational Behavior and Human Decision Processes* **38**(3): 318–341.
- Cyert RM, March JG. 1963. *A Behavioral Theory of the Firm*. Prentice-Hall: Englewood Cliffs, NJ.
- DeGeus AP. 1988. Planning as learning. *Harvard Business Review* **66**(2): 70–74.
- Denrell J, March JG. 2001. Adaptation as information restriction: the hot stove effect. *Organization Science* **12**(5): 523–538.
- Denrell J, Fang C, Levinthal DA. 2004. From T-mazes to labyrinths: learning from model-based feedback. *Management Science* **50**(10): 1366–1378.
- Einhorn HJ, Hogarth RM. 1985. Ambiguity and uncertainty in probabilistic inference. *Psychological Review* **92**(4): 433–461.
- Erev I, Roth AE. 1998. Predicting how people play games: reinforcement learning in experimental games with unique, mixed strategy equilibria. *American Economic Review* **88**(4): 848–881.
- Fang C, Levinthal D. 2008. The near-term liability of exploitation: exploration and exploitation in multi-stage problems. *Organization Science* (published online before print September 17, 2008, DOI: 10.1287/orsc.1080.0376).
- Forrester J. 1969. *Urban Dynamics*. MIT Press: Cambridge, MA.
- Gavetti G, Levinthal D. 2000. Looking forward and looking backward: cognitive and experiential search. *Administrative Science Quarterly* **45**(1): 113–137.

- Gibson FP. 2000. Feedback delays: how can decision makers learn not to buy a new car every time the garage is empty? *Organizational Behavior and Human Decision Processes* **83**(1): 141–166.
- Greene WH. 2000. *Econometric Analysis*. Prentice-Hall: Upper Saddle River, NJ.
- Hendricks KB, Singhal VR. 2001. The long-run stock price performance of firms with effective TQM programs. *Management Science* **47**(3): 359–368.
- Herriott SR, Levinthal D, March JG. 1985. Learning from experience in organizations. *American Economic Review* **75**(2): 298–302.
- Kahneman D, Slovic P, Tversky A. 1982. *Judgment under Uncertainty: Heuristics and Biases*. Cambridge University Press: Cambridge, U.K.
- Keating E, Oliva R, Reppenning N, Rockart S, Sterman JD. 1999. Overcoming the improvement paradox. *European Management Journal* **17**(2): 120–134.
- Leonard-Barton D. 1992. Core capabilities and core rigidities: a paradox in managing new product development. *Strategic Management Journal* **13**: 111–125.
- Levinthal DA. 1997. Adaptation on rugged landscapes. *Management Science* **43**(7): 934–950.
- Levinthal D. 2000. Organizational capabilities in complex worlds. In *The Nature and Dynamics of Organizational Capabilities*, Dosi G, Nelson R, Winter S. Oxford University Press: New York; 363–379.
- Levinthal D, March JG. 1981. A model of adaptive organizational search. *Journal of Economic Behavior and Organization* **2**(4): 307–333.
- Levinthal DA, March JG. 1993. The myopia of learning. *Strategic Management Journal* **14**: 95–112.
- Levitt B, March JG. 1988. Organizational learning. *Annual Review of Sociology* **14**: 319–340.
- Lipe MG. 1991. Counterfactual reasoning as a framework for attribution theories. *Psychological Bulletin* **109**(3): 456–471.
- Lord RG, Maher KJ. 1990. Alternative information-processing models and their implications for theory, research, and practice. *Academy of Management Review* **15**(1): 9–28.
- March JG. 1991. Exploration and exploitation in organizational learning. *Organization Science* **2**(1): 71–87.
- Meadows DH, Randers J, Meadows DL, Behrens WW. 1972. *The Limits to Growth: A Report for the Club of Rome's Project on the Predicament of Mankind*. Universe Books: New York.
- Milgrom P, Roberts J. 1990. The economics of modern manufacturing: technology, strategy, and organization. *American Economic Review* **80**(3): 511–528.
- Minsky M. 1961. Steps toward artificial intelligence. *Proceedings of the Institute of Radio Engineers* **49**: 8–30.
- Nelson RR, Winter SG. 1982. *An Evolutionary Theory of Economic Change*. Belknap Press of Harvard University Press: Cambridge, MA.
- Oliva R, Rockart S, Sterman JD. 1998. Managing multiple improvement efforts. In *Advances in the Management of Organizational Quality*, Vol. 3, Fedor DB, Gough S. JAI Press: Stamford, CT; 1–55.
- Raghu TS, Sen PK, Rao HR. 2003. Relative performance of incentive mechanisms: computational modeling and simulation of delegated investment decisions. *Management Science* **49**(2): 160–178.

-
- Rahmandad H. 2008. Effect of delays on complexity of organizational learning. *Management Science* **54**(7): 1297–1312.
- Repenning NP, Sterman JD. 2001. Nobody ever gets credit for fixing problems that never happened: creating and sustaining process improvement. *California Management Review* **43**(4): 64.
- Repenning NP, Sterman JD. 2002. Capability traps and self-confirming attribution errors in the dynamics of process improvement. *Administrative Science Quarterly* **47**: 265–295.
- Samuel A. 1957. Some studies in machine learning using the game of checkers. *IBM Journal of Research and Development* **3**: 210–229.
- Senge PM. 1990. *The Fifth Discipline: The Art and Practice of The Learning Organization*. Currency Doubleday: New York.
- Sengupta K, Abdel-Hamid TK, Bosley M. 1999. Coping with staffing delays in software project management: an experimental investigation. *IEEE Transactions on Systems Man and Cybernetics Part A* **29**(1): 77–91.
- Stata R. 1989. Organizational learning: the key to management innovation. *Sloan Management Review* **30**(3): 63–74.
- Sterman JD. 1989a. Misperception of feedback in dynamic decision making. *Organizational Behavior and Human Decision Processes* **43**: 301–335.
- Sterman JD. 1989b. Modeling managerial behavior: misperceptions of feedback in a dynamic decision making experiment. *Management Science* **35**(3): 321–339.
- Sterman JD. 1994. Learning in and about complex systems. *System Dynamics Review* **10**(2–3): 91–330.
- Sterman J. 2000. *Business Dynamics: Systems Thinking and Modeling for a Complex World*. Irwin, McGraw-Hill.
- Sterman JK, Repenning NP, Kofman F. 1997. Unanticipated side effects of successful quality programs: exploring a paradox of organizational improvement. *Management Science* **43**: 503–521.
- Sutton RS, Barto AG. 1998. *Reinforcement Learning: An Introduction*. MIT Press: Cambridge, MA.
- Walsh JP. 1995. Managerial and organizational cognition: notes from a trip down memory lane. *Organization Science* **6**(3): 280–321.
- Warren HC. 1921. *A History of the Association Psychology*. C. Scribner: New York.
- Watkins C. 1989. *Learning from delayed rewards*. PhD thesis, King's College, Cambridge, U.K.