

Near-Term Liability of Exploitation: Exploration and Exploitation in Multistage Problems

Christina Fang

Department of Management and Organization, Stern School of Business, New York University,
 New York, New York 10012, cfang@stern.nyu.edu

Daniel Levinthal

Department of Management, The Wharton School, University of Pennsylvania,
 Philadelphia, Pennsylvania 19104, levinthal@wharton.upenn.edu

The classic trade-off between exploration and exploitation reflects the tension between gaining new information about alternatives to improve future returns and using the information currently available to improve present returns. By considering these issues in the context of a multistage, as opposed to a repeated, problem environment, we show that exploratory behavior has value quite apart from its role in revising beliefs. We show that even if current beliefs provide an unbiased characterization of the problem environment, maximizing with respect to these beliefs may lead to an inferior expected payoff relative to other mechanisms that make less aggressive use of the organization's beliefs. Search can lead to more robust actions in multistage decision problems than maximization, a benefit quite apart from its role in the updating of beliefs.

Key words: exploration and exploitation; maximization; multistage problems; reinforcement learning; softmax choice rule

History: Published online in *Articles in Advance* September 17, 2008.

1. Introduction

The exploration and exploitation trade-off has been a central part of the discourse in organization studies stemming from March's (1991) original work and has a long history in decision sciences (DeGroot 1970), computer sciences (Holland 1975), and economics (Radner and Rothschild 1975). Decision makers need to make trade-offs between gaining new information about alternatives to improve future returns and using the information currently available to improve present returns (March 1991). Excessive exploitation of one's current knowledge in an effort to maximize immediate returns may yield an underinvestment in search and discovery and the development of new beliefs. However, the exploration and exploitation trade-off has been examined in a particular class of problem environments, where the choice situation is repeated over time and, critical to our present argument, an observed performance outcome is associated with each decision.

We consider the merit, and limits, of exploitive behavior in the context of multistage decision making in which a given decision need not lead to any tangible outcome but rather sets the stage for subsequent decision-making. Our results highlight an important facet of the familiar classic trade-off between exploration and exploitation. While there is a clear understanding in the management literature that exploitation of one's current beliefs may be dysfunctional in the long run, we show that, in a

multistage problem, exploitation can lead to an immediate decline in payoffs and not just a possible long-run penalty from insufficient learning about latent alternatives. In contrast, a decision policy that is mildly, but not strongly, exploitative is superior to an explicit maximization of perceived payoffs even in the *near* term.

As broad context for understanding our results, it is important to recognize that any conscious choice is logically preceded by another activity: representation and abstraction. Such representations or abstractions typically capture simplified features of the real-world problem in question and describes abstract relations of objects in the form of engineering diagrams, flow charts, chemical formulae, etc. (Simon 1966). These abstractions are "small worlds," structures that are built on a variety of simplifying assumptions (Savage 1954). Seen from this perspective, analytical efforts to identify the optimal solution to a decision problem, only identify the optimal solution to a model of that situation (Einhorn and Hogarth 1981). Decision makers choose and act on a representation of a real-world problem. When decision makers' representation does not align well with the real problem, as is likely in a multistage decision process, exploitation of the existing partial representation leads to lower payoffs on average than a less exploitative strategy in which perceived payoffs are not maximized. Less exploitative strategies are shown to lead to a more robust approach to problem solving in a multistage setting. Furthermore, this misalignment is not a question of

biases, but rather is a reflection of the fact that a representation may provide only a partial characterization of the problem. As such, this inferiority of exploitation should not be an anomalous property in complex problem environments.

This finding supplements our current understanding of the classic trade-off between exploration and exploitation. Exploration is frequently associated with revising beliefs and changing one's underlying model of the problem environment (Levinthal and March 1993, March 1991, Holland 1975). However, we show that the robustness of exploration in a multistage, as opposed to repeated, problem does not depend on the updating of beliefs; instead, it depends on the identification of "clues" or signals of value associated with intermediate steps in the multistage problem. While the classic trade-off between exploration and exploitation hinges on the distinction between making full use of current beliefs and enhancing one's understanding to make better choices in the future, we show that, in a multistage setting, even in the absence of any updating of beliefs, a more exploratory decision rule can be advantageous.

Our finding further points to a possible critique of maximization approaches. The question of whether individuals choose among actions so as to maximize their payoff has been a long-standing one in the social sciences (Savage 1954, Simon 1955, Friedman 1953). Empirical research on decision making has documented substantial deviations from optimal benchmarks, and critiqued the assumption of maximization on the grounds of behavioral realism (Simon 1955, Kahneman and Tversky 1973, March 1978). We argue that, to examine the normative validity of maximization as a choice mechanism, it is first necessary to examine the specific mechanism of representation. The way in which problems are formulated has much to do with the quality of the solutions that are found as the algorithms of choice applied to these representations (Simon 1986). If the representation captures faithfully the relevant features of the problem, then maximizing choices with respect to this representation will yield optimal results in the actual problem context. However, if the representation is not accurate, then no such assurance exists. The efficacy of maximization is therefore logically contingent on the model or representation of the problem.

2. Solving Multistage Problems

As Simon (1990, p. 7) noted, "human rational behavior is shaped by a pair of scissors whose two blades are the structure of task environments and the computational capabilities of the actor." Holding the computational capabilities of the actors constant, we explore problem settings in which actors need to develop more or less complex mental representations of their problem environment (Thagard 1996), which then serve as the

basis of their choices. To make salient the role of actors' representations of the environment, we focus on a class of multistage decision problems, where the ultimate payoff is only realized after a series of intermediate stages. In these settings, actions and outcomes are separated across stages. As such, the consequence of an action at one point may not be felt until many stages later.

To illustrate the challenge in a multistage problem, consider an individual learning how to solve Rubik's Cube (i.e., get all sides of a multifaceted cube to be of the same color). In trying to do so, the individual faces two challenges. First, no immediate objective information is available about whether individual moves bring the individual closer to the solution. Thus an individual has to make appropriate moves without being informed if they are good or bad. Second, even when the solution has been found, it is seldom obvious whether specific moves were good or bad. In particular, evaluating individual moves often requires recognition of their long-term implications. Yet, developing such understanding is difficult even with repeated experience. For example, most novices believe that getting one side correct is a useful subgoal. However, experts know that a move that achieves this state is not a step toward the ultimate goal and may constitute movement away from it.

This problem setting is characteristic of many organizational decision-making contexts. The very notion of strategic decisions revolves around the fact that this class of decisions has longer term consequences (Andrews 1965). In an organization, activities are often sequentially interdependent (Thompson 1967), where choices made in earlier stages influence the return to choices made in later stages. For instance, consider the product development efforts in an organization. A long sequence of activities has to take place, ranging from identifying a promising technology, allocating resources to development, testing of prototypes, large-scale manufacturing to marketing of the final products. This process often takes months, if not years, before market response is known. During the process, even though various indicators may be used to evaluate progress, there is no guarantee that these indicators are sufficiently correlated with the final outcome. In such a setting, learning what might constitute more or less favorable choices by an upstream activity is often confounded by the fact that the outcome of those choices are mediated by choices made in downstream activities and vice versa. Earlier activities "set the stage" for later ones. For instance, what the marketing department does to encourage the sales of a new product is very much enhanced or constrained by the prior initiatives of the R&D group, and the sort of product development efforts that prove fruitful may depend on the marketing initiatives that lie further downstream.

Consider further the innovation journey in biotechnology. Even after a drug candidate has been identified, a lengthy process of trial and error is still needed to

determine whether it is safe or effective for humans. A long sequence of downstream activities related to toxicology, process development, formulation design, clinical research, biostatistics, regulatory affairs, and finally marketing needs to unfold before the first product revenues can be realized (Pisano 2006). This lengthy process not only takes about a decade but also carries with it significant risks, because historically only one out of 6,000 synthesized compounds makes it to the market, and only 10%–20% of drug candidates beginning clinical trials are, ultimately, approved for commercial sale.

In both examples, there are two distinct temporal processes at work. One process is the standard process of feedback and experiential learning as “trials” are repeatedly carried out across time, leading to the revision of beliefs and improvement in performance and learning (Argote 1999). The other temporal process, a much less discussed one, is associated with a multistage decision. Decisions at one stage set up what might constitute more or less attractive actions at a subsequent stage. The processes, of course, may be linked. For instance, even though the idiosyncratic search for a specific new drug (e.g., a specific ulcer medication) is unlikely to be repeated exactly again, general lessons for how best to organize the multistaged drug discovery efforts are still learned. How best should the scientific teams be organized? What level of testing should be carried out before the drug is taken to full-scale clinical trials (clearly, there are regulatory constraints here as well as firm choices)? To what degree should a market analysis inform the selection among candidate research efforts? What are the best approaches to market introduction (direct advertising, linking with key professionals, etc.)? These are examples of the general lessons that can be learned in a multistage decision process.

The challenge in this class of multistage problems, known as the credit assignment problem in artificial intelligence (Minsky 1961, Samuel 1967, Holland 1975), is how to develop representations of the problem or milestones (Block and MacMillan 1985) to evaluate actions that may not have immediate outcome consequences. From a rational choice perspective, this challenge is captured by a dynamic programming framework (Bellman 1957). Formally, a set of dynamic programming techniques can be used to specify optimal action at each point in time, taking fully into account the potential stage-setting value of each action. In this way, an optimal path can be derived recursively. However, this formal approach poses formidable and often overwhelming computational burdens for their solution (Simon 1992). Even if the full decision tree can be carefully specified, computation of the dynamic programming solution can be infeasible as the space of possible futures expands exponentially with possible actions. Bellman (1957) himself acknowledges this as the “curse of dimensionality.” For instance, Deep Blue, the first computer chess

program to win a chess game against a reigning world champion in 1997, does not push through the full space of possibilities and identify an optimal action. Despite a computing power capable of evaluating 200 million positions per second, Deep Blue is still dwarfed by the enormous, though finite, state space in chess. There are, for example, more than 288 billion different possible positions after four moves. Rather, Deep Blue employs a sophisticated heuristic known as “selective deepening,” searching moves deemed *interesting* or *promising* far more deeply (Hsu 2002). In short, rational maximization solutions to multistage problems are at best elusive and unrealistic.

However, with a few notable exceptions (Brehmer 1995, Gibson et al. 1997, Sterman 1989), behavioral models of learning generally assume immediate, though possibly misleading, feedback about the consequences of actions (Cyert and March 1963; Lave and March 1975; Levinthal and March 1981; Levitt and March 1988; Herriott et al. 1985; Lant and Mezias 1990, 1992; Lant 1994; Roth and Erev 1995; Levinthal 1997; McKelvey 1999; Gavetti and Levinthal 2000; Rivkin 2000). In these feedback-based conceptions of learning, the primary mechanism is a process of reinforcement learning in which actions that lead to favorable outcomes are reinforced. Most real-life contexts, however, depart from this setup. In settings where there are multiple stages, as are modeled here, feedback is not immediately available. This implies that learning based solely on reinforcement or feedback (also known as hill climbing) would not prove effective (Denrell et al. 2004).

To examine the classic trade-off between exploration and exploitation in a multistage setting, we design a series of simulations that capture some of the basic properties of such contexts. In particular, we introduce a mechanism by which representations can be modeled in a multistage task, which is both behaviorally plausible and can capture the long run as well as immediate consequences of actions. In our setup, we first tune the level of accuracy and completeness of these representations by varying the length of experiential learning. We then examine the performance implications of exploiting or maximizing with respect to these beliefs. As such, while the specific mechanism of developing representations is an important component of our model (detailed in §3.2), this mechanism is nothing more than a tool to endow our organizations with a set of more or less sensible beliefs about the underlying payoff structure, and our results are not contingent upon the specifics of this mechanism.

3. Simulation Analysis

3.1. Task Setting

We model a multistage task as a performance landscape in the form of an N -dimensional landscape or hypercube (cf. Kauffman 1993, Levinthal 1997, Bruderer and Singh

1996). Each point in such a surface consists of a vector of choices associated with the N -dimensions. An $N + 1$ dimension consists of a measure of performance associated with the vector of N choices. We represent a vertex, or state S , as an N -element binary string, in which each element can take on the value of zero or one. With any given N , there are 2^N possible states. For instance, if N is set to 2, then there are 2^2 possible binary strings (namely, 00, 11, 01, 10). Associated with each binary string of choice configurations is a payoff that specifies the corresponding outcome feedback. An agent's task is to identify a sequence of actions that maximizes their payoffs over the long run. Search is local, where an action consists of flipping one element in the binary string from zero to one or vice versa.

To simulate a multistage problem, we consider a landscape in which only a subset of choice configurations yields a nonzero payoff. In other words, agents will find zero payoffs in all states except a selected few. To maximize long-term payoff, agents need to identify a path from an arbitrary starting position in the landscape to these selected configurations with nonzero payoffs. These states constitute goals or peaks in the landscape, while other states can be thought of as more or less useful antecedent steps toward them. Furthermore, these choice configurations carry differential nonzero payoffs. A reasonable measure of performance in this setting is whether agents learn to walk to peaks with higher rewards, because higher payoffs are only attained when agents identify paths from their starting points to the higher peaks.

We consider a simple scenario in which the landscape has two nonzero peaks, with one peak having a substantially higher payoff than the other. Visually, this problem can be represented by a flat surface with only two spikes, representing two states corresponding to distinct combinations of policy choices (Bruderer and Singh 1996). The challenge is how to develop representations of the problem to evaluate actions that may not have immediate payoffs.

In the next section, we detail a specific mechanism by which representations can be modeled. While the specifics of this mechanism are an important component of our model, we use it mainly as a tool to vary the accuracy and completeness of an organization's representation.

3.2. Evolution of Representations

To carefully model the mechanism by which a representation might develop, we build on one mechanism known as Q learning, in which a representation, or value function, evolves through repeated experience (Kaelbling 1993, Watkins 1989, Sutton and Barto 1998, Denrell et al. 2004).¹ This model extends the standard models of reinforcement learning (cf. Lave and March 1975

and others above) to allow actors' existing representation of the task environment to serve as an alternative basis for reinforcement learning (Samuel 1959, Sutton and Barto 1998). It explicitly simulates the evolution of representations by combining elements from both dynamic programming and feedback-based learning models. In short, dynamic programming equations are converted into simple updating rules, and strategies that produce success over time are reinforced more than those that do not produce success over time. This specific learning mechanism was initially developed in the field of machine learning as a means to identify solutions to optimal control problems (Sutton 1998, Sutton and Barto 1998). It has gained recent recognition in psychology (Berthier et al. 2005, Sutton and Barto 1981) and has produced results consistent with the robust property of the "law of effect" (Thorndike 1898), which has been observed in a large literature within experimental psychology. Recently, support has been found in neural science (Daw and Dayan 2004, McClure et al. 2003, O'Doherty et al. 2004) following a discovery in the primate ventral tegmental area (the front mid-brain) of neurons whose firing closely resembles the predicted patterns (Schultz et al. 1997, Montague et al. 1996).

First, we model representation in a very simple stylized manner by a state action value function known as a $Q(s, a)$ function (Watkins 1989), following the Q-learning method. A $Q(s, a)$ function captures an agent's beliefs about the immediate reward for taking action a in state s , as well as the long-run consequences of that choice. In our model, agents search locally, where an action consists of flipping from zero to one or vice versa. Given $N = 10$, we have 11 possible actions or columns: 10 actions to flip each one of the N elements and the 11th action being staying put. A $Q(s, a)$ function is implemented as a simple look up table with rows corresponding to all possible states (2^N) and columns corresponding to all possible actions ($N + 1$). As such, our state action space is a table with 1,024 rows and 11 columns, resulting in 11,264 cells. Each cell in this table contains a corresponding a $Q(s, a)$ value, which captures agent's beliefs about how much value can be generated from taking action a starting from state s .

Second, following Samuel (1959), we assume that an agent updates this representation successively over time, gradually uncovering the underlying problem structure (Kaelbling 1993, Watkins 1989, Sutton and Barto 1998, Denrell et al. 2004). Over time, the table is updated to reflect agents' evolving knowledge about the underlying problem. In particular, whenever a peak is found, the immediately preceding state and action pair (s, a) is positively updated because it has "led" to the peak. Each episode concludes with the agents walking to either the higher or lower peak. During subsequent episodes, this

particular (s, a) serves as a useful stepping stone or sub-goal to the goal states such that any other (s, a) pair, which “leads” to this (s, a) is also positively updated. In this way, “credit” is successively assigned to more and more distant (s, a) pairs.

Specifically, agents update their current representations of $Q(s, a)$, by weighting their beliefs about the current state s as well as future state s' by parameters α and γ as follows:

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

R represents the immediate payoffs from taking action a starting from a state s . The parameter γ weights the importance of $Q(s', a')$, the value of taking an appropriate action a' from a future state s' . It regulates the degree to which the prior state action pair, $Q(s, a)$, gets “credit” for the position (i.e., state s') that it has created. 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. The parameter α represents weights given to existing beliefs $Q(s, a)$, in relation to $Q(s', a')$. We set α at 0.8 and γ at 0.9.

Initially, all $Q(s, a)$ values are specified to be zeroes or uninformed. This “flat” prior belief structure is chosen because it does not discriminate among alternative state action pairs. As such, in the first trial, agents will choose actions randomly. Suppose that an agent has carried out action a , which was identified as desirable based on the actor’s current beliefs. Taking this action, the agent arrives at a new state s' . This new state may or may not provide some immediate payoff R . Regardless of whether an immediate reward is received, this new state s' is now a launching point for further actions. In other words, the *current* state s should be given some “credit” for helping to create a valuable *future* position. It follows that the perceived value of current state s is the sum of two terms: (1) an immediate payoff R and (2) perceived values of $Q(s', a')$, where s' is the future state that can be reached from s . If $Q(s', a')$ is positive, then an action that leads an agent from current state s to s' must be valuable as well. When a peak is found (by chance), however, R becomes positive. The immediately preceding state action pair, (s, a) , will therefore be positively updated and the action responsible for bringing the agent from state s to the peak will also be recognized as valuable. In subsequent trials, whenever the agent comes to this particular state s again, she knows that by taking that particular action a , she will find the peak as before. In this way, updating of the agent’s representation occurs not only when the solution is found, but also whenever the agent reaches the positively valued $Q(s', a')$ identified in the earlier trials. Over repeated experience, as distant state action pairs are discovered and updated positively, more and more long-term consequences of actions are compounded into the $Q(s, a)$ function.

This evolving $Q(s, a)$ function serves as the basis of agent’s actions. Because there are multiple actions available at each state, we need to tune the extent to which agents are sensitive to the different magnitudes of different actions available at the same state. On the one hand, agents can “maximize” by always choosing the action with the highest $Q(s, a)$ value. On the other hand, agents can explore by choosing an action that is “good,” yet not necessarily the “best.” Clearly, these alternative choice rules represent two extremes along a continuum, which varies along the extent to which an agent is sensitive to any differences in $Q(s, a)$ values associated with the various possible actions in a given state s . While agents who maximize are very sensitive to any difference in $Q(s, a)$ values, agents who randomize choose equally among all actions and decide independently of the $Q(s, a)$ values. To tune this sensitivity, we model the probability of choosing a given action using a decision rule developed by Luce (1959) that has been used widely in estimating learning models (cf. Camerer and Ho 1999, Gans et al. 2007, Weber et al. 2004). Choice is highly structured and heuristically guided by the current $Q(s, a)$ as follows:

$$\frac{\exp\{Q(s, a)/\tau\}}{\sum_{a=1}^n \exp\{Q(s, a)/\tau\}}. \quad (2)$$

A single parameter τ (from 0 to infinity), or “temperature,” regulates how sensitive the probability of choosing a given action is to the estimated $Q(s, a)$ values of alternative actions in the corresponding state. It operates by differentially choosing among actions, favoring those with higher $Q(s, a)$ values, which are perceived to be more attractive. For any positive τ value, the probability of selecting an action with a higher $Q(s, a)$ is greater than the probability of selecting an action with a lower $Q(s, a)$. In addition, the higher the value of τ , the probability of a given action being chosen will be less sensitive to the relative differences in the $Q(s, a)$ values. Actions will be chosen more uniformly and an action with lower $Q(s, a)$ values still has a positive probability of being chosen. Agents with high τ essentially do not take their own $Q(s, a)$ value very seriously, and frequently depart from what they believe to be the optimal behavior. On the other extreme, small τ values cause choices to become very sensitive to the estimated values of the various alternative actions. Agents with lower τ adhere closely to their representations, carrying out the best actions as dictated by their $Q(s, a)$ values. In the limit, as $\tau \rightarrow 0$, if an action has the highest $Q(s, a)$, it will be chosen with probability 1. As such, a lower τ value is equivalent to greater exploitation of one’s current beliefs, while a higher τ generates behavior that is more disconnected from one’s beliefs. In short, a high τ value corresponds to a higher degree of search and vice versa.²

3.3. Structure of Simulation

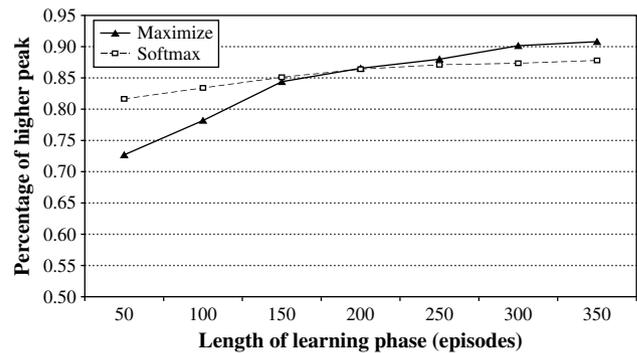
Because choice is based on an evolving representation, it is natural to design our simulation as a function of time. Early on in the simulation, organizations are presumed to sample broadly and to try alternative actions. This can be seen as a “learning phase” of trial and error search, as organizations identify attractive actions and develop a better representation of the underlying problem structure. During the learning phase, organizations explore and gradually build up their representations. As such, the longer the learning phase, the higher the quality of the representations or beliefs. We use the length of the learning phase as a tool to tune the degree of accuracy and informativeness of the organization’s representation of the problem environment. At the end of this “learning phase,” we contrast the efficacy of two alternative choice strategies: one in which agents maximize or exploit by locking into a preferred strategy and the other in which agents act in a way less constrained by their beliefs about what constitutes appropriate behavior.³ In this latter case, agents retain some degree of skepticism toward their own representations. As compared with fully exploiting their current beliefs, their choices are “softened” as they continue to search by deviating from perceived optimal actions. They are therefore less aggressive in using their current representations. In keeping with prior literature (Sutton and Barto 1998), we label this alternative strategy as “softmax.” In other words, agents follow either maximization or softmax as choice strategies. Exploration during the learning phase is approximated by setting τ to 20. At the end of the learning phase, maximization is approximated by setting τ to 0.1 and “softmax” is implemented by keeping τ at the same level of 20 as before. With this formal structure, we now examine the near-term performance implication of exploitation by varying systematically the length of the “learning phase.”

The results in the following analysis are based on the average behavior of organizations more than 2,000 independent histories of search. Each “history” comprises a “learning phase” of varying number of episodes and a subsequent episode in which we contrast the performance of organizations following the two alternative choice strategies.

3.4. Results

Recall that, in our problem setting, choices and outcomes are separated across time. Even though most actions are not associated with any immediate payoffs, they constitute more or less useful antecedent steps toward the two peaks. Furthermore, higher payoffs are only attained when organizations find a sequence of actions that leads to the higher, rather than lower peak. To examine the implications of alternative choice strategies, we measure performance by simply whether organizations learn to walk to peaks with higher payoffs.⁴

Figure 1 Performance Implications of Two Alternative Choice Strategies



In Figure 1, we contrast the number of times the higher peak is found by organizations that follow either “maximization” or “softmax.” Performance is measured after the end of the “learning phase,” which varies from 50 to 350 episodes. For instance, after 50 learning episodes, organizations carry out different choice strategies based on their representation developed in these 50 episodes. If the length of the learning phase is longer, agents are allowed more episodes to experiment and develop their representations, which then serve as the basis of their choice.

In Figure 1, we see that maximization systematically under or over performs relative to the baseline benchmark of “softmax,” depending on the length of the learning period. When the length of the “learning phase” is relatively short, maximization underperforms “softmax.” As more time is allowed to learn, the gap in performance narrows gradually. Eventually, maximization outperforms the benchmark. In short, we observe a maximization discount, followed by a maximization premium. Thus, a positive relationship exists between the relative efficacy of maximization and the length of the learning phase. Again, the length of the learning phase is a mechanism simply to tune the accuracy and informativeness of the organization’s beliefs. Thus, these results indicate that with relatively poor or incomplete beliefs, fully exploiting one’s beliefs by making choices based on what appears to be the best alternative will tend to reduce payoffs compared to a less aggressive use of these beliefs. However, with more informed beliefs, the relationship reverses itself. For instance, we see that after 50 episodes of learning, agents who maximize based on their representation identify the superior peak, on average, 73% of the time, while those who follow “softmax” and continue to operate with a higher τ value, have an average performance of 82%. After 150 episodes of learning, the size of this performance discount reduces to 1%.⁵

What underlies this maximization “discount”? Given that maximization is with respect to an evolving representation, the efficacy of maximization may depend

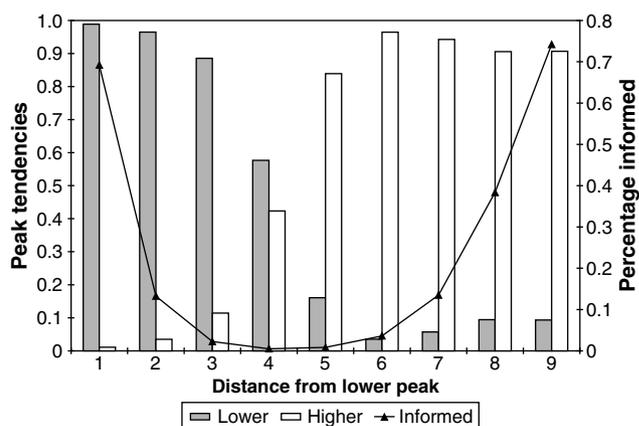
critically on the adequacy of the actor's representation. As such, it might be reasonable to expect that after even 50 episodes of exploration and experimentation, a decision maker's model of the problem is still fragmentary, inadequate, and perhaps misleading in important ways.

To see whether this is the case, we first need to develop an intuitive understanding of the mechanism by which a representation evolves in this problem setting. As mentioned earlier, the problem landscape consists of two spikes standing on a flat plain. Because there are no local cues serving as feedback, organizations cannot rely on any local gradient to guide them in their search for either one of the two peaks. To find paths from arbitrary starting points to the peaks, organizations have to construct their own gradient in their representation. This is done by gradually recognizing more and more antecedent states that were reached in prior steps. Over time, two slopes form (one from each peak) and cross each other at a dividing line, which can be termed a watershed. Just as small streams flow into different basins, depending on which side of an incline they lie, organizations are guided toward different peaks depending on the side of the watershed in their $Q(s, a)$ functions. More formally, the watershed characterizes all states whose $Q(s, a)$ values manifest a dominant tendency to walk toward one peak or another. It effectively divides the landscape into two separate regions of attraction. The exact location of the watershed depends on a set of parameters characterizing the nature of the updating process (α , and γ), the degree to which the choice process corresponds to maximizing behavior (as determined by τ), and the relative magnitudes of the rewards at the two peaks. For a broad set of these values, the resulting "watershed" appears at a point intermediate between the two peaks, and, in particular, is located more distantly from the higher peak. To achieve the higher payoff associated with the higher peak, organizations have to increase their odds of "walking" to the left of the watershed.

To gauge the adequacy of an organization's representation, we examine the average representation of 2,000 organizations to see whether it provides a correct direction for agents at each location. In Figure 2, we plot the percentage of states in which the dominant beliefs (i.e., maximum $Q(s, a)$ value) dictate a tendency to step toward one peak or the other at varying distance⁶ from the lower peak. Because it takes time for organizations to learn, and not all states may have been experienced, organizations may not have learned anything around certain locations. As such, we also plot the percentage of states with informed beliefs at various distances. We define an organization's beliefs about a state s to be informed if there is at least one action a such that the corresponding $Q(s, a)$ value is strictly positive.

As seen in Figure 2, organizations' average representation is incomplete in two important ways. First, a large

Figure 2 Quality of Representation at Various Locations in the Problem Space

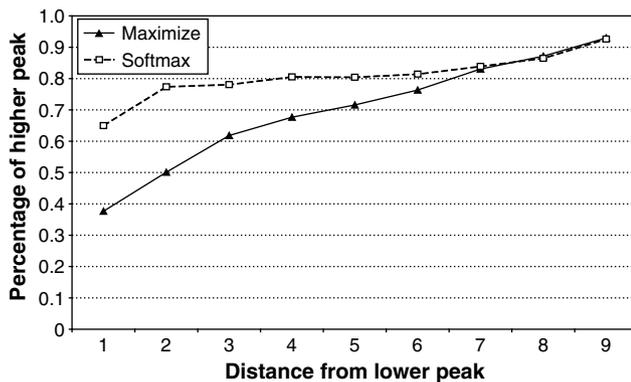


percentage of states across all distances are still uninformed, as organizations either fail to learn anything in those states or simply have not visited them. Thus they will not choose intelligently in these states. Second, we see clearly the presence of a watershed. In this case, states that are five steps away constitute a critical divide, because for these sets of states, the $Q(s, a)$ function is such that organizations are inclined to walk to the higher peak rather than the lower peak. These states delimit the watershed in the organization's representations. Whenever an organization arrives at this demarcation, or to its right, it will be guided to walk toward the higher peak, thereby attaining higher performance. However, this also implies that representations could be potentially misleading, especially if organizations find themselves in states that fall on the opposite side of the watershed. For instance, if we look at states that are two steps away from the lower peak, the predominant tendency is still to walk to the lower peak even after 50 episodes of learning.

However, incompleteness in the underlying representation cannot be the only explanation. Organizations' performances seem to differ, depending on their choice strategy, given the *same* incomplete representation. A more skeptical (or probabilistic) use of the same incomplete model systematically leads to superior performance, as indicated in Figure 1. To see a clearer mapping between the model and performance, we systematically vary the starting positions of organizations and contrast the performance of organizations, that follow either "softmax" or "maximization" strategies in Episode 51.

Given the same incomplete representation as reflected in Figure 2, we find that the potential inaccuracy in representations does not seem to pose much of a handicap for organizations following the "softmax" strategy. For instance, in Figure 3, when organizations are started three steps away from the lower peak, agents' own representation indicates that by a nine to one margin that

Figure 3 Performance Implications of Varying Starting Positions in the Problem Space



the appropriate course of action is to take an immediate step toward the lower peak. Given that this is to the left of the watershed, it would seem that the majority of organizations would walk to the lower peak. However, this is not the case. Despite these somewhat “erroneous” beliefs, more than 78% of the organizations manage to walk to the higher peak.

The basic intuition behind this apparent paradox lies again in the watershed imagery. Organizations can be thought of as traveling in a space of possible paths to two alternative solutions. The challenge is to find a path that leads to the superior solution in the absence of immediate feedback. Superior solutions offer stronger, localized outcome feedback. As a result, they tend to cast a broader shadow across representation of the problem space and push the watershed further away. On the one hand, if organizations maximize and adhere strictly to a given representation, they will be strongly guided by their beliefs as to what actions to take. They will be guided to the lower peak at least some of the time given the presence of the watershed. In contrast, under the alternative strategy “softmax,” some degree of experimentation is preserved as organizations are less sensitive to the relative magnitude of $Q(s, a)$ values. As such, they will be less constrained to carry out actions deemed to be the *best*. Rather than locking into an apparently attractive sequence of actions based on local fragmentary knowledge, organizations may wander off to areas where stronger signals of the superior solutions exist. Even if they are at the unfavorable side of the watershed, organizations increase their odds of success as if some imaginary “bridges” or shortcuts open access to the higher peak. This serves to mitigate the potential deficiency of an incomplete representation.

On the surface, such benefits of continual experimentation, in contrast to maximization, resembles closely the classic trade-off between exploration and exploitation. Exploration by gathering further information, helps avoid premature lock-in to inferior solutions (Levinthal and March 1981, Levitt and March 1988) and may be

beneficial in the long run. Our “bridging” mechanism, however, generates benefits in addition to, and independent of information gathering. To see this, we reexamine performance under the same conditions as in Figure 3, except that we turn off updating during the last episode. In this way, no further information is gathered by search and no new information is incorporated into the agent’s existing representations. As expected, performance benefits persist even in the absence of information gathering. For organizations following the softmax choice strategy, there is little difference in performance before and after information gathering is stopped. Both curves remain higher than that associated with fully exploitive behavior. As argued before, a more probabilistic use of one’s existing model generates randomness, which makes it possible to bridge the “watershed” that divides organizations’ mental landscape. This is distinct from the mechanism of information gathering. As expected, performance difference between “softmax” and maximization disappears at a higher distance (i.e., as we get closer to the higher peak). Therefore, search, at least in our context, does more than generate new information about the value of alternative states as in the classic trade-off between exploration and exploitation. It changes the starting point for the subsequent search process by building bridges or shortcuts that open access to the superior solution. As such, while this sort of sampling process also underlies the classic trade-off between exploration and exploitation, search generates benefits from a more robust use of existing (although imperfect) representations rather than from the updating of beliefs.

It is important to emphasize the flip side of maximization discount, however. For instance, in Figure 1, from 250 episodes onward, we observe a maximization *premium*. This is because after 250 episodes of learning, organizations have developed a sufficiently complete model of the underlying problem. As such, acting on such beliefs by maximization produces higher performance, whereas organizations end up engaging in excessive experimentation under the softmax approach.

3.5. Robustness Checks

The main result that maximization may lead to a performance discount holds over a range of parameter settings. To this point, we have used the length of the learning phase as a lever to manipulate the completeness of the organizations’ representations. More generally, one should consider the degree of difficulty associated with the underlying search problem. For a given degree of cognitive capabilities, the bounds of rationality will be more or less constraining, depending on the difficulty of the problem. If the problem is difficult, even sustained exploration is unlikely to generate an accurate model, and maximization becomes more hazardous. The converse property holds as problem difficulty diminishes.

Problem difficulty is contingent upon several parameters. In results not reported here, we find that as each one of these parameters increases (while holding other parameters constant), two things occur. First, the magnitude of the maximization discount decreases. Second, the point at which maximization becomes superior to softmax sets in earlier, as problem difficulty increases. Both results are expected and are consistent with our main findings. First, problem difficulty is influenced by N , the number of dimensions of the problem space. The higher the number of dimensions, the more challenging is the problem, holding other things constant. Second, difficulty is influenced by where organizations start their adaptive search relative to the problem and solutions. If one starts near a favorable solution, even in a large problem space, the challenge of identifying a sequence of steps to the solution is greatly mitigated. Therefore, if organizations are started closer to the lower peak, they face a more difficult problem of locating the other higher peak.

In addition, the informativeness of one's mental model is a function of the temperature parameter τ during exploration. Given the same problem, higher τ implies a higher degree of search and, consequently, agents will have developed more robust belief structures covering a wider array of possible actions and states. However, as τ becomes very high (e.g., around 30), softmax approximates random choice, and does not enable the organization to hold on to states or actions that have been discovered to be valuable. They do not "act on" those potentially superior beliefs. As a result, as τ increases beyond 30, maximization always outperforms softmax.

Furthermore, we find that our qualitative result is also robust to an alternative performance measure that incorporates search cost. Suppose that a cost is incurred with each move, then the relevant performance measure becomes the net payoff organizations receive after subtracting the total search costs. Find that the addition of move costs generates well-known dynamics that are also intuitive. In results not reported here, we find that search costs do not seem to impact the percentage of visits to the higher peaks. However, given that our new performance measure takes into account the efficiency of search, we find that as costs increase, maximization outperforms softmax both earlier and by a greater magnitude. This dynamic is well known and expected, because organizations try to avoid higher search costs, and therefore become more "exploitative" in their behavior. Nevertheless, the qualitative pattern holds.

One might imagine another sort of constraint on the search process in the form of a budget constraint.⁷ As a stylized representation of this, consider an organization that engages in softmax search behavior, but is subject to a budget. Once the budget is exhausted, the organization is forced to switch to the mode of fully exploiting its current beliefs or simply "gives up." We explore this alternative search constraint by keeping track of the number

of zero-payoff locations (defined either as the absence of external reward or as absence of positive $Q(s, a)$ values) the organization visits. Once a predetermined number of such states are reached, the organization reverts to fully exploitive behavior. A reasonable conjecture is that such a search policy would yield results intermediately between that of softmax and maximization and exploitation. We find that this conjecture holds at moderate to high levels of the search budget. At one extreme (i.e., high levels of budget), as expected, performance approaches that of softmax because the high budget is unlikely to be constraining, and therefore behavior would simply correspond to softmax. However, at very low budget levels, such a policy, in fact, yields performance results that are inferior to both policies. The intuition behind this result is that an organization is most likely to reach its budget constraint in the region where it has the least informed beliefs, which will generally correspond to the area intermediate between the two peaks and solutions. As a result, budget constrained organizations are likely to switch to exploitation precisely in those states where exploratory nonmaximization behavior would be most useful. Exploratory search in the middle of the problem space can help push the organization into the basins of attractions of the superior solution. In contrast, to maximize in the region with the least informed beliefs is problematic when the quality of beliefs is poor. As a consequence, when budget is very low, budget constrained organizations actually perform more poorly than the two pure forms of which this search process is composed.

Finally, it is important to note that while we believe that the Q-learning mechanism provides the appropriate modeling of representation in a multistage setting, our results are not contingent upon this specific mechanism. We have used Q learning to generate sensible beliefs and to vary their accuracy and completeness. Our analysis then examines the efficacy of maximization with respect to these beliefs. The qualitative pattern of our results are not specific to Q learning, and in results not reported, holds when we use other alternative mechanisms such as an ϵ -greedy strategy (Sutton and Barto 1998).

4. Discussion and Conclusion

Our results highlight an important facet of the familiar classic trade-off between exploration and exploitation (March 1991, Holland 1975). While there is a clear understanding in the management literature that exploitation of one's current beliefs may lead to a focus on near-term payoffs (Levinthal and March 1981) and proven alternatives (Levitt and March 1988), we show that, in a multistage problem, exploitation can lead to an immediate decline in payoffs and not just a possible long-run penalty from insufficient learning about latent alternatives. We show that a decision policy that

is mildly, but not strongly, exploitative is superior to an explicit maximization of perceived payoffs even in the *near* term. Less than full maximization leads to a robust approach to problem solving in a multistage setting. Furthermore, this robustness does not depend on the updating of beliefs; instead, it depends on the identification of “clues” or signals of value associated with intermediate steps in the multistage problem. Even in the absence of any updating of beliefs, a more exploratory decision rule can be advantageous.

However, it is important not to overinterpret these results. The systematic performance enhancement as a result of maximization is significant when representations are well aligned with reality, indicating possible dangers of excessive experimentation. We want to emphasize here the possibility of problematic consequence of maximization, not its inevitability. In particular, it is undoubtedly possible to rerun a simulation model with different parameter settings and different model formulations so as to produce a “better” picture for maximization. Such demonstrations, however, would not eliminate the possible dangers associated with maximization, which is central to our existence statement.

Our result that exploitation leads to immediate performance decline fits well with several related findings in the complementary disciplines of artificial intelligence and cognitive psychology. First, it is well known in the work on complex systems and the artificial intelligence literature (Kauffman 1993, Selman et al. 1994) that, on complex surfaces (e.g., rugged landscapes with many hills and valleys), the presence of many local optima make the application of local search methods problematic. Because a local search algorithm starts from a candidate solution and considers only alternatives solutions in the neighborhood of the current one, it often results in an adaptive system being trapped in the wrong hill. As such, additional mechanisms are needed to counterbalance this tendency. One such mechanism is randomness, or noise. The reason is that, in a complex landscape, there are multiple “attractors” (or hills) to which the system migrates over time. These attractors each commands a basin of attraction, a region of the problem space within which the system will settle toward the associated attractor. Given multiple attractors and basins of attraction, an adaptive system often exhibit sensitivity to initial conditions. In other words, small differences in starting values could result in the system settling into different basins of attractions, and therefore, different attractors. Yet, the amount of sensitivity to initial conditions is not uniform throughout space. The sensitivity is less in the basin of the attractor and more at the edge of the basin (Guastello 1995).

Noise, or randomness, capitalizes on this structural property of a complex landscape and allows an attractor to expand to its fullest range (Breedon et al. 1990; Jackson 1991a, b; Ohle et al. 1990). For a system that

is trapped on a local peak, noise introduces random movements away from the existing attractors. Precisely because the sensitivity to initial conditions is not uniform, the introduction of noise, on average, favors the discovery of superior attractors whose bases of attraction are more extensive. Indeed, studies have shown that the addition of randomized moves significantly improves the performance of a variety of local search algorithms (Selman and Kautz 1993, Selman et al. 1994, Fukunaga et al. 2004).

Second, our result shares another interesting and useful connection with studies of problem solving in cognitive psychology. Newell and Simon (1972) first introduced the concept of human problem solving as a search through a problem space. Problem spaces can be represented by two different extremes (Perkins 2000). On the one hand, simple problems have “homing” spaces, in which all contours lines demarcate a region where the solution can be found. On the other hand, complex problems are characterized by problem spaces in which the goal is buried in the midst of clues that are irrelevant, misleading, or without any clear direction. To successfully navigate through this kind of landscape, one needs a “precipitating event” that leads the search process to escape from traps and irrelevant cues (Perkins 2000). Random search, in this context, makes it possible to break through an existing boundary in the problem space.

Furthermore, supporting evidence of the “bridging” effect of exploration can be found in discussions of incubation, which refers to taking a break while struggling to solve a problem. Incubation may eventually speed up the solution process (Wallas 1926) and its benefits have been attributed to unconscious processing (Poincare 1929, Campbell 1960, Simonton 1995) and integration of external cues (Langley and Jones 1988, Yaniv and Meyer 1987). Recent evidence, however, suggests that no processing activity takes place during the break. Rather, the break’s only function is to divert the solver’s attention away from the problem, thus releasing the mind from the grip of a false organizing assumption (Kaplan and Simon 1990, Segal 2004, Seabrook and Dienes 2003, Simon 1966). Even though no new information is gathered and processed, the interruption allows decision makers to see the problem with “fresh eyes,” a mechanism similar to “bridging.”

Randomness in search produces performance benefits by capitalizing on the structural properties of the underlying representations. This echoes closely with our analysis. Randomness does more than generate new information about unsampled spaces. In a context in which one has useful but fragmentary knowledge about the world, exploration changes the starting point for the subsequent search process. It serves as an important robustness check and facilitates further examination of the fragments of existing knowledge in complex problem

spaces. While representations provide a powerful guide to action, they should be best viewed as “general” guidance rather than a specific sequence of behavior to be rigidly followed.

This image of tempered use of one’s beliefs about optimal action is central to the notion of robust action (Leifer 1991, Padgett and Ansell 1993). Leifer (1991), drawing from DeGroot’s (1965) classic studies of expert chess players, finds some striking examples of the value of not strictly adhering to a model of optimal behavior. DeGroot (1965) asked players to articulate two consecutive moves in a balanced chess situation. After a player selected her first move, the experimenter responded with an opponent’s move that had been anticipated by the players as expressed in their pregame analysis. If a player were confident in her strategy choice, she should move in an anticipated way by executing the previously declared strategy. Indeed, it took less expert players, on average, only a few seconds to execute their second moves. The most expert players such as grandmasters, however, took some 20 minutes to make the second move, despite the fact that the opponent’s move corresponded exactly to what they had predicted. A positive relationship was found between the skill level of the players and the amount of time they took to respond in their second move. Leifer (1991) suggests that expert players engage in local action that preserves opportunities⁸ and continually reevaluated appropriate strategies, despite a well-articulated attack plan. Only when favorable opportunities became clear do skilled players switched to an “intentional” mode and the pursuit of *ex ante* strategies (Leifer 1991, p. 66).

In a more natural setting, our analysis on the exploitation discount may provide a useful conceptual backdrop to the observation of the productivity paradox in the literature on process management. Consistent with our results, empirical evidence suggests that efforts at rationalizing an underlying business process, such as ISO 9000 certification, increasing yields, or reducing defects, are often linked to an immediate decline in financial performance (Serman et al. 1997, Garvin 1991) and new product innovations (Benner and Tushman 2002). In part, such perverse relationships may emerge as a result of a dysfunctional pursuit of subgoals that only imperfectly capture the path to the firm’s ultimate profitability. Yield, or product defects, may be important quality measures, but their link to firm performance need not be straightforward. Consistent with our analysis, Benner and Tushman (2003) find that it is important to sustain some degree of variability in the firm’s processes to be effective in new product developments.

Clearly, much work remains to flesh out the contingencies regarding the relative superiority of alternative choice mechanisms in different problem environments and belief structures. For instance, in a one-shot decision context, such as modeled in two- and *n*-armed bandit problems, maximization on the basis of an unbiased

set of beliefs will always provide the highest expected payoff in the current period. However, in the problem context modeled here, actions often take one to another point in the problem space and not to some ultimate payoff generating state. The multistage nature of the problem context impacts directly how performance is defined and how differential choice strategies are evaluated. In addition, we have implicitly assumed that problems are encountered with sufficient regularity, so that there is the opportunity for learning to accumulate across episodes. It would be interesting, in future work, to explore intermediate cases between such idiosyncratic histories and the fixed surface examined here. Intelligence in such settings would require the ability to generalize across prior distinct, but related, problem settings or domains. For instance, analogical reasoning is a powerful form of generalizing across experiences (Gavetti et al. 2005).

To conclude, intentionally rational decision-making hinges on both an organization’s model of their decision environment and the choice operator that is applied to these beliefs. As a research community, we have perhaps underattended to the interplay between these two considerations. Unbiased, but incomplete, representations may result in settings in which choice mechanisms that deviate from maximization are superior. We have shown in such settings where choice has an inherent dynamic quality, robust action may be a superior choice strategy to more fully exploitive behavior, even when measured in terms of immediate performance outcomes. More broadly, these results point to the limits to the normative superiority of maximization approaches in the face of incomplete representations of the problem context.

Acknowledgments

The authors are grateful for comments on a prior draft by Jerker Denrell, Anne Marie Knott, Nicolaj Siggelkow, Sidney Winter, and participants at the 2002 Academy of Management Conference in Denver. The authors also have benefited from the comments of three reviewers and the thoughtful critiques of Associated Editor Henrich Greve. All errors remain the authors’ own.

Endnotes

¹Denrell et al. (2004) examine how credit assignment can be approached when there is a single outcome using the Q-learning method. The issue we examine here of the possible inferiority of maximization as a decision strategy could not occur in the context of the single-peak structure considered in Denrell et al. (2004). Our work is concerned with how agents can learn to differentiate among multiple alternative outcomes.

²Other possible alternative strategies exist. For instance, one could use ϵ -greedy strategy by selecting the action with the maximum value most of the time, but to chose randomly among the remaining options with a small probability (ϵ). Recent experimental evidence suggests that softmax accounts for the pattern of subjects’ exploration behavior better than the ϵ -greedy strategy (Daw et al. 2006, Lee 2006). In results not

reported here, we implement the ϵ -greedy strategy and results are qualitatively the same.

³We use the terms maximize and exploit interchangeably here as exploitation corresponds to taking actions, which maximize payoffs according to the organization's beliefs.

⁴This performance measure implicitly assumes that there is no cost of moving in the problem space. In the robustness section, we introduce moving costs, which reduce the payoffs agents can attain. As we will discuss later, while this reduces the magnitude of the effects, the qualitative properties remain.

⁵Some readers may argue that we are measuring performance differently from how organizations themselves evaluate performance. An organization stop search whenever either peak is found and assign positive value to identifying the inferior peak (in our baseline model, the value of the inferior peak is set to 10 and that of the superior peak to 50), yet we measure performance outcomes as favorable only when the organization gets to the higher peak. We have analyzed a more general stochastic setup in which the two peaks generated rewards with different probabilities, but given the stochastic reward structure the agent cannot immediately infer whether the inferior or superior peak was reached by observing a single reward. For instance, suppose each peak yields a high or low payoff with probability p_1 and $(1 - p_1)$, respectively, for the superior peak and p_2 and $(1 - p_2)$ for the inferior peak. One peak is superior to the other to the extent that p_1 exceeds p_2 . However, experiencing a single outcome with either peak does not indicate which peak is superior. While we have run this stochastic version of the model and find similar results, we have chosen to present, for simplicity, a deterministic version (essentially with $p_1 = 1$ and $p_2 = 0$).

⁶More precisely, distance in a multidimensional space is known as hamming distance. It is a simple count of the number of different digits present in two strings to be compared. In other words, it measures the number of substitutions required to change one into the other. For instance, the hamming distance between 00001 and 00000 is 1; while the hamming distance between 11111 and 00000 is 5.

⁷The authors thank Associate Editor Henrich Greve for making this suggestion.

⁸For this reason, Padgett and Ansell (1993) use the term robust action.

References

- Andrews, K. 1965. *The Concept of Corporate Strategy*. Irwin Homewood, IL.
- Argote, L. 1999. *Organizational Learning: Creating Retaining and Transferring Knowledge*. Springer-Verlag, New York.
- Bellman, R. 1957. *Dynamic Programming*. Princeton University Press, Princeton, NJ.
- Benner, M., M. Tushman. 2002. Process management and technological innovation: A longitudinal study of the photography and paint industries. *Admin. Sci. Quart.* **47** 676–706.
- Benner, M., M. Tushman. 2003. Exploitation, exploration and process management: The productivity dilemma revisited. *Acad. Management Rev.* **28**(2) 238–257.
- Berthier, N., M. Rosenstein, A. Barto. 2005. Approximate optimal control as a model for motor learning. *Psych. Rev.* **112**(2) 329–346.
- Block, Z., I. MacMillan. 1985. Milestones for successful venture planning. *Harvard Bus. Rev.* **63**(5) 184–190.
- Breeden, J., F. Dionkelacker, A. Huber. 1990. Noise in the modeling and control of dynamical systems. *Physical Rev. A* **42** 5827–5836.
- Brehmer, B. 1995. Feedback delays in complex dynamic decision tasks. P. Frensch, J. Funke, eds. *Complex Problem Solving, The European Perspective*. Erlbaum Associates, Hillsdale, NJ.
- Bruderer, E., J. Singh. 1996. Organizational evolution, learning and selection: A genetic algorithm-based model. *Acad. Management J.* **39** 1322–1329.
- Camerer, C., T. H. Ho. 1999. Experience-weighted attraction learning in normal form games. *Econometrica* **67** 837–874.
- Campbell, D. T. 1960. Blind variation and selective retention in creative thought processes. *Psych. Rev.* **67** 380–400.
- Cyert, R., J. March. 1963. *A Behavioral Theory of the Firm*. Prentice-Hall, Englewood Cliffs, NJ.
- Daw, N., P. Dayan. 2004. Matchmaking. *Science* **304** 1753–1754.
- Daw, N., J. O'Doherty, P. Dayan, B. Seymour, R. Dolan. 2006. Cortical substrates for exploratory decisions in humans. *Nature* **441**(15) 876–879.
- DeGroot, A. D. 1965. *Thought and Choice in Chess*. Mouton Publishers, The Hague.
- DeGroot, M. H. 1970. *Optimal Statistical Decisions*. McGraw-Hill, New York.
- Denrell, J., C. Fang, D. Levinthal. 2004. Learning from t-mazes to labyrinths: Learning from model-based feedback. *Management Sci.* **50**(10) 1366–1378.
- Einhorn, H., R. Hogarth. 1981. Behavioral decision theory: Processes of judgment and choice. *J. Accounting Res.* **19**(1) 1–32.
- Friedman, M. 1953. *Essays in Positive Economics*. University of Chicago Press, Chicago.
- Fukunaga, A., G. Rabideau, S. Chien. 2004. Robust local search for spacecraft operations using adaptive noise. *Proc. 4th Internat. Workshop Planning Scheduling Space*, Darmstadt, Germany.
- Gans, N., G. Knox, R. Croson. 2007. Simple models of discrete choice and their performance in bandit experiments. *Manufacturing Service Oper. Management* **9**(4) 383–408.
- Garvin, D. 1991. How the baldrige award really works. *Harvard Bus. Rev.* **69**(6) 80–93.
- Gavetti, G., D. Levinthal. 2000. Looking forward and looking backward: Cognitive and experiential search. *Admin. Sci. Quart.* **45**(1) 113–137.
- Gavetti, G., D. Levinthal, J. Rivkin. 2005. Strategy making in novel and complex worlds: The power of analogy. *Strategic Management J.* **26** 691–712.
- Gibson, F., M. Fichman, D. Plaut. 1997. Learning in dynamic decision tasks: Computational model and empirical evidence. *Organ. Behav. Human Decision Processes* **71** 11–35.
- Guastello, S. 1995. *Chaos, Catastrophe and Human Affairs*. Lawrence Erlbaum Associates, Mahwah, NJ.
- Herriott, S., D. Levinthal, J. G. March. 1985. Learning from experience in organizations. *Amer. Econom. Rev., Papers Proc. 97th Annual Meeting* **75**(2) 298–302.
- Holland, J. 1975. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI.
- Hsu, F. H. 2002. *Behind Deep Blue: Building the Computer that Defeated the World Chess Champion*. Princeton University Press, Princeton, NJ.

- Jackson, E. 1991a. On the control of complex dynamic systems. *Physica D* **50** 341–366.
- Jackson, E. 1991b. Controls of dynamic flows with attractors. *Physical Rev. A* **44** 4839–4853.
- Kaelbling, L. 1993. *Learning in Embedded Systems*. MIT Press, Cambridge, MA.
- Kahneman, D., A. Tversky. 1973. On the psychology of prediction. *Psych. Rev.* **80** 237–251.
- Kaplan, C. A., H. A. Simon. 1990. In search of insight. *Cognitive Psych.* **22** 374–419.
- Kauffman, S. 1993. *The Origins of Order*. Oxford University Press, New York.
- Langley, P., R. Jones. 1988. A computational model of scientific insights. R. Sternberg, ed. *The Nature of Creativity: Contemporary Psychological Perspectives*. Cambridge University Press, New York, 177–201.
- 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, Hillsdale, NJ.
- Lant, T., S. Mezas. 1990. Managing discontinuous change: A simulation study of organizational learning and entrepreneurship. *Strategic Management J.* **11**(4) 147–179.
- Lant, T., S. Mezas. 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 Social Sciences*. Harper & Row, New York.
- Lee, D. 2006. Best to go with what you know? *Nature* **441**(15) 822–823.
- Leifer, E. 1991. *Actors as Observers: A Theory of Skill in Social Relationships*. Garland, New York.
- Levinthal, D. 1997. Adaptation on rugged landscapes. *Management Sci.* **43**(7) 934–950.
- Levinthal, D., J. March. 1981. A model of adaptive organizational search. *J. Econom. Behav. Organ.* **2** 307–333.
- Levinthal, D., J. March. 1993. The myopia of learning. *Strategic Management J.* **14** 95–112.
- Levitt, B., J. March. 1988. Organization learning. *Annual Rev. Sociol.* **14** 319–340.
- Luce, R. 1959. *Individual Choice Behavior*. John Wiley and Sons, New York.
- March, J. 1978. Bounded rationality, ambiguity and the engineering of choice. *Bell J. Econom.* **9**(2) 587–608.
- March, J. 1991. Exploration and exploitation in organization learning. *Organ. Sci.* **2**(1) 71–87.
- McKelvey, B. 1999. Avoiding complexity catastrophe in co-evolutionary pockets: Strategies for rugged landscape. *Organ. Sci.* **10**(3) 294–321.
- Minsky, M. 1961. Steps towards artificial intelligence. *Proc. Inst. Radio Engineers* **49**(1) 8–30.
- Montague, P. R., P. Dayan, T. J. Sejnowski. 1996. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *J. Neuroscience* **16** 1936–1947.
- McClure, S., N. Daw, P. Montague. 2003. A computational substrate for incentive salience. *Trends in Neurosciences* **26** 423–428.
- Newell, A., H. Simon. 1972. *Human Problem Solving*. Prentice-Hall, Englewood Cliffs, NJ.
- O’Doherty, J., P. Dayan, J. Schultz, R. Deichmann, K. Friston, R. Dolan. 2004. Dissociable roles of ventral and dorsal striatum in instrumental conditioning. *Science* **304** 452–454.
- Ohle, F., F. Dinkelacker, A. Huber, M. Welge. 1990. Adaptive control of chaotic systems. Technical Report CCER-90-13, Department of Physics, Beckman Institute, University of Illinois, Urbana-Champaign.
- Padgett, J., C. Ansell. 1993. Robust action and the rise of the medici, 1400–1434. *Amer. J. Sociol.* **98**(6) 1259–1319.
- Perkins, D. 2000. *The Eureka Effect: The Art and Logic of Breakthrough Thinking*. W. W. Norton and Company, New York.
- Pisano, G. 2006. Can science be a business? *Harvard Bus. Rev.* (October) 1–12.
- Poincaré, H. 1929. *The Foundations of Science*. Science House, New York.
- Radner, R., M. Rothschild. 1975. On the allocation of effort. *J. Econom. Theory* **10** 358–376.
- Rivkin, J. W. 2000. Imitation of complex strategies. *Management Sci.* **46**(6) 824–844.
- Roth, A., I. Erev. 1995. Learning in extensive form games: Experimental data and simple dynamic models in the intermediate term. *Games and Econom. Behav.* **8** 164–212.
- Samuel, A. 1959. Some studies in machine learning using the game of checkers. *IBM J. Res. Development* **32** 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.
- Savage, L. 1954. *The Foundations of Statistics*. John Wiley and Sons, New York.
- Schultz, W., P. Dayan, P. R. Montague. 1997. A neural substrate of prediction and reward. *Science* **275** 1593–1599.
- Seabrook, R., Z. Dienes. 2003. Incubation in problem solving as a context effect. *Proc. 25th Meeting Cognitive Sci. Soc.*, Lawrence Erlbaum Associates, Mahwah, NJ.
- Segal, E. 2004. Incubation in insight problem solving. *Creativity Res. J.* **16**(1) 141–148.
- Selman, B., H. Kautz. 1993. An empirical study of greedy local search for satisfiability testing. *Proc. Amer. Assoc. Artificial Intelligence*, 46–51.
- Selman, B., H. Kautz, B. Cohen. 1994. Noise strategies for improving local search. *Proc. Amer. Assoc. Artificial Intelligence*, 337–343.
- Simon, H. 1955. A behavioral model of rational choice. *Quart. J. Econom.* **69**(1) 99–118.
- Simon, H. A. 1966. Scientific discovery and the psychology of problem solving. R. Colodny, ed. *Mind and Cosmos*. University of Pittsburgh Press, Pittsburgh, 22–40.
- Simon, H. 1986. Decision making and problem solving. *Research Briefings 1986: Report of the Research Briefing Panel on Decision Making and Problem Solving*. National Academy of Sciences, National Academy Press, Washington, D.C.
- Simon, H. 1990. Invariants of human behavior. *Annual Rev. Psych.* **41** 1–19.
- Simon, H. 1992. Rational decision-making in business organizations. A. Lindbeck, ed. *Nobel Lectures, Economics 1969–1980*. World Scientific Publishing, Singapore.
- Simon, D. 1995. Foresight in insight? A Darwinian answer. R. J. Sternberg, J. E. Davidson, eds. *The Nature of Insight*. MIT Press, Cambridge, MA, 465–494.

- Sterman, J. 1989. Misperceptions of feedback in dynamic decision making. *Organ. Behav. Human Decision Processes* **43**(3) 301–328.
- Sterman, J., N. Repenning, F. Kofman. 1997. Unanticipated side effects of successful quality programs: Exploring a paradox of organizational improvement. *Management Sci.* **43**(4) 503–521.
- Sutton, J. 1998. *Philosophy and Memory Traces: Descartes to Connectionism*. Cambridge University Press, Cambridge, UK.
- Sutton, R., A. Barto. 1981. An adaptive network that constructs and uses an internal model of its world. *Cognition Brain Theory* **4**(3) 217–246.
- Sutton, R., A. Barto. 1998. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.
- Thagard, P. 1996. *Mind: Introduction to Cognitive Science*. MIT Press, Cambridge, MA.
- Thompson, J. D. 1967. *Organizations in Action*. McGraw-Hill, New York.
- Thorndike, E. 1898. Animal intelligence: An experimental study of the associative processes in animals. *Psych. Rev. Monograph Supplements* 8.
- Wallas, G. 1926. *The Art of Thought*. Harcourt, New York.
- Watkins, C. 1989. *Learning from Delayed Rewards*. Kings' College, Cambridge, UK.
- Weber, E., S. Shafir, A.-R. Blais, 2004. Predicting risk sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psych. Rev.* **11** 430–445.
- Yaniv, I., D. Meyer. 1987. Activation and metacognition of inaccessible stored information: Potential bases of incubation effects in problem solving. *J. Experiment. Psych. Learn., Memory, Cognition* **13** 187–205.