What makes dynamic strategic problems difficult? Evidence from an experimental study

Rahmandad Hazhir (Orcid ID: 0000-0002-2784-9042)

What makes dynamic strategic problems difficult? Evidence from an experimental study

Hazhir Rahmandad*
Associate Professor of System Dynamics, MIT Sloan School of Management
Room E62-442, 100 Main St., Cambridge, MA 02142
hazhir@mit.edu, +1-617-258-8912, orcid: 0000-0002-2784-9042

Jerker Denrell
Professor of Behavioral Science, Warwick Business School, University of Warwick, Scarman Road, Coventry CV4 7AL, United Kingdom.
denrell@wbs.ac.uk, +44 (24) 76522119, orcid: 000-0001-9628-1924

Drazen Prelec
Massachusetts Institute of Technology, Sloan School of Management, Department of Economics, Department of Brain and Cognitive Sciences,
Cambridge, MA, USA
Room E62-54, 100 Main St., Cambridge, MA 02142
dprelect@mit.edu, +1-617-253-2833

*: Corresponding Author
Keywords: Learning, managerial cognition, Dynamic Decision-making, Persistent Performance Differences, Experiment

Research Summary: Managers regularly deal with dynamic tasks, where decisions impact immediate payoffs as well as long-term capabilities. Research shows that people do poorly in dynamic tasks, but the underlying mechanisms are unclear. These may range from unsystematic problem-solving to rational learning in complex environments. In a series of experiments, we tease apart alternative explanations, showing that poor performance is due to behavioral difficulties. Remarkably, we find that people do poorly even if provided with complete information about the payoff function, thus eliminating any need for learning. They unsystematically search among possible solutions and end up with inefficient heuristics. The results show that differences in thinking through a dynamic problem may lead to substantial variation in performance, even if common sources of complexity and ambiguity are excluded.
Managerial Summary: Why do people, including managers, have difficulty managing systems where taking action today impacts future outcomes? Difficulty of learning in a complex environment has been proposed as the key challenge. Using experiments, we show that people find such tasks difficult even when all relevant information is provided to them and there is nothing to learn. Using trial and error most participants learn satisfactory, but inferior, heuristics. Those who

This is the author manuscript accepted for publication and has undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1002/smj.3254

systematically think through tradeoffs over time significantly outperform others even in a simple task, suggesting such thinking adds value in realistic managerial settings as well.

INTRODUCTION

Central to many strategic choices is that decisions impact not only immediate payoffs but also the payoff of future actions (Dierickx and Cool 1989; Ghemawat 1991; Leiblein et al. 2018). For example, allocating resources to learning about an emerging technology is costly but offers flexibility in future choices and enhances absorptive capacity (Cohen and Levinthal 1990), reducing future learning costs. So, learning about a new technology not only has direct costs and benefits, like any investment, but also changes the payoff that could be expected in future. Many other strategic actions similarly impact the future benefits of taking the same, or other, actions, including investments in innovation, acquisitions, capability building, and advertising. In fact, scholars have argued that such temporal interdependences are a hallmark of what makes a decision strategic (Leiblein et al. 2018). The literature in decision-making and cognitive science (e.g., Busemeyer 2002; Osman 2010) use the label *dynamic* to refer to this class of tasks. The label *dynamic* here does not refer to volatile environments but to tasks where taking an action today changes the payoff of the same or other actions in the future.

The purpose of this paper is to deepen the individual micro-foundations of strategy by studying when and why people can effectively solve dynamic tasks. Strategy researchers have increasingly taken a behavioral and cognitive approach, informed by behavioral decision-research and cognitive psychology, to understand how managers differ in making important strategic decisions (Gavetti 2012; Helfat and Peteraf 2015; Menon 2018). By using surveys and experimental methods, researchers have refined our understanding of strategically relevant topics, such as when excess entry occurs (Artinger and Powell 2016; Cain et al. 2015), when entrepreneurs are overconfident (Cassar and Friedman 2009), how biases impact corporate resource allocation (Bardolet et al. 2009; Shapira and Shaver 2014), how individuals differ in their ability to forecast new venture performance (Csaszar and Laureiro-Martínez 2018), and in their ability to profit from trade in competitive markets (Levine et al. 2017). Less explored in behavioral strategy are the challenges of dynamic decision-making.

Case studies suggest that managers may struggle with implementing the best course of action in dynamic tasks, ranging from product development to process improvement and maintenance (Lyneis and Sterman 2016; Rahmandad and Repenning 2016; Repenning and Sterman 2002). However, it is difficult to determine the cause of such behaviour from case study data. Does the blame lie with faulty incentives, which pressure managers to focus on short-term effects (Laverty 1996; Stein 1989) even if understanding the beneficial long-term effects? Alternatively, do managers have incorrect expectations about the value of investing in capabilities? Based on unlucky past experience with investments, a manager may come to underestimate the benefits of investing in capabilities (Denrell and March 2001; Levinthal and March 1993). Or, could it be that learning the trade-offs in such dynamic tasks and using this knowledge to derive the optimal policy are inherently so challenging that it vexes sophisticated managers even if incentives are aligned and expectations are unbiased? Some findings show that even rational learners who follow an optimal learning and problem solving strategy but are uncertain about how actions impact future payoffs will show poor performance for a very long period (Sims et al. 2013).

Past work has not teased apart these alternative mechanisms, but they matter for theory and practice. If incentive problems are to blame for myopic behavior, better incentive design is the remedy. If individual biases and priors explain heterogeneity in performance, that underlines the role of individual cognition in the micro-foundations of strategy, and points to potential benefits of training and de-biasing; and if poor performance is inevitable even for the best algorithms, then luck and random idiosyncrasies may be better explanatory mechanisms for understanding heterogeneity in managerial choice.

In this paper we examine whether poor performance in dynamic tasks is i) due to behavioral difficulties, including flawed learning and poor problem solving, ii) an inevitable outcome of rational learning and choices, or iii) due to computational complexity and ambiguities that render the tasks hard for sophisticated algorithms and individuals alike. To that end, we design experimental conditions that allow us to distinguish between those alternative explanations and gain a better understanding of how people actually learn in dynamic tasks. We use a simple and

established dynamic task, the Harvard Game, which offers us the required control to tease apart alternative explanations.

Our results show that poor performance cannot be explained by rational learning. People do poorly even when provided with complete information about the payoff function and the number of periods in the future that an action can affect. With this information there is nothing to learn and any analytically sophisticated agent could quickly identify the optimal strategy. People do poorly even in this complete information condition, and importantly, they show limited improvement compared to the more complex conditions that require complex learning. This suggests that poor performance in dynamic tasks may persist even if the complex learning part of the problem was removed. Exploring how people tackle the task, we find they try to solve the problem by mentally "muddling through" (Lindblom, 1959), i.e., by unsystematically trying out different action policies to see if some work better than others. While this can be an effective heuristic in some problems, in dynamic tasks people who muddle through can get stuck on local optima in the policy space. Consistent with muddling through, we find that even participants who perform well in the task show limited understanding of the true payoff structure.

Our results contribute to the behavioral view of strategy formation in organizations (Gavetti 2012; Levinthal 2011), to the importance of managerial capabilities, and to the literature on dynamic decision-making. Broadly, our results show that computational complexity, causal ambiguity, and difficulties in learning are not necessary for explaining persistent variation in performance. In a well-defined dynamic task much of the variation remains even after removing those barriers to learning. Differences in insight into, or ability to think through a dynamic problem, i.e., differences in what can be called managerial cognition (Helfat and Peteraf 2015), are sufficient to explain persistent differences in performance in rather simple dynamic tasks.

DYNAMIC TASKS AND STRATEGY

Intertemporal trade-offs in strategy

Many strategic decisions involve inter-temporal trade-offs in the sense that payoffs are distributed over time. Dynamic decision-making tasks include one type of inter-temporal trade-offs,

where taking an action today changes the payoff of the same or other actions in the future (Busemeyer 2002; Gonzalez et al. 2017). A simple investment, where a manager pays a sum today expecting a future return, is not necessarily a dynamic task if the distribution of incoming returns over time is known and the action taken today does not change the payoffs to future actions or investments.

Many classical strategic problems involve dynamic tasks. Consider the learning curve strategy: pricing low to gain a large market share and drive costs down via the learning curve. In this case, producing one item today impacts the cost of producing one item in the future, due to the learning curve. Or consider the strategy of "getting big fast": pricing low to gain large market share and increase the future value of the product through network externalities. In the age of digital platforms and machine learning, many firms invest to become a well-known hub and to accumulate data on consumers. While the basic logic for these, and many other strategies with dynamic components, are well-known, assessing their effectiveness requires solving a dynamic task. For example, Pfeffer (1994) argues that firms could profit from treating their employees better, which may reduce profits in the shorty-run but increases productivity in the long-run. But how widely this strategy applies? Could it co-exist with other firms ignoring their employees and driving labor costs down (Rahmandad and Ton 2020)? And what level of 'treating better' would cease to be profitable? Answering such questions requires solving the underlying dynamic task.

Dynamic tasks make accurate learning difficult because past as well as current actions may impact current performance. This is known as the credit assignment problem (Minsky 1961): how should we distribute the credit for the current outcome among recent and past choices? Consider a manager who observes that productivity is high. Is this the result of actions implemented recently or due to a sequence of actions taken over a long period? The credit-assignment problem is distinct from the well-known exploration-exploitation trade-off. By exploring a manager may find out more about an action, which can improve choices in the future. In dynamic tasks choosing an action does not only provide information about future payoffs but also *changes* the payoffs. The challenges of dynamic tasks are also conceptually distinct from issues related to uncertainty in outcomes and

dynamically changing environments that exogenously alter the task's payoffs. Even if outcomes from a given set of action sequences are deterministic and stable over time, the credit assignment problem remains extremely challenging. For example, consider a novice who loses a chess game. What move should the loss the attributed to? The last, the first, or the whole sequence? Uncertainty and environmental dynamism are of course important in strategic settings and further complicate learning in dynamic tasks, yet in this paper we focus on deterministic dynamic tasks in stable environments to tease out the sources of challenges specific to dynamic tasks.

Evidence on individual decision-making in dynamic tasks

To examine if people can successfully manage dynamic tasks, and to understand learning in these tasks, management and psychology scholars have relied on experimental methods. Initially, the tasks used were often designed to be realistic "microworlds"(Gonzalez et al. 2005) mimicking, at a small scale, real life decision-problems. The overall conclusion from this line of research is that people perform relatively poorly in dynamic tasks. In a variety of realistic tasks from supply chain management (Sterman 1989) to firefighting (Brehmer and Allard 1991), allocation (Gonzalez 2004), resource management (Moxnes 1998), and competitive strategy (Kunc and Morecroft 2010), among others, individual performance is rather poor, compared to the maximum possible (for reviews see (Gonzalez et al. 2017; Hotaling et al. 2015; Osman 2010)). Many of these tasks were designed to represent real-life business settings. For example, Rahmandad and Gary (Forthcoming) designed a simulation of a service firm based on data from multiple case studies (Rahmandad and Ton 2020). Participants made decisions about worker compensation and task design and observed data on compensation, sales, and profits. The resulting performance landscape (modelled after real cases) had two peaks. The results showed that even participants with managerial background converged mainly to the peak with high short-run profits but low long-term profits.

While such realistic designs have many advantages, they also make it difficult to understand *why* people tend to do poorly. First, these tasks tend to be complex, with many interacting variables, which makes it difficult to disentangle what causes poor performance. Second, using realistic tasks

has the disadvantage that people may bring inappropriate expectations to the task; expectations that may be accurate in reality but may not hold in the particular task. For example, Rahmandad and Gary (Forthcoming) find that experienced managers do worse than inexperienced subjects, because they focus on inefficient parts of strategy space in the simulation environment, presumably because of what they had learned in their own business environment. Finally, in a realistic microworld it is often very hard to establish fair performance benchmarks against which one can assess individual performance. And absent a solution to the optimal learning problem in a dynamic task, how could we assess if observed performance is inefficient? Because of these disadvantages, the literature has more recently started to focus on more simple and abstract tasks (Gonzalez et al. 2017).

One of the established tasks in this space is the so called 'Harvard game' (See Prelec 2014 for a review). This task provides a very simple set up to examine trade-offs between immediate and long-turn returns and as a result allows for a better examination of the underlying learning mechanisms. To build on this prior literature we use variants of this task in the current study. Here we first explain the task before discussing relevant prior findings.

The Harvard Game

In the Harvard game participants repeatedly choose between two alternatives (here we call them # and @) and receive a 'profit' (p) after each choice. They only observe the profit from the alternative they chose. Their goal is to maximize their cumulative profits over T periods of the game ($\pi = \sum_{t=1}^{T} p(t)$). The task is dynamic: the profit from each choice depends on the current choice as well as the N previous choices. Specifically, the profit from each choice depends on the fraction of @ choice in the N previous periods (we denote this fraction by $r_{@}$). In its most common incarnation profits depend on the past N=10 periods using the deterministic and linear profit functions in equations 1 and 2 (also shown in Figure 1)

$$f_{@}(r_{@}) = 6r_{@} \quad (1)$$

$$f_{\#}(r_{@}) = 3 + 6r_{@} \quad (2)$$

< Insert Figure 1 around here>

Here $f_{@}(r_{@})$ is the profit in the period t from choosing alternative @ when the proportion of choices of alternative @ in the most recent ten periods (t-10,t-9,...,t-1) was $r_{@}$. As Figure 1 shows, the profit from choosing both @ and # is higher when $r_{@}$ is higher. Importantly, for any given value of $r_{@,}$ the profit from choosing # is higher than that of choosing @. This makes the # choice attractive. Choosing #, however, reduces the past proportion choices of @, $r_{@}$, and thus reduces the profitability from both alternatives in the next ten periods. At the start of the experiment, there has not been any prior choices and the fraction of choices of @ is not defined. Following common practice, we initialize the system so that the previous 10 choices have alternated between @ and # (resulting in an initial fraction of @ choices equal to $r_{@}$=0.5).

To maximize profitability in the Harvard game participants must trade-off the immediate gain from choosing the 'tempting' alternative # (with higher immediate profits) against the long-term benefits from choosing alternative @ which increases $r_{@}$ and thus increases the profitability from both alternatives in the next ten periods. In the long-run, choosing only # pays off 3 points per period, while choosing only @ pays off 6. In fact, given the linear profit functions the optimal strategy is to always choose @. Mixing the two choices will result in expected profits ranging between 3 and 6 per period in proportion to the fraction of @ choice in the mix. The exception is if it is known that there are only a few periods left. If there are only six periods left, the optimal choice during those six periods is to choose # in the remaining periods.

The Harvard game was designed to make the contrast between immediate profits and long-term profits as stark as possible (Prelec et al. 1986): long-term profits is maximized by choosing the alternative, @, that minimizes immediate profits. While abstract, this task captures the essence of many managerially relevant problems such as deciding whether to cut corner to improve short-run productivity or to invest in capabilities that improve long-run capabilities (Rahmandad and Repenning 2016). If managers observe changes in productivity, they cannot be sure whether the changes are due to recent activities or investments some time ago. They have to learn the payoffs from alternative actions and based on this figure out the optimal policy.

A similar learning challenge exists in Harvard game. To illustrate this, suppose the initial $r_@$ is 0.5. A participant selects @ and gets a payoff equal to 3 (6*0.5 = 3). If the action 10 periods ago was #, the @ choice will raise $r_@$ to 0.6. Suppose the participant then tries action #. The observed payoff will then be 3+0.6*6 = 6.6. Now the participant has tried both actions and the payoff from # was higher. The participant may then continue to choose #. By doing so the payoff from # will decrease (as the fraction of @ choices in the last ten periods decreases). How will the participant react to this change? Will the change be attributed to stochastic variation or unexplained systematic changes in the underlying payoff-generating function? Or will the participant start suspecting that payoffs depend on actions chosen in the past? Even so, will the participant be able to figure out the rule and increase the proportion of @? This is challenging: for example, by selecting @ again, the observed immediate profit will yet again be lower for @ than the most recently observed payoff from #. Only by repeated choices of @ will a participant discover its long-term superiority.

Prior research has shown that even after several hundred periods of choice many subjects do not learn the optimal policy (i.e. only choosing @; see (Prelec 2014) for a review). Most participant continue to choose # about 40% of the time. Several moderators of performance are known. The value of N, the number of past periods that matter, is important. If N is larger the problem becomes more difficult (the impact of an action is spread over a larger number of future actions) and performance deteriorates (Warry et al. 1999). If the difference in the immediate profitability of the two actions (which was 3 in the above version, i.e., $3 + 6r_@$ versus $6r_@$) is larger, performance also deteriorates (Warry et al. 1999). Performance is also worse if payoffs are stochastic instead of deterministic (Gureckis and Love 2009a). Performance is significantly worse in the "delay" version of the game (Herrnstein et al. 1993). In this version, the two choices (@ and #) pay the same amount but they lead to different delays until another choice is available. That delay is longer for @, as well as for lower $r_@$, leading to similar overall payoff structure, but a more visceral perception of payoff differences. If the task is redesigned to make learning easier, performance also improves (Brown and Rachlin 1999; Gureckis and Love 2009b; Stillwell and Tunney 2009; Tunney

and Shanks 2002). Nevertheless, even after 1400 periods of learning, with feedback every 100 periods, and an early (cost-free) practice round of 100 choices, about half the participants fail to exclusively choose the optimum (@) button (Tunney and Shanks 2002).

EXPLAINING POOR PERFORMANCE IN DYNAMIC TASKS

Why do people perform poorly in dynamic decision-making tasks, including the Harvard Game, typically focusing on alternatives with high short-run payoffs? Several explanations may be relevant.

Misaligned Incentives: Incentives emphasizing short-term outcomes is an important contributor to myopia in many real-life settings. Incentives based on short-run profitability encourages managers to focus on short-term and skimp on investments even if they do understand the long-term negative consequences of doing so (Laverty 1996). However, experiments with dynamic tasks show that a bias towards actions with rewarding short-term outcomes persist even when incentives are based on long-term outcomes (Gureckis and Love 2009b; Rahmandad and Gary Forthcoming).

Poor impulse control: In personal decisions it is well-known that individuals are often tempted by alternatives (e.g. junk-food) that offer short-run benefits even if they have known costly long-term effects. Postrel and Rumelt (1996) argue that a similar problem of poor impulse control can explain short-termism in some strategic settings. Poor impulse control could possibly also explain the short-term bias in dynamic decision-making tasks: people may know that @ is better in the long-run but fail to choose it because the increase in payoff from # is tempting or salient. Poor impulse control cannot fully explain the short-term bias, however, partly because participants do improve over time and also improve their performance if the task is redesigned to make learning easier (Brown and Rachlin 1999; Gureckis and Love 2009b; Stillwell and Tunney 2009; Tunney and Shanks 2002).

Sub-optimal learning: The most common explanation of poor performance in dynamic decision-making tasks is sub-optimal learning (Gonzalez et al. 2017). The credit assignment

problem makes accurate learning difficult in dynamic tasks because it is not clear what action caused an observed change in performance. In the Harvard game, for example, participants should learn whether and how past actions impact future payoffs. Effective algorithms exist for such learning problems (Sutton and Barto 2018) and some experimental evidence suggest that people use heuristics that share some features of those algorithms (Fang 2012; Gureckis and Love 2009b; Simon and Daw 2011; Walsh and Anderson 2011). Implementing those algorithms in complex problems, however, is cognitively challenging because a lot of information needs to be accurately remembered and updated (Fu and Anderson 2008). As a result, it is likely that people rely on simple representations of dynamic problems. A poor match between these representations and the true dynamics can explain poor performance in dynamic tasks. For example, students in the beer-game may fail to understand that it is their own actions that cause fluctuations in demand, blame others, and order more as a precaution, which in turn amplify the fluctuations (Sterman 1989). Showing that performance is improved when the learning challenge is simplified, several studies support the interpretation that poor performance is due to challenges in learning. For example, providing a cue that is correlated with the state of the system ($r_{@}$) (Gureckis and Love 2009b), frequent benchmarking of subjects' performance against the best they could do (Stillwell and Tunney 2009), transforming the task to one of loss-minimization (Tunney and Shanks 2002), and letting subjects know that @ choice improves future payoffs (Herrnstein et al. 1993), enhance learning and improve performance.

The sub-optimal learning explanation faces two challenges. First, it encompasses a broad set of distinct mechanisms, spanning poor formulation of the problem, limits to memory, incorrect updating, and limited exploration to name a few. Even if we accept that sub-optimal learning is important, we need to establish which mechanisms are the source of the challenge. Second, recent research has argued that labeling performance as poor may be due to lack of appreciation for the complexity of the task. Instead, the observed human performance in these tasks may well be a product of sensible or even rational learning.

Unknown Dynamics and Rational Avoidance: What appears to be short-termism and poor performance can be the outcome of an optimal policy of learning and exploration when participants are uncertain about the dynamics of the system (Denrell 2007; Denrell and Le Mens 2020). For example, suppose a decision-maker is uncertain about whether the payoff from an alternative is always low or low initially but increasing over time. If the decision-maker comes to believe that the payoff does not increase over time, it makes sense to avoid this alternative. By avoiding it, however, the decision-maker will not be able to find out if its payoff is indeed increasing and if it would be the superior choice in the long-run. The decision-maker may thus end up persistently underestimating the long-run potential of an alternative with an initially low but increasing payoff. Indeed, Denrell and Le Mens (2020) prove that a similar bias occurs even if the decision-maker is rational, foresighted, and follows an optimal strategy for balancing exploration and exploitation.

Unknown dynamics can similarly explain what appears as short-termism in dynamic tasks. In typical experimental set-ups, including the common version of the Harvard Game, participants are not informed about the nature of the tasks or given any instructions beyond trying to maximize payoff from choosing between two buttons. Suppose, for example, that participants come to believe that each action, @ or #, only has an immediate payoff and does not impact the future payoff of any action. What would such participants learn? They would observe that # seemingly pays off more in the sense that whenever they switch to it then payoff increases. Choosing # repeatedly will decrease the payoffs to both options, but this trend may be interpreted as stochastic and exogenous variation. In brief, sensible learning within a mis-specified model (e.g. that actions do not impact future payoffs) can explain the tendency to choose the sub-optimal action #. In fact, such rational biases may persist even if a participant believes that past actions impact the payoff to future actions but does not know for how long this effect lasts, i.e., does not know the value of N (see Appendix A for a formal analysis). In short, even rational behavior, in the presence of task uncertainty, can be mistaken for biased and poor performance.

Computational complexity: A related possibility is that dynamic tasks are so complex that even our best algorithms would take a very long time to learn the optimal policy and would show poor performance initially. Without knowledge of the form of the payoff function, learning in a dynamic task is computationally challenging. There are many possible shapes the payoff function could take: it may depend on the past period action, on actions many periods before, or on the sequence of actions chosen the last N periods. Consider, for example, the Harvard Game. Even if participants knew that current payoffs depend on the previous N=10 actions, this knowledge does not tell them how payoffs depend on the sequence of the past 10 actions. If the order of actions matters, there would be 2^N (=1024) alternative action sequences. To estimate how the payoff from actions @ and # depend on past actions participants would need to try each of these 1024 sequences and then choose @ and # respectively. Thus, one needs 2048 experiments to fully map the state-action-profit combinations. In the actual task, because the payoffs only depend on the proportion of choices in the past ten periods, the relevant states are far fewer, only 11 ($0 \leq r_{@} \leq 1$, with increments of 0.1), but this is not known to participants. Absent this knowledge it would take a very long time to explore the full state space, so it is perhaps not surprising that people seem to learn slowly in the Harvard game; and given that most realistic tasks are far more complex than the Harvard game and often include shorter trial periods, observed performance shortfalls in past research may have been unavoidable outcomes of extremely complex learning tasks.

Indeed, Sims and colleagues (Sims et al. 2013) showed what appears to be sub-optimal behavior in the Harvard game could in fact be the result of a rational Bayesian learning algorithm, facing the problem of a large state-space. They showed that even the best algorithms require several thousands of periods to learn the optimal policy in the (stochastic version of) the Harvard game. In fact, some human subjects' performances were close to such "rational" benchmarks. This raises the possibility that computational complexity, instead of poor learning or decision-making, is the root cause of the seemingly low performance in the Harvard game. This type of complexity is not limited to the Harvard game; the majority of research on dynamic decision making uses tasks with

complex payoff generating functions that are not known to experimental subjects (Osman 2010). Perhaps participants in these experiments are doing well compared to what is computationally feasible? That possibility has different managerial implications compared to learning failures.

Unsystematic Problem solving: Finally, the reason for poor performance may have little to do with learning. Even if people knew the payoff function in a dynamic task, they might find it challenging to compute the optimal long-term solution. People have difficulty solving problems where reaching the goal requires steps that seemingly moves them further away from the goal (Newell and Simon 1972). Using backward induction to solve dynamic programming problems is also known to be challenging. Hey and Knoll (2011) show that only 12% of all participants behave rationally and use backward induction to find an optimal strategy in a sequential decision-making task, although another 20% use a simplified version of backward induction. There is also evidence that people have difficulty making effective use of relevant information about dynamics when there is no need to learn, e.g. failing to deduce the number of customers in a department stores based on inflow and outflow rates (Cronin et al. 2009).

In the Harvard game, finding the optimal policy is not difficult if the problem is represented as a comparison between the immediate benefit and the long-run gain. Consider a given period: is it better to choose @ or choose #? The choice in this period has an immediate impact (the profitability in that period) and a long-term effect. The immediate impact is that choosing # generates 3 units higher profits than choosing @. The long-term effect is that choosing @ increases $r_{@}$ by 0.1 for the next ten periods (compared to a choice of #). This boost of $r_{@}$ increases payoffs (from either choice) by 0.6 for each of the next 10 periods, leading to a total of 6 units of long-term benefit for choosing @. Since 6 > 3, it is profitable to always choose @ (if there are enough periods remaining to experience the long-term benefits of $r_{@}$). People may not intuitively understand how to represent

this problem in a transparent manner.[1] Instead they may rely on non-analytical strategies (Beach and Mitchell 1978), using a simple heuristic such as hill climbing in the action policy space (changing actions while performance increases in response).

Research gap

Poor performance in dynamic tasks have been blamed on sub-optimal learning (Gonzalez et al, 2017), but could also be due to unsystematic problem solving (with no learning involved), reflect limits that apply to all actors and algorithms in novel and unknown environments, or even be explained as rational learning. Previous research has not clearly distinguished between these very different explanations.

In fact, even the explanations that focus on sub-optimal learning, and tackle simpler tasks such as Harvard game, have not pinned down the specific mechanisms. Consider the problem of uncertainty about the dynamics of the system. Almost all studies tell participants nothing about the underlying system. One exception is experiment 5 in Herrnstein et al. (1993) where participants were told that taking the @ action repeatedly would increase payoffs from both actions, and taking the # action repeatedly would decrease payoff from both actions. Providing this information did improve performance substantially, even though the hint did not specify the number of periods in the future that the action would impact. Past experimental work has also only partly addressed the issue of computational complexity. Brown and Rachlin (1999) designed a different task with a much-reduced state-space (N=1). They also separately displayed the impact of each choice on immediate and future profitability. This design significantly simplified the task, and improved performance. However, the much smaller state-space and the display of both the short-term and the

[1] It may be argued that the corner solution to the Harvard game, where @ should always be chosen, is not intuitive. Still, poor performance has also been observed in versions of the Harvard game with an intermediary optimal solution (Hernstein et al, 1993, Experiment 2).

long-term contribution of each choice effectively removes the credit-assignment problem central to dynamic tasks. Similarly, Stillwell and Tunney (2009) reduced the state space and also displayed the impact of each choice on immediate and future profitability, again removing the credit-assignment problem. Their design also involved feedback (comparing actual vs. maximum feasible performance) every 100 trials. Performance improved, but it is not clear why. Finally, studies have provided hints about the relevant state variable in the Harvard game (i.e., the proportion, $r_{@}$). Herrnstein et al. (1993) (Experiment 1) showed an arrow pointing to this fraction and Gureckis and Love (2009b) had 11 lights corresponding to the 11 possible fractions. These manipulations modestly improved performance but it is not clear why: is it the reduced combinatorial complexity (from knowing that the proportion matters), signaling the long-term value of actions (inferable from changes in the $\underline{r_{@}}$) or some other reason? Finally, past research has not clearly teased out explanations based on inability to solve the problem from the learning challenges in the Harvard game.

The contribution of this paper is to design a series of studies that allows us to infer when and why people perform poorly in the Harvard Game. We carefully manipulate the computational complexity of the tasks, the information provided about the dynamics of the task, and whether participants know the payoff function or need to learn it. Different explanations make different predictions about what manipulations should increase performance. If computational complexity is important, performance should improve when we reduce the computational complexity by reducing the state-space. If uncertainty about the dynamics of the task matters, performance should improve when we inform participants about how long into the future an action impacts payoffs. If sub-optimal learning is the problem, performance should improve when no learning is involved. Finally, if unsystematic problem solving is what matters, then performance should remain poor even when we inform people about the payoff function and thus remove the learning component.

METHODS

We use the Harvard Game as our experimental paradigm and conduct three experiments. The first experiment is the primary one where we manipulate the computational complexity of the task and the information provided to participants. The second experiment introduces two interventions to enhance learning and better understand the heuristics people use in dynamic tasks. The third experiment assesses if performance may improve in the one-shot version of the task, where subjects are given full information on task payoffs and structures but are not exposed to experiential learning. Below we elaborate on common design features of the experiments before explaining the three experiments and their results in detail.

Experimental task

The task structure and payoff function we used is explained above. In the first two experiments, participants make 500 choices. In all experimental conditions we initialize the system so that the previous 10 choices have alternated between @ and # (resulting in $r_{@}$=0.5). Actual experiments are counter-balanced, so @ and # are swapped for half the participants. However, for clarity throughout the text, including above, we use @ to designate the long-term choice and # for the one paying better in the short-term

Participants and procedures

Participants- For the experiments we recruited participants from Amazon Mechanical Turk (AMT). AMT offers a diverse participant pool that represents the general population better than typical participants on university campuses. On the other hand, AMT participants may have more limited attention spans and put more emphasis on the financial rewards of participating in an experiment compared to in-person participants. Overall, prior comparative research has shown that when properly recruited and incentivized AMT participants offer reliable, even superior, sources of data on a variety of tasks (Mason and Suri 2012; Paolacci et al. 2010). We therefore combined financial incentives for performance and a rigorous recruitment and training procedure (below) to leverage the benefits of AMT participant pool while minimizing the risks.

Procedures and incentives- Participants were recruited from AMT pool and restricted to those in the U.S. (to ensure English proficiency) and with at least 100 previous tasks on AMT (to ensure quality). They joined to participate in "A Repeated Choice Experiment" that consisted of two stages. First, they had to read the instructions for the task (which varied based on the experimental condition into which they had been randomized; see appendix B for details) and take part in a qualification test. They were paid $1 for this stage as long as they were attentive to the instructions, but had to correctly answer 4 or more questions in a stringent six-question test to qualify for the second stage, the actual experiment. In all but the one-shot condition experiment three they were told that their payments for part two consists of a minimum of $2 for completing the second part (500 periods of choice) and between 0 and $3 in performance bonuses. In the one-shot condition they were paid between 0 and $2 in performance bonus based on the payoffs expected from their proposed strategy (details under experiment three). They learned that this bonus depended on their cumulative profits compared to best/worst performances (and the quality of their answer to mental model elicitation question (details below) where it applied). Use of cumulative profits for performance bonus excludes rational discounting of later performance and simplifies the interpretation of results. In practice the paid performance bonus was proportional to a participant's cumulative profit scaled between the minimum and maximum feasible profit scores. In cases a participant spent significantly more time on the task than typical (28% of participants) we paid them an extra (and un-advertised) effort bonus to keep participants' hourly wages at decent levels.

Following this procedure 1173 participants took part in the qualification survey and 547 qualified to participate in the experiments. 526 started the experiments and 515 completed the experiments and constitute the union of 3 samples (218 in experiment 1, 187 in experiment 2, and 110 in experiment 3). Demographic variables (age, gender, education, employment, income; all self-reports) had limited impact on qualification while time spent on instruction was a strong predictor. The qualification test for one-shot condition also proved more stringent. In the remainder of the paper we focus on the 515 participants who completed the experiment. The average (standard

deviation of) age for this group is 36.3 (10.5). This was a heterogeneous group in terms of education (40% with 4-year degrees; 36% with some college; 13% high-school graduates; 9% with Professional degrees; 1% with lower than high-school education). Gender mix was slightly towards male participants (44% female). The majority were working full time (67%), followed by 16% part-time workers, 8% unemployed, and 5% home-makers. Income also varied but most participants were distributed uniformly between $10,000 and $80,000 in annual income. Participants (excluding experiment 3, i.e. the one-shot condition) spent an average (standard deviation) of 6.8 (7.3) minutes on the instruction stage and 14.5 (22.6) minutes on the experiment. One-shot participants spent 9.8 (6.3) minutes on instructions and the single choice task. On average participants (excluding one-shot condition) earned $4.57 ($0.96) in the experiment which resulted in the average hourly rate of $16.4 ($6.1). One-shot participants earned $16.6 ($9.0) per hour. These rates are significantly higher than typical AMT compensation rate (Paolacci et al. 2010). Overall, stringent recruitment, low drop-out rate, a diverse sample, attractive financial incentives, and a significant level of engagement with the task suggests that the recruitment procedure has yielded an informative sample leveraging the benefits of AMT and minimizing its downsides.

EXPERIMENT 1: DISENTANGLING CHALLENGES TO LEARNING

In the first experiment we manipulate the size of the state space and the information provided to participants, to examine whether poor performance is due to i) uncertainty about dynamics, ii) computational complexity iii) sub-optimal learning or iv) unsystematic problem solving. The online appendix B provides detailed instructions for participants and the user-interface of the experimental conditions. Figure 2 shows the interface for the condition with the most comprehensive information (Full Knowledge) which is helpful to explain the other conditions as well. Table 1 summarizes the conditions in experiment 1, the learning challenges relevant in each, and how the impact of each challenge can be assessed by comparing results across different conditions.

<Insert Table 1 Around Here>

<Insert Figure 2 Around Here>

Classical- This first condition is similar to the one used in most prior studies. Participants are only told that they will be choosing repeatedly between two buttons with the goal of maximizing their cumulative profits over T = 500 periods. Besides these rather uninformative instructions the user interface excludes items 1-3 from those in Figure 2.

History- The History condition removes the uncertainty about the dynamics by informing participants that their profit function is deterministic, stable, and depends on the combination of their choices in the past N=10 periods. It also ameliorates memory challenges by showing the history of previous 10 choices (item 1 in Figure 2) but leaves out other information. This condition is significantly less ambiguous than classical and thus less susceptible to misspecification: participants now know that payoffs depend on past actions and know the relevant window of past actions (the last ten) that matter. They also know payoffs are deterministic and stable, so variation in payoffs cannot be explained as random variation or environmental shifts. This condition remains complex computationally, however, requiring 2048 actions to fully map out all possible state-action payoffs. Of course, the learner may suspect that some payoff functions are more likely than others. Yet, depending on ones' priors even a Bayesian learner would need a few hundred data-points to move close to a good solution (Sims et al. 2013). Comparing the performance in the Classical and History conditions allows us to find out whether the poor performance observed is due to uncertainty about the dynamics (Compare Columns 1 and 2 in Table 1).

Informed History- The Informed History condition informs participants that profits depend only on the fraction of choices of @ in the previous ten periods ($r_{@}$) and is stable and deterministic. Moreover, the current $r_{@}$ is shown on the interface along with previous ten choices (items 1 and 2 in Figure 2 are included). Participants are not told, however, how $r_{@}$ impacts the payoffs. Still, telling participants that the payoff from an action depends on $r_{@}$ simplifies the learning problem: participants need to discover 22 payoff values, one pair for each of the 11 possible states (i.e., a $r_{@}$ of 0, 0.1,…, 1). This can most efficiently be done by taking 11 @ decisions, followed by 11 # (or

vice versa). After that, a sophisticated participant has all the information needed to calculate the optimal choice. Comparing the performance of participants between the History and the Informed History conditions can thus tell us about the relevance of computational complexity (See columns 2 and 3 in Table 1). While prior work has provided hints about the relevance of showing the state variable ($r_@$), our design is the first that explicitly informs participants that payoffs only depends on this variable.

Full Knowledge- In this condition participants are fully informed about the structure and dynamics of the task and the immediate and long-term profit resulting from each choice. To achieve this, we show participants, at each choice opportunity, a table of how profitability for the two actions depends on $r_@$. That is, we show them Figure 1 in a table format. As in the Informed History condition participants are informed that the payoffs only depend on $r_@$. We also show them the next-period profits from each action. Providing participants with information about the payoff function eliminates the learning challenge but the problem of deriving the optimal solution remains. Comparing this condition with Informed History clarifies the role of unsystematic problem solving in accounting for poor performance in the Harvard Game.

Outcome and process measures

We measure the proportion of periods in which participants choose the optimal long-term action and the cumulative profit (π). To these commonly used outcome measures we added three additional measures:

Learned strategies (s_L and s_{SH})- The choices of a participant over the later periods reflect the strategy she has learned from experience. We use two summary measures of learned strategies to reflect the short and long-term learning. We use the average action over periods 50-90 ($s_{SH}^i = \sum_{t=51}^{t=90} a_i(t)/40$), where $a_i = 1$ if the optimal long-term action is chosen, as the short-term measure of learning. The long-term counter-part focuses on the last 90 periods ($s_L^i = \sum_{t=401}^{t=490} a_i(t)/90$). We do

not include the last 10 periods to avoid confounding steady-state optimal strategy (@ actions) with end-of-horizon choices (which could include # on the optimal path).

Exploration level (*xl and xs*)- The task includes 11 distinct values of $r_@$ combined with the two choices which participants (unless informed through instructions) need to experience before they have fully explored the profit function. Ideal exploration would visit all those combinations in 22 periods of choices (e.g. 11 @ choices followed by 11 # choices) and thus learn the entire payoff function. The *exploration level* can be measured as the fraction of the 22 unique state-action combinations observed up to some point in the experiment. This measure is normalized relative to two benchmarks. One is the ideal exploration level $xl^*(t)$ obtained through the policy outlined above, which reaches an exploration level of 1 (explored all alternatives) after 22 choices. The other is a policy of random choice, $xl^r(t)$, which can be estimated with Monte-Carlo simulations. The random choice policy does not lead to high exploration levels as random choices rarely visit extreme values of $r_@$ (0 or 1). Exploration level ($xl_i(t)$) for individual i changes over time, so we also create a single exploration score, xs_i, which measures where between the two benchmark values the exploration level stands (averaged over the total number of periods, T). Specifically:

$$xs_i = (\sum_{t=1}^{500} \frac{xl_i(t) - xl^r(t)}{xl^*(t) - xl^r(t)})/500$$

This measure equals 1 if a participant explores optimally, 0 if they follow the random exploration policy, and could become negative if very few $r_@$ values are experienced.

Mental Model errors (e_{MM})- Besides participants' choices we elicited data on emerging mental models of participants in conditions where they were not informed about the exact payoffs in the task. Specifically, after 500 choices participants were informed (reminded, if they were already informed) about the dependence of their profit on $r_@$, and were asked to estimate that relationship. They could either complete two graphs with $r_@$ as X-axis and $f_@$ and $f_\#$ as Y-axis, or enter their perceptions in a table. These mental models are compared against the correct profit functions (eqs. 1 and 2) and the mean absolute error between true *f* and perceived one (over full range of $r_@$) is

calculated as a measure of mental model quality for each choice, i.e. $e_{MM}^{@}$ and $e_{MM}^{\#}$. This measure allows us to assess if people who found the optimal strategy really understood how the task works.

Replicability

To provide maximum transparency and facilitate replication and enhancements in the online appendices B and C we provide detailed instructions, qualification survey, computer code for building and hosting the experimental conditions, full (anonymous) data from this experiment, and the statistical analysis codes.

Results-Table 2 summarizes the main outcomes across different experimental conditions. Simple t-test comparisons of Classical condition show little difference from History, but also a few differences with the Informed History and Full Knowledge conditions. Results also show notable within-condition heterogeneity in the learned strategies. Mental model errors across the first three conditions show limited difference ($p>0.1$ for t-test comparisons). Next, we compare learning across the four conditions to inform the mechanisms regulating the effectiveness of individual learning in dynamic tasks.

<Insert Table 2 Around Here>

What explains learning trajectories in Harvard game? Individuals in all conditions learned to improve their performance by increasingly favoring the long-term option, @. Figure 3-A reports these learning trajectories over time (averaged within each condition) as well as the optimal trajectories for Full Knowledge (taking @ from beginning till period 495, then switching to #) and Informed History (initially taking 11 periods on each of two actions, then switching to @ until 495, ending with #). The optimality of switching to # in the last 6 periods is due to end-of-horizon response. Therefore, s_L excludes the last 10 periods, averaging choice in periods 400-490. In this interval, across the four conditions, participants chose the long-term (@) choice 68% of the time, higher than chance, and yet below optimal (i.e. 100%). There was also significant variation across

participants: whereas the 10th percentile of the participants had s_L=0.39, the 90th percentile participant had completely converged to choosing @ all the time (s_L=1). In fact, 20% of participants had s_L=1, with largest fractions in the Classical and Informed History (Figure 3-B, columns on the right). While those with s_L=1 are a minority across all conditions, the results offer evidence for learning. However, it is not clear if these participants had fully realized the optimal policy. Specifically, none of those participants with s_L=1 switched to # actions in periods 495-500, which is the optimal policy. Nevertheless, some participants intuitively incorporated some end-of-horizon optimizing in their actions (thus the drop in actions at the end).

We next use regressions to further explore the differences in learned strategies across various conditions. These differences are especially important because we introduced two conditions (Informed History and Full Knowledge) that, analytically, are much simpler than conditions studied in prior research. By removing the ambiguity and large state-space problems, these conditions make the task tractable and optimal strategy can be pursued either from the beginning (under Full Knowledge) or after just 22 periods of exploration (in Informed History). Therefore, a large difference in learning between the first two conditions and the last two shows that the ambiguity and large state-space problems are critical learning bottlenecks. Small gaps, on the other hand, point to the key role of unsystematic problem solving.

<Insert Figure 3 Around Here>

Model 1 in Table 3 reports how different conditions predict s_L. Putting the Informed History, which has the highest long-term learned strategy, as the reference condition, we note that only the Classical condition is different from it with a 0.1 reduction in s_L (p=0.02). The modest difference is driven by a notable minority of participants who have converged to low values of s_L in the Classical condition (See Figure 3-B). No other differences in long-term learned strategies are notable. Variations in cumulative performance (over 500 periods) are similarly small and only marginally different between Classical vs. Full Knowledge (p=0.07 in a T-test). In models 3 and 4 (Table 3) we

include process measures that help tease out some of the intermediate drivers of learning across different conditions. Full knowledge condition is not included in these two models because, knowing the actual payoff function, mental models are not elicited for subjects in this condition. Qualification scores are not a predictor of learning or performance, suggesting that the learning challenges are not due to variations in understanding task instructions. Exploration score, x_s, emerges as the only important predictor of finding the best strategy ($p<0.001$), and in fact it explains away the difference in learned strategy between classical and Informed History conditions. This may suggest that complexity of exploring different state-action-profit combinations is indeed a notable challenge to learning. However, that conclusion is tempered by two related observations. First, the quality of participants' mental models at the end of experiment has no bearing on learning and performance. If exploration helped learning through building a better understanding of state-action-profit mapping, then we would have expected $e_{MM}^{@}$ and $e_{MM}^{\#}$ to strongly predict performance. The fact that none of those coefficients are significant ($p>0.4$) is surprising and points to a disconnect between how well people understand the task structure and how successful they are in managing it. Second, exploration is not a concern in the Full Knowledge condition (participants fully know the state-action-profit map), and still participants in that condition do not perform any better than the Informed History participants ($p=0.53$). It seems exploration is important, but it is not operating through teaching participants about the state-action-profit mapping.

<Insert Table 3 Around Here>

These results provide a more nuanced view about several mechanisms that prior research had presumed central in explaining the complexity of learning in this dynamic task (Sims et al. 2013). Specifically: A) There is no major difference between Classical and History settings. Therefore, uncertainty about the dynamics (i.e., not knowing that the task is deterministic and stable, not knowing the number of periods an action impacts) does not explain the variations in performance. B) The difference between History and Informed History is very modest, and limited to lower exploration scores in the History case. Therefore, differences in the size of the state space (1024 in

History vs. 11 relevant states in Informed History) cannot explain much. Thus, computational complexity (which is much larger in History with 1024 states) does not really explain the results. C) Most participants remain far from optimal even when the states and the payoffs are known. This suggests that learning is not really the underlying problem. Poor performance remains even when we removed the learning component: people have a hard time figuring out the optimal policy even if they know the payoff function and have 500 periods to figure out the solution. Finally, even those participants who performed well do not have an accurate understanding: the quality of mental models is not correlated with the learning and performance outcomes. This suggests that the primary source of the underlying difficulty is unsystematic problem solving.

Experiment 2- Alternative learning heuristics

In experiment 1 we established that poor performance in Harvard task cannot be the result of rational (optimal) learning under uncertainty about the payoff structure. In the second experiment we focus on delineating between the alternative problem-solving heuristics people may be using in this task. We focus on two types of heuristics potentially applicable in dynamic tasks. First, people may attempt heuristics that resemble dynamic programming precepts: tracking the value of states and identifying valuable actions as those that make valuable states more likely. Prior experiments show that a subset of subjects may follow such heuristics (Hey and Knoll 2011).

A second possibility is that people use learning heuristics that do not track the value of different states. Instead of trying to identify the value of states and evaluate actions based on whether an action makes a valuable state more or less likely, people may directly explore the policy space. They may simply test different *policies,* observe the payoffs from those policies, and settle on a more promising policy through some search process in the policy space. Here policies could be defined as fixed sequences of choices, e.g. 'repeat @#@@#' would be a policy, and so would be 'repeat #'. This heuristic does not try to identify the value of each state but simply notes the payoffs associated with particular sequences of actions. Such methods are called "policy optimization methods" in machine learning and computer science, because they try to directly optimize the

policy instead of estimating the value of different states. Policy optimization is easy to implement but hard to generalize to new situations and often inefficient computationally because of the exceedingly large size of potential policy space. Suppose you tried to use this method in chess. You note that a given sequence of moves was associated with a loss or a win. Because the number of possible sequences of moves is enormous, and it is unlikely you observe the same sequence of moves again, information about the payoff associated with past sequences will not help you much in the future. Still, if the number of possible sequences is manageable, adopting such a heuristic may prove effective.

In the second experiment we design two interventions that help tease out these alternative interpretations of how people learn in dynamic tasks and inform ways to enhance that learning.

Experimental design and conditions- We compare learning in two new conditions against the classical condition. 187 Subjects who had successfully completed the training (see online appendix B for detailed instructions and test) and continued to simulation were randomized into one of the three conditions. Classical (66 subjects) was explained before, so we introduce the two new conditions, *Value Tip* and *Direction Tip* here.

Value Tip- This condition is designed to assist learners who may be attempting to follow a forward-looking strategy similar to dynamic programming by computing the immediate profitability of an action and comparing this to its impact on future profits. Specifically, at every period we provide learners with a tip about the increase in the future payoffs expected from choosing @ (compared to # choice). Learners are instructed that:

"*Besides the immediate profit they generate, the two choices also impact later payoffs. One choice enhances future profits more than the other. On the interface the choice that enhances future profits more than the other is identified. You can also see how much choosing that button will improve future profits beyond the alternative."*

Thus, participants will see how much choosing @ will enhance future profits. This amount is 6 in all periods except for the last 11 periods when this number goes down in increments of 0.6. If

participants use a forward-looking heuristic, and try to compute whether choosing @ is worth it, they only need to compare the increase in future profits (i.e., 6) with the decrease in immediate profits due to choosing @, which is 3 at all times and thus easy to learn. The resulting optimal choice is always @, except for the last 6 periods when the value tip falls below 3. To stay close to the comparison (Classical) condition we only alter the interface to include this value tip but do not offer any additional information.

Direction Tip- This condition is designed to assist learners who use heuristics closer to policy optimization (i.e., trying out different sequences of actions and observing their payoffs) by nudging them to search for policies that emphasis the @ choice more. Specifically, they are informed that one of the choices also enhances future payoffs, and that choice is specified on the screen:

"*Besides the immediate profit the two choices also impact later payoffs with one enhancing future profits more than the other; that future-enhancing choice is identified in the interface.*"

Note that this condition provides *less* information than the value tip. It does not provide information about how much choosing @ will enhance future profits. So, the challenges of identifying and comparing the impact of @ on immediate and future profits, central to applying dynamic programing heuristics, remain. However, the direction tip may encourage people who follow a policy optimization heuristic to explore policies that include @ more frequently and discover the value of those heuristics.

Results- Table 4 and Figure 4 report the results. Both interventions show some promise, especially early on. They improve the strategies learned in the short-run (s_{SH}) ($p<0.001$; from 0.38 in classical to 0.56 for both value and direction tip) and enhance cumulative profit ($p=0.02$). The long-term impact (s_L) is modest (going from 0.63 (Classical) to 0.65 (Value Tip; $p=0.34$) and 0.74 (Direction Tip; $p=0.02$)). The mediocre long-term performance of Value Tip is mainly due to a larger fraction of subjects continuing with exploring strategies with medium range s_L values. Closer inspection suggests many of those subjects have adopted strategies that consist of cycling between a

few successive # choices followed by a few @ ones. It may be that by informing participants about how much choosing @ will enhance future profits the Value Tip condition primed learners to realize those additional profits by periodically choosing #.

<Insert Table 4 and Figure 4 around here>

Overall, the results show that the Direction tip is superior: its short-term performance is similar to the Value Tip and it has better long-term performance. The moderate effectiveness of Direction Tip in enhancing learning came at limited costs: this type of tip is usually easy to provide (as it only highlights the temporal tradeoffs) and requires little training or calculation. We also experimented with a more subtle intervention, suggesting the short vs. long-term tradeoff by labelling the buttons as "Cake" and "Exercise", but found limited impact on learning outcomes (See appendix D for details).

These results provide further evidence that many participants seem to adopt a heuristic close to a policy optimization where they explore and choose between sequences of actions rather than trying to understand the value of different states. We found that the simple Direction tip, which encourages exploration of @, improved performance more than the Value Tip, which was most helpful for dynamic programming heuristics trading off long-term and short-term impact of different actions explicitly. Moreover, to the extent that Value tip improved performance, the impact could be attributed to encouraging the exploration of policies that emphasize @ instead of helping people more effectively apply dynamic programming heuristics.

Experiment 3- Would subjects fair better in solving the task analytically?

The fact that people seem to experiment with policies, rather than trying to identify the value of states, suggests that people may be under-performing their potential in the more analytically tractable cases. Rather than thinking through the problem and trying to find the optimal solution, people start experimenting with the task, trying out some sequences, and making small changes to sequences that have been associated with high performance in the past. The temptation to adopt such heuristics may limit how well they do in tasks which are in fact analytically tractable. Thus, it

is possible that people would do better by focusing more on thinking through the task analytically and less on experimenting with alternative policies. In the third experiment we examine if performance would improve if we remove the repeated choice aspect of the problem, provide them with full information about the payoff function, and encouraged participants to think thorough the problem and come up with a solution. This experiment also provides a test of whether impulsivity explains the poor performance. Impulsivity could play a role if people plan to select the long-run maximizing alternative but are tempted by the recent gains from trying the short-run alternative. To reduce such temptation, we instruct participants to think through the problem and plan a sequence of future choices without going through experimental trial and errors.

Design and procedures- Participants are randomly assigned to one of the two conditions: Classical (as control condition to ensure comparability with prior experiments) or One-Shot. The one-shot condition provides participants with the same information as in the Full-knowledge case, and asks them to think through the problem and offer their recommended strategy for playing the Harvard game without actually playing it. Detailed instructions for participants are reproduced in appendix B. Their strategy can recommend any sequence of # and @ choices, to be repeated indefinitely. Absent any experimentation this is a significantly shorter task. They are paid $1 for completing the instruction phase, and a performance bonus of $0-$2 based on the expected payoff from their recommended strategy. The large (in hourly compensation rate) performance bonus is designed to elicit cognitive effort from participants to solve the problem as well as they can.

Results-110 participants completed this experiment (53 in One-shot condition, and 57 in Classical). Those in classical condition performed similar to the previous two experiments, with slightly lower overall payoff (cumulative payoff of 2222 (448) compared to 2318 and 2271 in the first two experiments). Those completing the One-shot condition offered strategies that, when normalized to 0-1 range (so that it is directly comparable to S_{SH} and S_L), averaged at 0.52 with standard deviation of 0.33. In fact, those strategies on average cannot be distinguished from a random strategy yielding a performance of 0.5 (p=0.5). 15% of subjects in the one-shot condition identified the profit-

maximizing strategy (only @) and 13% proposed the worst possible strategy (only #). Comparing with other conditions, including those in experiments 1 and 2, subjects in the one-shot condition outperformed those in Classical condition in the short-run (i.e. compared with S_{SH}; values of 0.38 and 0.4; p=0.001), but performed worse than all conditions, including Classical, in the long-run (p-values <0.05).

These results provide evidence that 1) The trial and error experience may bias individuals in the short-term beyond what they would deduce analytically (or achieve by chance); yet 2) even the simplest version of the task is too hard for typical subjects to think through and solve analytically. This finding further limits what we can expect people to achieve in dynamic tasks: even when there are no uncertain dynamics, the payoff structure is fully known, and when subjects may not be misled by trial and error learning, they cannot perform better than chance in this task.

DISCUSSION

How do these results inform when and why managers might fail to learn in tasks where going for immediate payoffs impacts future opportunities? Prior experiments have showed that people do poorly in such dynamic tasks. In the Harvard task they choose the optimal action only about 60 percent of the time after 500 periods of experience. What is less clear from past work is why. In particular, is the poor performance observed due to ambiguity and computational complexity implying that even the best algorithms would not do any better (Sims et al. 2013)? Or is the poor performance due to sub-optimal behavior? And if so, what mechanisms explain such sub-optimal behavior?

To distinguish these possibilities, and to learn more about how people approach this task, we designed a series of experimental conditions that gradually removed the problems of uncertainty about the dynamics and computational complexity. In the Informed History condition, participants learned that payoffs depend on the fraction of choices in the previous ten periods; this reduced the state and search space, and allowed in principle for a complete solution of the problem in 22 periods. Moreover, in the Full Knowledge condition, participants were told about the payoff functions, reducing the task to a simple math problem. These changes, however, have very limited

impact on the performance of human participants compared to the classical task. Large differences in how well people perform in the Harvard Game remain even if we remove the learning, computational, and informational challenges that prior work has argued are important. Thus, the computational challenges to which some prior work has attributed the poor performance (e.g. Sims et al, 2013) do not explain much of the observed poor performance. Indeed, even when the states, actions, and the payoff functions are fully known most participants fail to identify the optimal solution.

We find that behavioral factors --from learning heuristics to poor analytical solutions-- explain much of the performance variations in rather simple and well-defined dynamic tasks. This matters because if the main reason for poor performance was computational complexity, there would be fewer opportunities for improvement (even the best algorithm could not do very well) and variations in performance would mainly be due to luck. We show that there is substantial variation in performance even in simpler set-ups. Thus learning difficulties due to causal ambiguity or novelty are not necessary for explaining variations in performance. Much of the variation in performance remains even when those barriers to learning are removed and even if people have substantial experience with a simple dynamic task.

Implications for understanding the nature of managerial learning in dynamic tasks

How can these results shed light on findings such as managers skimping on documentation to get more work done in the short-run even though it leads to more errors in the long-run (Rahmandad and Repenning 2016)? First, our results as well as prior work indicate that learning is not entirely "myopic" in the sense that it completely ignores the long-term and only focuses on the short-run. A learning algorithm which believed that each action only had a short-run impact would also converge to choosing # in all periods (see appendix A for a proof). More psychologically realistic learning models make a similar prediction. For example, the instance-based learning algorithm (Gonzalez et al. 2003), among the best accounts of human learning in static repeated choices (Erev et al. 2010),

predicts learners will quickly converge to the # choice in this task[2], a result we only observed among a small minority of participants.

While learning is not fully myopic, poor performance persisted even with the full knowledge of task structure and payoffs. Participants face a significant challenge deriving the optimal solution from knowledge of the payoff functions. Instead, most people seem to approach the task in a way quite different from the heuristics built on the dynamic programming principles. Dynamic programming derives the optimal solution by evaluating both the immediate reward generated by a choice and also the value of the "state" it leads to (Bertsekas 2007). In the Harvard game this would involve estimating the immediate payoff consequence of an action as well as its impact on the relevant state ($r_{@}$). Learning algorithms taking this approach have to accomplish two things: first, they should identify the relevant states and actions and the state-action-payoff relationships, and then trade-off between the immediate reward from each choice, and the changes in the value of states that result from the choice.

Our results suggest that people do not intuitively try to identify the relevant states or consider the impact of choices on future states. First, among those in the simpler conditions (Informed History and Full Knowledge), only a minority found that sticking to the @ choice is the right strategy. Even those who did, seem to have done so through a channel different from separating the value of states from actions. For example, they did not adjust their actions to incorporate end-of-horizon considerations, which dictated a switching to # for the last 6 periods of the experiment. Moreover, the quality of participants' mental models of state-action-payoff relationships had no bearing on their performance in the task, again suggesting a different mechanism to be at play. Finally, note that people seem to do well (relative to algorithms) in the Harvard game when dealing with the more complex version (the classical condition) with unknown states and under-specified determinants of outcomes. Yet, people do not perform well in the much simpler setting when the

[2] We are thankful to Coty Gonzales and Don Morrison for running the instance-based learning on this task.

task can be easily solved using a dynamic programming approach. What learning process may explain these contrasting learning outcomes?

While we have no direct evidence on the cognitive mechanisms people rely on, we propose that most subject use a simple process of trial and error. People try different policies, observe their performances, and tend to select policies that performed well in the past. For example, in the Harvard game continuously pushing #, or @, would each be a policy that our participants could try and evaluate by observing its past (average) payoff; so would be alternating between # and @ choices, and so on. This learning heuristic differs from dynamic programming because it does not attempt to understand the dynamics of the task, it simply tries sequences of actions and repeats those with favorable outcomes. This heuristic can be relatively effective when learners know very little about the task and thus the task is computationally very complex (Sims et al. 2013). However, when more is known about the task, following this heuristic ignores additional available information, leading to no improvement as we find.

A few observations support the hypothesis that people follow this heuristic. First, the state-action-payoff data offered to participants in the Full Knowledge condition is very useful for solving the dynamic program, but has limited value if one is using a policy optimization method. This may explain why Full Knowledge does not enhance performance beyond Informed History. Second, we noted that exploration was indeed valuable for learning, yet its effect was not moderated through improving the mental models of state-action-payoffs. A policy optimization heuristic depends on exploration, but does not utilize it to update the state-action-payoff mappings, instead, the relevant updates are in policy-payoff space. Third, as we observed among our participants, a policy trial and error heuristic would not recognize end-of-horizon adjustments, unless it has observed multiple instances of such end-of-horizons and has incorporated the end states in a modified policy function.

Policy trial and error may be a useful heuristic offering good payoffs when there is little known about task structure (e.g. Classical condition), but it cannot solve the task optimally. Subjects cannot systematically explore all sets of policies, including choosing @ eleven periods in a

sequence. Rather, they may try a few possible, often short, sequences. Depending on the order in which these sequences are tried, a learner may end up believing that a suboptimal sequence is superior to a sequence with only choices of @. Consider, for example, a learner who examines sequences of four choices. Remember that the Harvard game is initialized with choices alternating between @ and #. If the learner initially tries the sequence ####, the observed average per period payoff will be 5.4. If the learner then immediately tries the sequence @@@@ the observed average per period payoff will only be 2.1 (because the fraction of choices of @ is now low). Because 2.1 < 5.4, the minimum average payoff from choosing #### repeatedly, the learner may never revert back to @@@@ even if they continue with #-only choices and discover its long-term value to be only 3 points per period. Of course, if the learner explores, and tries repeated @ choices again, she may discover its superiority, which is consistent with our results that variation in exploration explains part of the variation in performance. This mechanism is also consistent with evidence that knowing about the maximum long-run payoff possible (i.e., 6 in our set-up) improves performance (Tunney and Shanks 2002). A learner who knows that the maximum payoff is 6 would be less likely to settle for a payoff of 3.

If people do follow a policy trial and error heuristic this can also explain why they do reasonably well in the classical condition, even if this condition provides little information about the relevant state space and thus is exceedingly complex analytically. By trying out a few sequences, people may realize that including a few choices of @ increase the payoff even from the choice of #. As a result, most people tend to alternate or include some choices of @, although very few understand the dynamics of the task and find the optimal strategy. In fact, the optimal policy in this task has an extremely simple structure (just push one button repeatedly), and thus likely to be explored (albeit not for long enough by many subjects). Using a policy trial and error heuristic would be less likely to find a good policy if the optimal strategy required a complex arrangement of actions (e.g. cycling through the sequence "##@##@@@@").

These results suggest that managers may not be myopic in the sense of ignoring the long-term. Rather, what seems like myopic behavior may occur because managers use policy optimization heuristics that "muddle through" (Lindblom 1959) complex dynamic problems. Unless managers could afford substantial exploration, those heuristics are unlikely to discover the superiority of policies that rely on alternatives with poor immediate payoffs. Moreover, even if managers are informed about relevant characteristics of the task (e.g. $r_@$ in our task), such information will not impact behavior much because they do not directly inform the typical learning heuristic. But that does not mean they could do better when restricted not to use their baseline learning heuristics. Experiment 3 shows that when forced to find the solution by thinking through the problem, subjects perform even worse, and no better than chance.

To improve learning managers may take a few distinct paths. First, they may engage in substantial exploration which would allow them to observe the payoffs from alternative policies including policies that persist with the long-term option. Case study evidence from Rahmandad and Repenning (2016) provides one such example: the software group that sustained quality processes had managers who had experienced the negative consequences of skimping on process, and had internalized the lesson of not doing so. This path is costly, and many would-be managers in such failed projects will not be promoted to utilize their hard-won lessons. Alternatively, managers may switch to a more analytical mode of thinking, in which they try to understand the dynamics of the task they face and solve the resulting optimization problem. Taking this route comes with the challenges of a) building simple models of complex organizational problems which could be tackled analytically, b) collecting and processing the data needed to parameterize that model reliably, and c) solving that model. Our experiments do not inform the first two steps, but show that without specific training subjects are unlikely to succeed in the third. A third, complementary, path is through valuing the intermediate states directly. Organizational measurement systems that assign value to factors other than bottom-line, such as balanced scorecard (Kaplan and Norton 1996), may help overcome the challenge of learning complex value functions by individual managers. Yet, the

value of intermediate steps would still need to be calculated before it can be embedded in organizational measurement systems or culture. Therefore, this path largely helps role out solutions found through other methods, but cannot fully address the challenge on its own.

Implications for understanding heterogeneity in strategies and performance

Scholars of organizations have long debated the sources of heterogeneity in firms' strategies and outcomes in similar markets (Gibbons and Henderson 2012; Syverson 2011). The resource based framework has advanced a theoretical understanding of such heterogeneity rooted in variations of organizational resources and capabilities that are valuable, irreplaceable, rare, and hard to imitate (Barney 1991). It is well recognized that organizational resources may induce heterogeneity if their development and usage is ambiguous, path-dependent, complex, or subject to rapid environmental change (Dierickx and Cool 1989; Levinthal 1997; Teece 2007). The literature has assumed a corollary to those requirements: that simpler managerial tasks where the underlying technology and market conditions are stable and well-understood are unlikely to be relevant for understanding heterogeneity among organizations. After all the solutions to those problems are easy and managers would learn or imitate the best solution across the board, removing heterogeneity. However, our findings illustrate that differences can persist even in simple dynamic tasks with a small state space and full information about the payoff function. Despite ample opportunity to identify the optimal strategy, most participants fail to do so. Various additional information and hints do little to change the outcome when they do not support the dominant learning heuristics managers bring to the task. These results add to a recent literature that shows that differences in performance can emerge even in simple trading settings without entry barriers and need for special knowledge (Levine et al. 2017).

Overall, three sources seem relevant in explaining heterogeneity in performance in dynamic tasks. First, luck matters. Whether early exploratory steps would expose a learner to the benefits of insisting on the long-term option is as much a function of insight as it requires luck. Second,

differences in insights or ability to think through a problem, i.e., differences in what can be called managerial cognition (Helfat and Peteraf 2015), would also matter. For example, learners advised about the long-term value of @ choices (in Direction Tip condition), performed better than others with a minimal intervention; and some 15% of subjects figured out the optimal solution in our One-shot condition (though we can't rule out luck for this group). Third, heterogeneity may also be informed by the fit between managerial learning heuristics and the task at hand. For example, one can argue that, in light of ambiguity and complex state-space, human learners do a great job in the classical version of Harvard task (Sims et al. 2013). Policy optimization heuristic offers a viable, and simple, path to discovering a policy of only choosing @. However, one could conceive of tasks that are much harder to deal with effectively using policy optimization heuristics. The issue of fit between common learning heuristics and the structure of dynamic managerial tasks thus brings up an important area of research for scholars of behavioral strategy: what are the other learning heuristics, beyond policy optimization or elaborating on its features, that people may apply in dynamic tasks? And what are the structures of common dynamic managerial tasks? When optimal strategy can be approximated by simple rules intuitively explored in policy optimization, learned heuristics may perform well, even compared to sophisticated algorithms (Sims et al. 2013). In contrast, when optimal strategies are complex and non-monotonic mappings from various states to organizational choices, adopted strategies may fall significantly short of the optimal (but then if nobody can find the optimal strategy, who can tell which strategy is subpar?) Exploring the fit between common heuristics and realistic dynamic tasks can help us distinguish tasks where managers can evolve simple rules for analytically daunting problems (Sull and Eisenhardt 2015) from those where even well understood managerial insights do not stick (Repenning 2002).

Limitations and Future Research Directions

Our experimental study focused on a few variants of the Harvard task. The strength of this experimental paradigm is in capturing temporal tradeoffs in a simple, analytically tractable, setting.

Within this paradigm we only leveraged the payoff function most often used in previous research, but other functions can offer opportunities to deepen our understanding of how people learn in dynamic tasks, and which types of tasks lend themselves to effective human learning. Specifically, teasing out if people follow heuristics that are close to dynamic programming, policy optimization, or another solution approach is an important area for future research. This can partially be achieved by collecting more process data (e.g. using verbal protocols, by allowing participants to change the state of the system for a price at various points in the game, or by collecting post-experiment detailed debriefs). Other promising pathways include using payoff functions with smaller state spaces but more complex optimal strategies. Dynamic programming approximations should be rather robust to such changes whereas policy optimization methods would falter. A third set of interventions could be focused on finding ways to improve participants' learning in these tasks. Successful interventions not only inform our understanding of why people fail in these tasks, but also may offer promising practical implications. It would be interesting to test if more sophisticated participants, e.g. those with training in dynamic programming, would significantly outperform the general population. Participants may also benefit from having benchmark for how well they can perform (Stillwell and Tunney 2009); in the absence of such benchmarks they may have stopped exploration too early. Other visualization interventions may both simplify and make more prominent the current state of the system in the hope that participants will focus on building a value function for different levels of the state. Performance incentives in the current study were relatively strong, but not extremely so. Even stronger performance incentives, e.g. including significant tournament bonuses for best performers in each experimental condition, may stimulate increased effort and exploration among all participants. Exploration incentives could also be enhanced by tying the bonus to the performance of a final strategy the participant offers, rather than their cumulative performance over time. This set up frees participants to focus only on exploration and learning during the main task, and then offer their ideal strategy at the end. A fourth direction is in identifying individual traits that may impact learning and performance in dynamic tasks, a

promising area given observed impact of individual characteristics in other strategic choice settings (Levine et al. 2017). Fifth, we focused on interdependencies over time, it is interesting to see how results may change when interdependencies are among choices with no intertemporal dimension. Other research trajectories could examine heuristics used in realistic dynamic managerial tasks and opportunities to build better decision rules in those settings. In fact, many managerial choices do not include the number of trials offered to our subjects. Identifying how people learn in dynamic tasks when samples are more limited is an important topic for future research. Overall, persistent suboptimal performance in a simple task even when all ambiguity and complexity is removed, highlights the challenges to learning in dynamic settings and offers a robust behavioral foundation for understanding heterogeneity in the strategies that managers adopt.

Acknowledgements: We are grateful to participants in TOM 2018 conference and seminar participants at MIT, INSEAD, Chicago Booth, CMU, UIUC Geis, Yale and Cornell for providing helpful comments. We are thankful to the anonymous reviewers and the AE for their excellent feedback. Financial support for this research was provided by MIT Sloan JFRAP program.

REFERENCES

Artinger, S., T.C. Powell. 2016. Entrepreneurial failure: Statistical and psychological explanations. *Strategic Management Journal* **37**(6) 1047-1064.

Bardolet, D., C.R. Fox, D. Lovallo. 2009. Naïve diversification and partition dependence in capital allocation decisions: field and experimental evidence. *Strategic Management Journal*.

Barney, J. 1991. Firm Resources and Sustained Competitive Advantage. *Journal of Management* **17**(1) 99-120.

Beach, L.R., T.R. Mitchell. 1978. A contingency model for the selection of decision strategies. *Academy of Management Review* **3**(3) 439-449.

Bertsekas, D.P. 2007. *Dynamic programming and optimal control*, 3rd ed. Athena Scientific, Belmont, Mass.

Brehmer, B., R. Allard. 1991. Dynamic decision making: The effects of task complexity and feedback delay. J. Rasmussen, B. Brehmer, J. Leplat, eds. *New technologies and work. Distributed decision making: Cognitive models for cooperative work*. John Wiley, Oxford, England, 319-334.

Brown, J., H. Rachlin. 1999. Self-control and social cooperation. *Behavioural Processes* **47**(2) 65-72.

Busemeyer, J.R. 2002. Dynamic decision making. N.J. Smelser, P.B. Bates, eds. *International encyclopedia of the social and behavioral sciences: Methodology, mathematics and computer science*. Elsevier, Oxford, England, 3903-3908.

Cain, D.M., D.A. Moore, U. Haran. 2015. Making sense of overconfidence in market entry. *Strategic Management Journal* **36**(1) 1-18.

Cassar, G., H. Friedman. 2009. Does self-efficacy affect entrepreneurial investment? *Strategic Entrepreneurship Journal* **3**(3) 241-260.

Cohen, W.M., D.A. Levinthal. 1990. Absorptive Capacity: A New Perspective on Learning and Innovation. *Adminstrative Sceince Quareterly* **35** 128-152.

Cronin, M.A., C. Gonzalez, J.D. Sterman. 2009. Why don't well-educated adults understand accumulation? A challenge to researchers, educators, and citizens. *Organizational Behavior and Human Decision Processes* **108**(1) 116-130.

Csaszar, F.A., D. Laureiro-Martínez. 2018. Individual and organizational antecedents of strategic foresight: A representational approach. *Strategy Science* **3**(3) 513-532.

Denrell, J. 2007. Adaptive learning and risk taking. *Psychological Review* **114**(1) 177-187.

Denrell, J., G. Le Mens. 2020. Revisiting the competency trap. *Industrial and Corporate Change* **29**(1) 183-205.

Denrell, J., J.G. March. 2001. Adaptation as information restriction: The hot stove effect. *Organization Science* **12**(5) 523-538.

Dierickx, I., K. Cool. 1989. Asset Stock Accumulation and Sustainability of Competitive Advantage. *Management Science* **35**(12) 1504-1511.

Erev, I., E. Ert, A.E. Roth, E. Haruvy, S.M. Herzog, R. Hau, R. Hertwig, T. Stewart, R. West, C. Lebiere. 2010. A Choice Prediction Competition: Choices from Experience and from Description. *Journal of Behavioral Decision Making* **23**(1) 15-47.

Fang, C. 2012. Organizational Learning as Credit Assignment: A Model and Two Experiments. *Organization Science* **23**(6) 1717-1732.

Fu, W.-T., J.R. Anderson. 2008. Solving the credit assignment problem: explicit and implicit learning of action sequences with probabilistic outcomes. *Psychological research* **72**(3) 321-330.

Gavetti, G. 2012. Toward a Behavioral Theory of Strategy. *Organization Science* **23**(1) 267-285.

Ghemawat, P. 1991. *Commitment : the dynamic of strategy*. Free Press ; Maxwell Macmillan Canada ;Toronto.

Gibbons, R., R. Henderson. 2012. Relational Contracts and Organizational Capabilities. *Organization Science* **23**(5) 1350-1364.

Gonzalez, C. 2004. Learning to make decisions in dynamic environments: Effects of time constraints and cognitive abilities. *Human Factors* **46**(3) 449-460.

Gonzalez, C., P. Fakhari, J. Busemeyer. 2017. Dynamic Decision Making: Learning Processes and New Research Directions. *Human Factors* **59**(5) 713-721.

Gonzalez, C., J.F. Lerch, C. Lebiere. 2003. Instance-based learning in dynamic decision making. *Cognitive Science* **27**(4) 591-635.

Gonzalez, C., P. Vanyukov, M.K. Martin. 2005. The use of microworlds to study dynamic decision making. *Computers in Human Behavior* **21**(2) 273-286.

Gureckis, T.M., B.C. Love. 2009a. Learning in noise: Dynamic decision-making in a variable environment. *Journal of Mathematical Psychology* **53**(3) 180-193.

Gureckis, T.M., B.C. Love. 2009b. Short-term gains, long-term pains: How cues about state aid learning in dynamic environments. *Cognition* **113**(3) 293-313.

Helfat, C.E., M.A. Peteraf. 2015. Managerial Cognitive Capabilities and the Microfoundations of Dynamic Capabilities. *Strategic Management Journal* **36**(6) 831-850.

Herrnstein, R.J., G.F. Loewenstein, D. Prelec, W. Vaughan. 1993. Utility Maximization and Melioration - Internalities in Individual Choice. *Journal of Behavioral Decision Making* **6**(3) 149-185.

Hey, J.D., J.A. Knoll. 2011. Strategies in dynamic decision making–An experimental investigation of the rationality of decision behaviour. *Journal of Economic Psychology* **32**(3) 399-409.

Hotaling, J.M., P. Fakhari, J.R. Busemeyer. 2015. Dynamic decision making. *International encyclopedia of the social & behavioral sciences* 709-714.

Kaplan, R.S., D.P. Norton. 1996. Using the balanced scorecard as a strategic management system. *Harvard Business Review* **74**(1) 75-&.

Kunc, M.H., J.D.W. Morecroft. 2010. MANAGERIAL DECISION MAKING AND FIRM PERFORMANCE UNDER A RESOURCE-BASED PARADIGM. *Strategic Management Journal* **31**(11) 1164-1182.

Laverty, K.J. 1996. Economic ''short-termism'': The debate, the unresolved issues, and the implications for management practice and research. *Academy of Management Review* **21**(3) 825-860.

Leiblein, M.J., J.J. Reuer, T. Zenger. 2018. What makes a decision strategic? *Strategy Science* **3**(4) 558-573.

Levine, S.S., M. Bernard, R. Nagel. 2017. Strategic Intelligence: The Cognitive Capability to Anticipate Competitor Behavior. *Strategic Management Journal* **38**(12) 2390-2423.

Levinthal, D.A. 1997. Adaptation on rugged landscapes. *Management Science* **43**(7) 934-950.

Levinthal, D.A. 2011. A Behavioral Approach to Strategy-What's the Alternative? *Strategic Management Journal* **32**(13) 1517-1523.

Levinthal, D.A., J.G. March. 1993. The Myopia of Learning. *Strategic Management Journal* **14** 95-112.

Lindblom, C.E. 1959. The Science of Muddling Through. *Public Administration Review* **19**(2) 79-88.

Lyneis, J., J. Sterman. 2016. How to Save a Leaky Ship: Capability Traps and the Failure of Win-Win Investments in Sustainability and Social Responsibility. *Academy of Management Discoveries* **2**(1) 7-32.

Mason, W., S. Suri. 2012. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior research methods* **44**(1) 1-23.

Menon, A. 2018. Bringing cognition into strategic interactions: S trategic mental models and open questions. *Strategic Management Journal* **39**(1) 168-192.

Minsky, M. 1961. Steps towards artificial intelligence. *Proceedings of the Institute for Radio Engineers* **49**(1) 8-30.

Moxnes, E. 1998. Not only the tragedy of the commons: Misperceptions of bioeconomics. *Management Science* **44**(9) 1234-1248.Newell, A., H.A. Simon. 1972. *Human problem solving*. Prentice-Hall Englewood Cliffs, NJ.Osman, M. 2010. Controlling Uncertainty: A Review of Human Behavior in Complex Dynamic Environments. *Psychological Bulletin* **136**(1) 65-86.

Paolacci, G., J. Chandler, P.G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making* **5**(5) 411-419.

Pfeffer, J. 1994. Competitive advantage through people. *California management review* **36**(2) 9.

Postrel, S., R.P. Rumelt. 1996. Incentives, routines and self-command. *Industrial and Corporate Change* **1**(3) 397-425.

Prelec, D. 2014. Consuming at the Wrong Rate: Lessons from the Harvard Game. U. Alistair, D. Southerton, eds. *Sustainable Consumption: Multi-disciplinary Perspectives in Honour of Professor Sir Partha Dasgupta*. Oxford University Press, USA, 161-174.

Prelec, D., R.J. Herrnstein, W.J. Vaughan. 1986. An intra-personal prisoners' dilemma *IX Symposium on the Quantitative Analysis of Behavior*.

Rahmandad, H., M.S. Gary. Forthcoming. Delays impair learning and can drive convergence to inefficient strategies. *Organization Science*.

Rahmandad, H., N. Repenning. 2016. Capability erosion dynamics. *Strategic Management Journal* **37**(4) 649-672.

Rahmandad, H., Z. Ton. 2020. If Higher Pay Is Profitable, Why Is It So Rare? Modeling Competing Strategies in Mass Market Services. *Organization Science*.

Repenning, N.P. 2002. A simulation-based approach to understanding the dynamics of innovation implementation. *Organization Science* **13**(2) 109-127.

Repenning, N.P., J.D. Sterman. 2002. Capability traps and self-confirming attribution errors in the dynamics of process improvement. *Administrative Science Quarterly* **47**(2) 265-295.

Shapira, Z., J.M. Shaver. 2014. Confounding changes in averages with marginal effects: How anchoring can destroy economic value in strategic investment assessments. *Strategic Management Journal* **35**(10) 1414-1426.

Simon, D.A., N.D. Daw. 2011. Neural correlates of forward planning in a spatial decision task in humans. *Journal of Neuroscience* **31**(14) 5526-5539.

Sims, C.R., H. Neth, R.A. Jacobs, W.D. Gray. 2013. Melioration as Rational Choice: Sequential Decision Making in Uncertain Environments. *Psychological Review* **120**(1) 139-154.

Stein, J.C. 1989. Efficient Capital-Markets, Inefficient Firms - a Model of Myopic Corporate-Behavior. *Quarterly Journal of Economics* **104**(4) 655-669.

Sterman, J.D. 1989. Misperception of feedbak in dynamic decision making. *Organizational Behavior and Human Decision Processes* **43** 301-335.

Stillwell, D.J., R.J. Tunney. 2009. Melioration behaviour in the Harvard game is reduced by simplifying decision outcomes. *Quarterly Journal of Experimental Psychology* **62**(11) 2252-2261.

Sull, D.N., K.M. Eisenhardt. 2015. *Simple rules : how to thrive in a complex world*. Houghton Mifflin Harcourt, Boston.

Sutton, R.S., A.G. Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Syverson, C. 2011. What Determines Productivity? *Journal of Economic Literature* **49**(2) 326-365.

Teece, D.J. 2007. Explicating dynamic capabilities: The nature and microfoundations of (sustainable) enterprise performance. *Strategic Management Journal* **28**(13) 1319-1350.

Tunney, R.J., D.R. Shanks. 2002. A re-examination of melioration and rational choice. *Journal of Behavioral Decision Making* **15**(4) 291-311.

Walsh, M.M., J.R. Anderson. 2011. Learning from delayed feedback: neural responses in temporal credit assignment. *Cognitive, Affective, & Behavioral Neuroscience* **11**(2) 131-143.

Warry, C.J., B. Remington, E.J.S. Sonuga-Barke. 1999. When more means less: Factors affecting human self-control in a local versus global choice paradigm. *Learning and Motivation* **30**(1) 53-73.

Figures

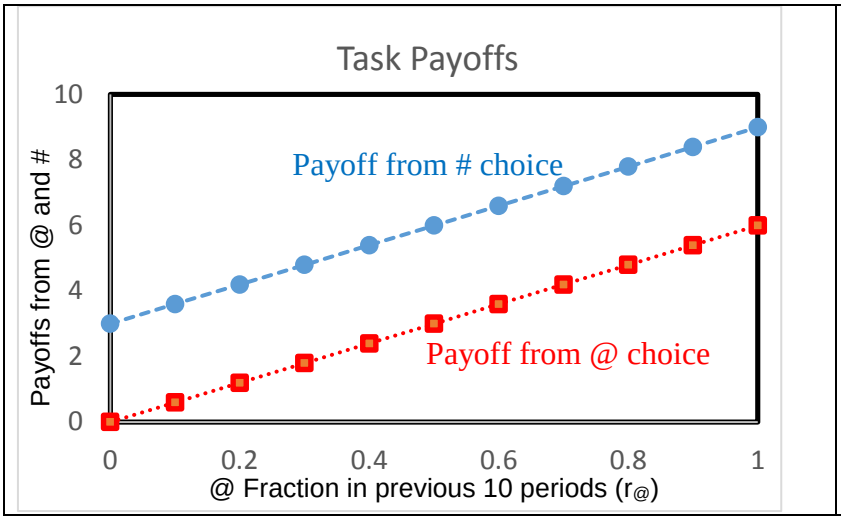

Figure 1- Payoff as a function of # (dashed blue circle) and @ (dotted red square) choices and the fraction of previous N=10 choices that were @ ($r_{@}$)

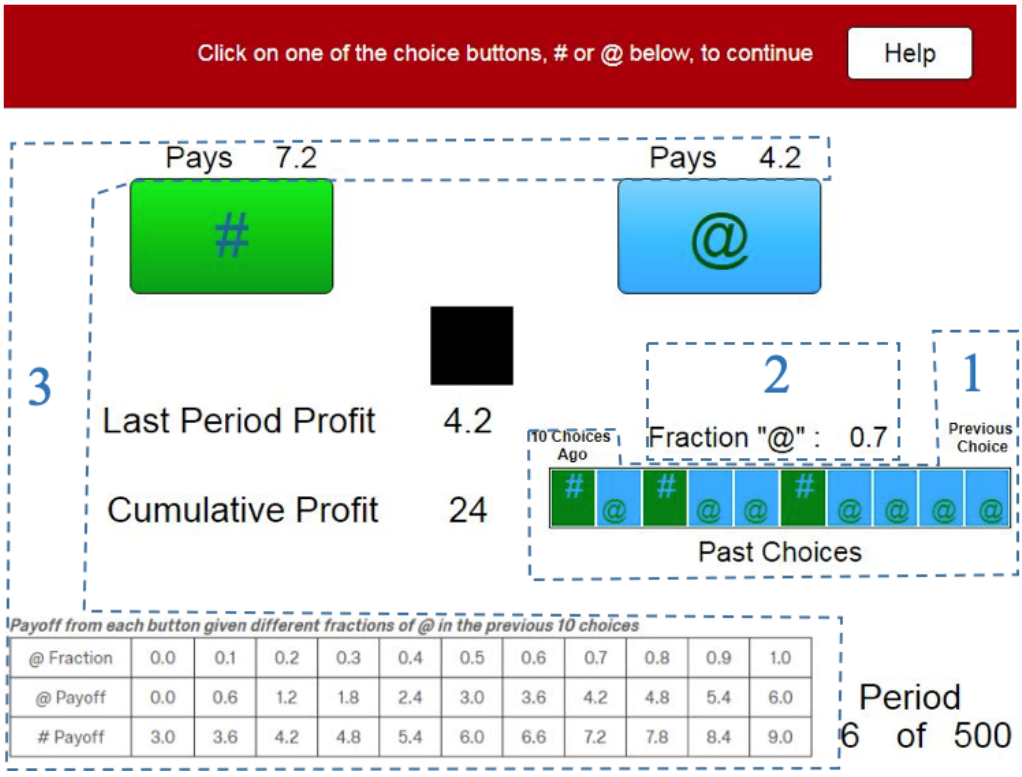

Figure 2- Experimental interface in Full Knowledge condition. Three components identified by blue dashed lines and marked 1, 2 and 3 are varied across the first four conditions. Classical excludes all three; History includes only 1; Informed History includes 1 and 2; 3 is unique to Full Knowledge.

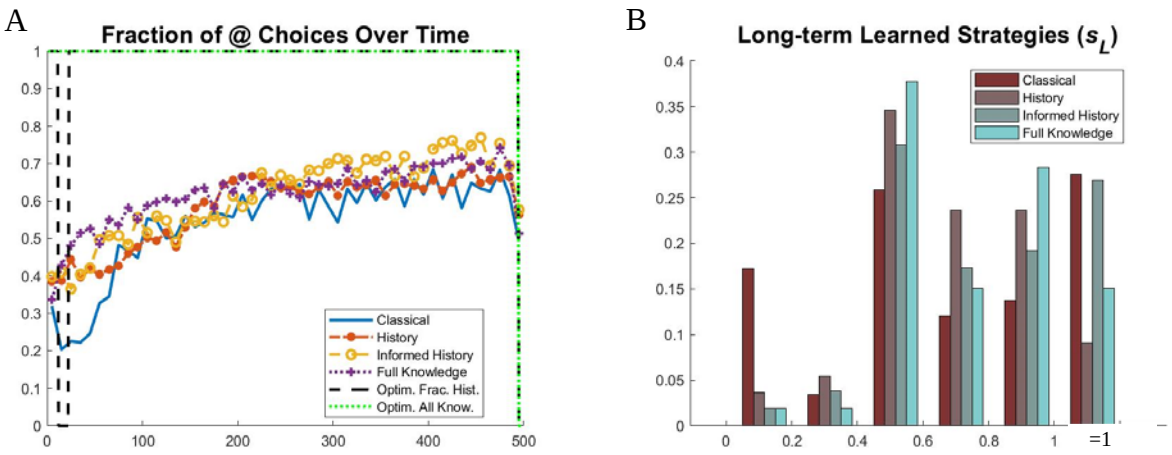

Figure 3-Learning outcomes in experiment 1. A) Actions over time, averaged across participants in condition and 10-period intervals (except for optimal actions which are graphed every period) B) Long-term learned strategies (i.e. actions averaged between 401-490 periods). The columns on the right represent those with $s_L=1$.

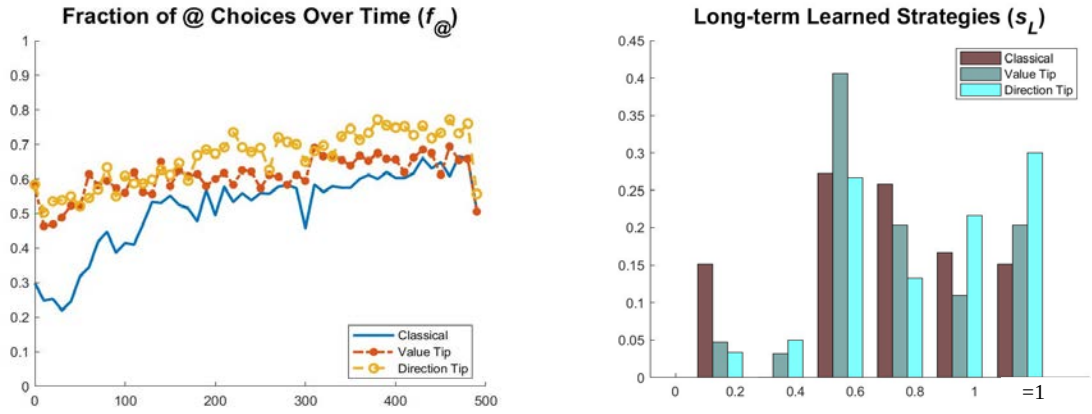

Figure 4- Performance over time (left) and long-term learned strategies (right) in experiment 2

Tables

Table 1-Challenges in different conditions of experiment 1

Challenge \ Conditions	Classical	History	Informed History	Full Knowledge
Unknown dynamics	Yes	**No**	**No**	**No**
Computationally Complex	Yes	Yes	**No**	**No**
Credit Assignment Problem	Yes	Yes	Yes	**No**
Deriving Optimal Solution	Yes	Yes	Yes	Yes
Optimal Policy Available	No	Approximately	Yes	Yes
Length in Periods	500	500	500	500

Table 2-Overview of outcome measures in experiment 1. Those different from Classical at p≤0.1 in a T-test are bold.

Measure \ Condition	**Classical**	**History**	**Info. Hist.**	**Full Know.**
Cumulative Profits (π)	2318 (409)	2373 (306)	2421 (302)	**2432 (302)**
Exploration Score (*xs*)	0.33 (0.67)	0.24 (0.56)	**0.53 (0.45)**	**0.087 (0.77)**
Learned Strategy, Short (s_{SH})	0.4 (0.37)	0.43 (0.29)	**0.5 (0.31)**	**0.54 (0.23)**
Learned, Long Strategy (s_L)	0.64 (0.34)	0.66 (0.23)	**0.74 (0.25)**	0.7 (0.23)
$e_{MM\#}$	3.18 (2.01)	3.58 (2.26)	3.11 (1.63)	
$e_{MM@}$	3.21 (2.46)	3.26 (2.2)	2.98 (2.36)	
Number of Participants	60	55	52	54

Table 3- Long-term learned strategies and performance as a function of conditions in experiment 1 (Informed History is reference condition) and process controls

Long-term Learning and Performance Outcomes				
	(1)	(2)	(3)	(4)
	s_L	π	s_L	π
Condition = Classical	-0.099	-103	-0.056	-50.1
	(0.051)	(63.9)	(0.048)	(60.8)
Condition = History	-0.077	-47.4	-0.019	37.5
	(0.051)	(64.7)	(0.050)	(63.5)
Condition = Full Knowledge	-0.033	11.3		
	(0.052)	(65.3)		
Qualification Score			0.026	12.8
			(0.024)	(30.3)
Exploration Score (xs)			0.24	302
			(0.033)	(42.4)
Mental Model Error # (eMM#)			-0.0046	-7.70
			(0.011)	(14.3)
Mental Model Error @ (eMM@)			-0.0076	-6.12
			(0.0097)	(12.3)
Constant	0.74	2,421	0.55	2,256
	(0.037)	(46.4)	(0.11)	(133)
Observations	218	218	162	162
R-squared	0.021	0.018	0.275	0.263

Standard errors in parentheses

Table 4-Various outcome measures in experiment 2.

Condition / Measure	**Classical**	**Value Tip**	**Direction Tip**
Cumulative Profits (π)	2271 (356)	2394 (301)	2482 (326)
Exploration Score (*xs*)	0.38 (0.61)	0.46 (0.56)	0.37 (0.61)
Learned Strategy, Short (s_{SH})	0.38 (0.34)	0.56 (0.29)	0.56 (0.3)
Learned Strategy, Long (s_L)	0.63 (0.3)	0.65 (0.25)	0.74 (0.26)
Number of Participants	66	62	59

Title: What makes dynamic strategic problems difficult? Evidence from an experimental study

Running head: What makes dynamic strategic problems difficult?

Authors:

Hazhir Rahmandad*
Associate Professor of System Dynamics, MIT Sloan School of Management
Room E62-442, 100 Main St., Cambridge, MA 02142
hazhir@mit.edu, +1-617-258-8912, orcid: 0000-0002-2784-9042

Jerker Denrell

Professor of Behavioral Science, Warwick Business School, University of Warwick, Scarman Road, Coventry CV4 7AL, United Kingdom.

denrell@wbs.ac.uk, +44 (24) 76522119, orcid: 000-0001-9628-1924

Drazen Prelec
Massachusetts Institute of Technology, Sloan School of Management, Department of Economics, Department of Brain and Cognitive Sciences,
Cambridge, MA, USA
Room E62-54, 100 Main St., Cambridge, MA 02142
dprelect@mit.edu, +1-617-253-2833

*: Corresponding Author
Keywords: Learning, managerial cognition, Dynamic Decision-making, Persistent Performance Differences, Experiment