Dynamic Decisions with Short-term Memories

Li, Hao

University of Toronto

Sumon Majumdar

Queen's University

July 2, 2005

Abstract: A two armed bandit problem is studied where the decision maker can only recall the most recent outcomes of his past decisions. Optimal learning strategies are shown to involve random and periodic experimentation (choosing the risky arm). We find that any optimal strategy is necessarily time inconsistent, unless it calls for experimentation with probability one or zero regardless of history. We show through an example that the decision maker can benefit from memory manipulation, i.e. not recording outcomes accurately.

Acknowledgments: We thank Tilman Börgers and James Dow for comments.

I. Introduction

Learning in a society is often generational. Each generation can make their own decisions, sometimes against the advice of previous generations, as they can be skeptical of the experiences and stories of previous generations as told to them by their parents. In this paper we study issues of time inconsistency and memory manipulation in "generational learning."

The formal model we use is a single decision maker with short term memories playing a two armed bandit. We choose the two armed bandit problem, with one safe arm of a known, constant period payoff and one risky arm of a stochastic period payoff and an unknown mean, because it is a welled studied dynamic learning model. In each period the decision maker chooses whether or not to experiment, i.e. play the risky arm. By "short term memories," we mean that the decision maker can only recall the payoff associated with his last period decision. We use this particular form of imperfect recall to capture an essential feature of generational learning. Clearly, without imperfect recall, the issues of time inconsistency and memory manipulation would not arise.

We show that optimal learning strategies generally involve random and periodic experimentation. The probability of experimentation after getting an unfavorable payoff from the risky arm can be strictly between zero and one. Without such randomization, the decision maker with short term memories would be forced to make a tough choice between stopping experimentation right after the first unfavorable payoff form the risky arm and continuing experimentation even after repeated negative information about the risky arm. Instead, the optimal strategy carefully calibrates the probability of experimentation to balance the need to engage in some experimentation and the need to respond to negative information. In periodic experimentation, the decision maker adopts a positive probability of resuming experimenting after having drawn the safe arm in the previous period. Optimal strategies require the right combinations of periodic experimentation with random experimentation as a response to the constraint of short term memories. It turns out that there is always an optimal strategy that uses random experimentation alone, so periodic experimentation can be optimal but is not implied by optimality.

In deriving optimal learning strategies, we assume that at the start of the time the decision maker can commit not to modify his plan alone the entire learning process. Such commitment is extreme in the context of generational learning. We show that optimal strategies are generally time inconsistent if the decision maker is introspective in spite of the constraint of short term memories. That is, generally there exists a sequence of events along the path under any optimal strategy at which point the decision maker would want to change his experimental decision, if he updates his belief about the risky arm based on the experience that he recalls and the knowledge that he has acted according to the optimal strategy. Only when an optimal strategy calls for experimentation with probability one or zero regardless of information, would it be time consistent. This happens when the prior about the risky arm is either extremely optimistic or extremely pessimistic, so that the decision maker optimally disregards any information. Thus, responsiveness to new information and incentive to deviate from ex ante optimal learning go hand-in-hand in generational learning.

Random and periodic experimental decisions in an optimal learning strategy reflects the need of the decision maker to "manage" memory in order to retain the flexibility in how to make use of information. This raises the possibility that the decision maker may benefit from "memory manipulation" in the sense of not recording his experience truthfully. Of course we assume that the decision maker knows how he would want to manipulate his past experience, so a simple relabeling has no effect on his ex ante welfare. We demonstrate through an example how memory manipulation can work without assuming that the decision maker engages in any form of self deception. In this example, the decision maker has a positive probability of retaining the "clean slate" of null history instead of recording the outcome resulting from the most recent experimentation. This form of memory manipulation allows the decision maker to enrich the state space of his strategy, and helps improve his ex ante welfare by responding better to new information.

The two armed bandit problem without the short term memory constraint is a simple example of a class of problems studied by Gittins (1989). The short term memory constraint considered here is a type of complexity constraints that focus limited memory. See Lipman (1995) for a survey of the literature. A standard way of modeling limited

memory is a finite automaton, which consists of a finite set of memory states, an action rule that maps the set of states to a finite sets of choices, and a transition rule that maps the set of states and a finite set of outcomes to the set of states. See Rubinstein (1986) for an application of finite automata to repeated games and for references to the literature on finite automata. A feasible strategy for our decision maker with short term memories can be thought of as a finite automaton with the set of period payoffs as the set of states and the fixed transition rule that gives the state of the next period as the current payoff resulting from the current choice. To our knowledge, there is no work on finite automata playing bandits; the closest is a paper by Börgers and Morales (2004), who have a bandit model with perfectly revealing outcomes and limited scope for learning. The present paper is motivated by issues of time consistency and memory manipulation in generational learning, and we find the assumption of short term memories more natural than generic finite automata. In particular, the issue of memory manipulation can not be addressed with a finite automaton approach, as the meaning of each memory state is optimally chosen and is therefore endogenous with such an approach.[1] Our assumption of short term memories is a form of imperfect recall. The need for randomization and the problem of time inconsistency under imperfect recall have already been pointed out by Piccione and Rubinstein (1997).[2] We add to this literature by characterizing the solution to a well-studied dynamic learning model under an intuitive constraint on memory capacity and providing its time consistency properties.

This paper is organized as follows. In the next section, we describe the two armed bandit problem with short term memories. In section 3 we characterize optimal learning strategies and show that random experimentation and periodic experimentation can be optimal. In section 4 we show that any optimal strategy is necessarily time-inconsistent,

[1] The same is true for decisions models with limited memory and one time decisions, such as Wilson (2004). In Meyer (1991), decisions are also one time, but memory states have fixed, exogenous meanings. Her characterization of optimal recording of outcomes in coarse learning shares similarities with what we call memory manipulation here. Note that the issue of time inconsistency does not arise in these models as they involve once for all decisions.

[2] Studies of randomization under imperfect recall go back to Kuhn (1953). Kalai and Salon (2003) define "non interactive" Markov decision problems, and show that under imperfect recall optimal strategies generally require randomization, but not in the action rule. Our two armed bandit problem is interactive because the decision maker controls the set of the possible outcomes through his choice in each period.

unless it calls for experimentation with probability one or zero regardless of history. In section 5 we show through an example that memory management can improve the ex ante welfare of the decision maker. Section 6 lists some topics for further research. Detailed proofs can be found in the appendix.

1. A Two Armed Bandit Problem with Short Term Memories

Consider an infinite horizon two armed bandit problem, with discrete time $t = 1, \ldots, \infty$. A safe arm gives a certain period payoff of 0. A risky arm has either high average payoffs (state h) or low payoffs (state l), with the decision maker's prior probability in period 0 equal to η for state h. We assume a symmetric binary signal structure from the risky arm: the normalized period payoff is either 1 or -1, with $\Pr[1|h] = \Pr[-1|l] = q$ and $\frac{1}{2} < q < 1$. In each period a decision maker must choose between the risky arm (experimentation, e) and the safe arm (stop, s). The decision maker maximizes the period 0 discounted sum of his expected utility , with a discount factor $\delta \in (0, 1)$.

Without any memory constraint, the decision maker's optimal learning strategy is given by the solution to a Bellman equation. Let p denote the current belief for state h, and $U(p)$ denote the optimal value of the decision maker's objective function. The Bellman equation for this problem is

$$U(p) = \max\{\delta U(p), (2p-1)(2q-1) + \delta(pq + (1-p)(1-q))U(p(+)) + \delta((1-p)q + p(1-q))U(p(-))\},$$

where

$$p(+) = \frac{pq}{pq + (1-p)(1-q)};$$
$$p(-) = \frac{p(1-q)}{p(1-q) + 1 - p}$$

are Bayesian updates of the belief after getting payoffs of $+1$ and -1 from the risky arm. It is straightforward to establish the following: (i) there is a unique function $U(p)$ that satisfies the Bellman equation; (ii) $U(p)$ is increasing and convex; and (iii) there exists $\hat{p} < \frac{1}{2}$ such that $U(p) = 0$ (and the optimal choice is s) if $p \leq \hat{p}$ and $U(p) > 0$ (and the optimal choice is e) if $p > \hat{p}$.

In the present paper we assume that the decision maker has short term memories in that he can remember the experience in the previous period only. To model this memory constraint, we assume that there are four memory states: null memory ($\emptyset$), a positive payoff of 1 from the risky arm ($+$), a negative payoff of -1 from the risky arm ($-$) and a payoff of 0 from the safe arm (c). Denote a memory state as $m \in \{\emptyset, +, -, c\}$. A pure strategy sends each memory state m to a choice of experiment (e) or stop (s). A behavioral strategy β maps each m to a probability β_m of playing e.[3] The decision maker chooses $(\beta_\emptyset, \beta_+, \beta_-, \beta_c)$ to maximize his period 0 discounted sum of expected utilities.

2. Optimal Learning Strategies

Fix a strategy β. Suppose that the state is h. From the perspective of period 0, the probability X_t^h of choosing the risky arm in period $t = 1, 2, \ldots$ satisfies

$$X_{t+1}^h = (1 - X_t^h)\beta_c + X_t^h(q\beta_+ + (1-q)\beta_-).$$

Denoting

$$B^h = q\beta_+ + (1-q)\beta_- - \beta_c,$$

We have

$$X_{t+1}^h = B^h X_t^h + \beta_c.$$

Using the above formula recursively and $X_1^h = \beta_\emptyset$, we obtain

$$X_t^h = \beta_\emptyset (B^h)^{t-1} + \frac{\beta_c(1 - (B^h)^{t-1})}{1 - B^h}.$$

Symmetrically, in state l from the perspective of period 0 the probability X_t^l of e is given by

$$X_t^l = \beta_\emptyset (B^l)^{t-1} + \frac{\beta_c(1 - (B^l)^{t-1})}{1 - B^l},$$

[3] Since this is a decision problem with imperfect recall, Kuhn's (1953) theorem of equivalence of behavioral and mixed strategies does not hold. A mixed strategy in our model is a period 0 randomization of pure strategies. It is easy to see that mixed strategies will not improve over pure strategies given the von-Neumann expected utility formulation here.

where

$$B^l = (1-q)\beta_+ + q\beta_- - \beta_c.$$

The expected payoff to experimentation in any period t is $2q-1$ in state h, and $-(2q-1)$ in state l. Thus, the decision maker's period 0 discounted sum of expected utilities from the strategy β is given by

$$V(\beta) = (2q-1)(\eta V^h(\beta) - (1-\eta)V^l(\beta)),$$

where

$$V^h(\beta) = \sum_{t=1}^{\infty} \delta^t X_t^h;$$
$$V^l(\beta) = \sum_{t=1}^{\infty} \delta^t X_t^l.$$

Completing the geometric sums, we have

$$V(\beta) = \delta(2q-1)\left(\beta_\emptyset + \frac{\delta\beta_c}{1-\delta}\right)\left(\frac{\eta}{1-\delta B^h} - \frac{1-\eta}{1-\delta B^l}\right).$$

An optimal strategy β maximizes $V(\beta)$ subject to $\beta_m \in [0,1]$ for each $m \in \{\emptyset, +, -, c\}$.

To characterize optimal strategies, we need the following three threshold values for the prior. Define

$$\eta_0 = \frac{1-\delta q}{2-\delta}$$

and

$$\eta_1 = q.$$

Note that η_0 and η_1 satisfy

$$\frac{1-\eta_0}{\eta_0} = \frac{1-\delta(1-q)}{1-\delta q};$$
$$\frac{1-\eta_1}{\eta_1} = \frac{1-q}{q}.$$

Since $q > \frac{1}{2}$, we have $\eta_0 < \eta_1$ and $\eta_0 < \frac{1}{2}$. Define η_* such that

$$\frac{1-\eta_*}{\eta_*} = \left(\frac{1-\eta_1}{\eta_1}\right)\left(\frac{1-\eta_0}{\eta_0}\right)^2.$$

It is straightforward to verify that $\eta_* \in (\eta_0, \eta_1)$ because $q > \frac{1}{2}$ and $\delta < 1$. Next, for each $\eta \in [\eta_*, \eta_1]$, define $K(\eta)$ such that

$$\left(\frac{1-q}{q}\right)\left(\frac{1+\delta q K(\eta)}{1+\delta(1-q)K(\eta)}\right)^2 = \frac{1-\eta}{\eta}.$$

Note that K is a strictly decreasing function in η, with $K(\eta_*) = 1/(1-\delta)$ and $K(\eta_1) = 0$.

We have the following characterization of optimal learning strategies:[4]

PROPOSITION 2.1. *An optimal strategy β satisfies: (i) (no experiment) $\beta_\emptyset = \beta_c = 0$ for $\eta \leq \eta_0$; (ii) (pure experiment) $\beta_\emptyset = \beta_+ = 1$ and $\beta_- = \beta_c = 0$ for $\eta \in (\eta_0, \eta_*]$; (iii) (random and periodic experiment) $\beta_\emptyset = \beta_+ = 1$, and β_- and β_c satisfy $(1-\beta_-)/(1-\delta(1-\beta_c)) = K(\eta)$ for $\eta \in (\eta_*, \eta_1]$; and (iv) (always experiment) $\beta_\emptyset = \beta_+ = \beta_- = 1$ for $\eta > \eta_1$.*

Thus, a pure strategy is uniquely optimal in cases (i), (ii) and (iv) above. For sufficiently pessimistic prior (case (i), $\eta \leq \eta_0$), the optimal strategy calls for no experimentation from the start and no experimentation ever. In the opposite extreme when the prior is sufficient strong (case (iv), $\eta > \eta_1$), the optimal strategy calls for experimentation from the start and continuing experimentation regardless of the payoff from the risky arm. For intermediate priors just above the *no experiment* region (case (iii), $\eta \in [\eta_0, \eta_*)$), the optimal strategy calls for initial experimentation, continuing experimentation until the first negative payoff from the risky arm and no experimentation thereafter.

The most interesting region is the intermediate priors just below the *always experiment* region. Here there exists a continuum of optimal strategies that can exhibit random and periodic experimentations. Note that random and periodic experimentations apply only to the memory states $-$ and c. From the expression of V, we can see that it is always optimal to set $\beta_\emptyset$ to 0 or 1. Further, it is intuitive that β_+ is either 0 or 1, as the memory state $+$ is the most favorable so β_+ should be set to 1 if there is a positive probability of experimentation in any memory state.[5] By "random experimentation," we mean that β_-

[4] We do not give the value of β_m for an optimal strategy if m occurs with 0 probability under the strategy. Thus, β_+ and β_- are unrestricted in case (i) below and β_c is unrestricted in case (iv).

[5] The proof of Proposition 2.1 makes this point formal by showing that the derivative of V with respect to β_+ is strictly positive whenever the derivatives of V with respect to β_c or β_- are weakly positive.

is strictly between 0 and 1, while by "periodic experimentation," we mean that β_c is great than 0. Optimal strategies require the right combinations of periodic experimentation with random experimentation, so that[6]

$$\frac{1-\beta_-}{1-\delta(1-\beta_c)} = K(\eta).$$

Since $K(\eta)$ is a decreasing, a more favorable prior about the risky arm tends to increase both β_- and β_c. However, due to multiplicity of optimal strategies, the experimentation probabilities in memory states $-$ and c are not necessarily monotone in the prior η. Instead, the two variables β_- and β_c are carefully calibrated to balance the need to engage in some experimentation and the need to respond to negative information.

Since $K(\eta)$ satisfies

$$0 \leq K(\eta) \leq \frac{1}{1-\delta}$$

for all $\eta \in [\eta_*, \eta_1]$, the constraint on β_c and β_- can always be satisfied by $\beta_c = 0$ and $\beta_- = 1-(1-\delta)K(\eta)$. Thus, there is always an optimal strategy that uses random experimentation alone. Periodic experimentation can be optimal but is not implied by optimality. On the other hand, for a range of values of η in the *random and periodic experiment* region, there is an optimal learning strategy that does not use random experimentation. Define η_{**} such that $K(\eta_{**}) = 1$, or

$$\left(\frac{1-q}{q}\right)\left(\frac{1+\delta q}{1+\delta(1-q)}\right)^2 = \frac{1-\eta_{**}}{\eta_{**}}.$$

Since $K(\eta)$ is a decreasing function, we have $\eta_* < \eta_{**} < \eta_1$ and $1 \leq K(\eta) \leq 1/(1-\delta)$ for all $\eta \in [\eta_*, \eta_{**}]$. Then, for all $\eta \in [\eta_*, \eta_{**}]$, we can find $\beta_c \in [0,1]$ such that

$$\frac{1}{1-\delta(1-\beta_c)} = K(\eta).$$

Thus, in this range random experimentation can be optimal but is not implied by optimality.

[6] Since $\beta_+ = 1$, how frequent a learning strategy plays the risky arm is determined by β_- and β_c. Intuitively, the ratio $(1-\beta_-)/(1-\delta(1-\delta_c))$ measures how frequent the learning strategy plays the safe arm. The constraint on β_- and β_c below shows that β_- and β_c matter only through their effects on this ratio.

3. Time Inconsistency

In this section we ask whether any of the optimal strategies characterized in the previous section is time consistent. To answer this question, we need to assume that the decision maker is "introspective" in spite of the short memory constraint. This assumption requires that the decision maker remember the strategy he is carrying out, and be capable of updating his belief about the risky arm based on the memory state and the knowledge that he has acted according to the optimal strategy. The issue of time consistency of an optimal strategy then reduces to the question of whether there is a memory state along the path at which the decision maker wants to deviate from his the prescribed choice if his updated belief is taken as the prior.

The short term memory constraint means that the decision maker can not recall the calendar time except at the very first period, i.e. when the memory state m is $\emptyset$. Thus, we have $\Pr[h|\emptyset] = \eta$, and there remain three updated beliefs to compute, $\Pr[h|m]$ for $m = +, -, c$. To define how the belief about risky arm is updated under any given strategy β, we use the concept of "consistent beliefs" a la Piccione and Rubinstein (1997). The idea is to use the "Bayes rule" to compute the updated beliefs along the path implied by β, even though the constraint of short term memory implies that the numbers assigned to events are not probability numbers because they can exceed 1. Further, due to an infinite horizon in our model, these numbers can be infinity. We resolve this issue by introducing a small probability τ in every period that the decision problem terminates in that period after the choice between e and s is made, and then take τ to zero in the limit.[7] Then, we have

$$\Pr[h|+] = \lim_{\tau \to 0} \frac{\eta \sum_{t=1}^{\infty} \tau(1-\tau)^t q X_t^h}{\eta \sum_{t=1}^{\infty} \tau(1-\tau)^t q X_t^h + (1-\eta) \sum_{t=1}^{\infty} \tau(1-\tau)^t (1-q) X_t^l}.$$

The interpretation is the decision maker assesses the belief about the risky arm conditional on that the decision problem has stopped and the memory state is $+$. Using the expressions for X_t^h and X_t^l and taking the limit, we have

$$\Pr[h|+] = \frac{\eta q(1-B^l)}{\eta q(1-B^l) + (1-\eta)(1-q)(1-B^h)}.$$

[7] We are inspired by Wilson's (2004) model of limited memory capacity with one time decisions and an exogenous termination probability.

Similar calculations lead to

$$\Pr[h|-] = \frac{\eta(1-q)(1-B^l)}{\eta(1-q)(1-B^l) + (1-\eta)q(1-B^h)},$$

and

$$\Pr[h|c] = \frac{\eta(1-\beta_c - B^h)}{\eta(1-\beta_c - B^h) + (1-\eta)(1-\beta_c - B^l)}.$$

We have the following result regarding time consistency of optimal learning strategies.

PROPOSITION 3.1. *An optimal strategy for prior η is time consistent if and only if $\eta \in [0, \eta_0] \cup [\eta_1, 1]$.*

One can easily verify that

$$\left(\frac{1-q}{q}\right)\left(\frac{1-B^h}{1-B^l}\right) \leq 1$$

for any β, with equality if and only if $\beta_- = 1$ and $\beta_c = 0$. Therefore,

$$\Pr[h|+] \geq \eta$$

and the decision maker always becomes more optimistic about the risky arm after a positive payoff regardless the strategy he is using (not just the optimal strategies). Note that the optimal strategies given by Proposition 2.1 have the properties that β_+ is either 0 or 1, and whenever $\beta_+ = 1$ for some η then $\beta_\emptyset = 1$ for higher priors. Since a positive payoff never depresses the decision maker's belief, if an optimal strategy calls for experimentation after a positive payoff, he would not want to change the decision if he takes the updated belief as his prior. Therefore, the issue of time inconsistency does not arise after a positive payoff from experimentation.

Time consistency issue does not necessarily arise after a negative payoff from the risky arm. When the decision maker starts with a very optimistic belief (in the *always experiment* region, it turns out that his updated belief after a negative payoff remains sufficiently upbeat so he will not deviate from the prescribed choice of e based on the updated belief. However, time consistency problem occurs for all intermediate values of the prior, for different reasons, depending on whether the prior is in the *pure experiment*

region or the *random and periodic experiment* regions. In the *pure experiment* region, the decision maker is supposed to stop at the first instance of a negative payoff, but the updated belief would suggest experimentation is optimal. In fact, the updated belief is equal to the prior η—according to the optimal strategy in this region, the first negative payoff could be either after a series of positive payoffs from the risky arm, which would lead to a rather favorable belief, or actually the first payoff, which would result in an unfavorable belief. The situation in the *random and periodic experiment* region is more complicated. Essentially, since the probability of experimentation at the beginning of the decision process (i.e. for the null history $\emptyset$) is either 0 or 1 in any optimal strategy, random and periodic experimental decisions after getting a negative payoff from the risky arm or drawing the safe arm or can not be time consistent.

Thus, an optimal strategy is time consistent only in the *never experiment* and *always experiment* regions. These two regions are precisely where the decision maker does not respond to new information, and there is no learning going on. In our model of dynamic decisions with short term memory, optimal learning and time consistency are necessarily linked to each other. Since η_0 decreases with q and η_1 increases with q, the incidence of time inconsistency in optimal learning increases with the quality of signal. Further, since η_0 decreases with δ, time inconsistency in optimal learning is more likely to arise with a more patient decision maker.

4. Memory Manipulation

If we model the behavioral strategies of the decision maker with short term memories as finite automata, then we have considered only varying the action rule while exogenously fixing the transition rule from a memory state to another. However, the characterization of optimal learning strategies in Proposition 2.1, and in particular, random and periodic experimentations, strongly suggests the decision maker may want to vary the transition rule as well. In our model of two armed bandit with the short term memory constraint, optimizing over the transition rule amounts to manipulating the meanings of memory states.

In general, different forms of memory manipulation may be considered. For example, the decision maker may record a negative payoff from the risky arm as a positive payoff. Since we assume that the decision maker can recall his own strategy, including possible manipulations of memory states, a relabeling of memory states will not have any effect. In this section we consider incentives of the decision maker not to replace the memory state at the start of the period, which is the experience from the choice made in the previous period, with the experience resulting from the current period decision. This may be thought of as "endogenous forgetfulness." In particular, we investigate whether the decision maker can improve his period 0 discounted sum of expected utilities by retaining the "clean slate" of null history (i.e., the memory state $\emptyset$) instead of recording the payoff from the most recent experimentation. The interpretation in generational learning would be that the generation that has made their choice does not always admit this to the next generation of decision makers.

Formally, when the beginning of period memory state is $\emptyset$, for each current period outcome $m \in \{+, -, c\}$, let γ_m be the probability of replacing the memory state $\emptyset$ with m. Memory manipulation with respect to the null history state $\emptyset$ occurs when $\gamma_m < 1$ for some $m \in \{+, -, c\}$. We assume that there is no kind of memory manipulation, so that when the beginning of period memory state is any m other than $\emptyset$, with probability 1 the decision maker replaces m with the current period outcome. Denote $\gamma = (\gamma_+, \gamma_-, \gamma_c)$. The decision maker now chooses γ as well as β to maximize $W(\beta; \gamma)$, the period 0 discounted sum of expected utilities.

Fix a strategy β and γ. Suppose that the state is h. Let P_t^h, N_t^h, Z_t^h and F_t^h be the ex ante probability (i.e., from period 0 perspective) of the memory state $+$, $-$, c and ϕ, respectively, at the beginning of period t, $t = 1, 2, \ldots$, before the experimental decision and memory manipulation. The evolution of $(P_t^h, N_t^h, Z_t^h, F_t^h)$ is determined by the following transition matrix:

$$\begin{array}{c} \\ P_{t+1}^h \\ N_{t+1}^h \\ Z_{t+1}^h \\ F_{t+1}^h \end{array} \begin{array}{c} \begin{array}{cccc} P_t^h & N_t^h & Z_t^h & F_t^h \end{array} \\ \begin{bmatrix} \beta_+ q & \beta_- q & \beta_c q & \beta_\emptyset q \gamma_+ \\ \beta_+(1-q) & \beta_-(1-q) & \beta_c(1-q) & \beta_\emptyset(1-q)\gamma_- \\ 1-\beta_+ & 1-\beta_- & 1-\beta_c & (1-\beta_\emptyset)\gamma_c \\ 0 & 0 & 0 & \Lambda^h \end{bmatrix} \end{array}$$

where

$$\Lambda^h = (1-\beta_\emptyset)(1-\gamma_c) + \beta_\emptyset(q(1-\gamma_+) + (1-q)(1-\gamma_-)).$$

Note that $\Lambda^h = 0$ if there is no memory manipulation. The initial values are given by $P_1^h = N_1^h = Z_1^h = 0$ and $F_1^h = 1$. It follows from the transition matrix that

$$F_t^h = (\Lambda^h)^{t-1}$$

for each t.

Define

$$X_t^h = P_t^h\beta_+ + N_t^h\beta_- + Z_t^h\beta_c + F_t^h\beta_\emptyset$$

as the aggregate probability of experimentation in period t from period 0 perspective. We claim that

$$X_{t+1}^h = B^h X_t^h + \beta_c + G^h F_t^h$$

for each $t \geq 1$, where B^h is as defined in section 3 and

$$G^h = (\beta_\emptyset - \beta_+)\beta_\emptyset q(1-\gamma_+) + (\beta_\emptyset - \beta_-)\beta_\emptyset(1-q)(1-\gamma_-) + (\beta_\emptyset - \beta_c)(1-\beta_\emptyset)(1-\gamma_c).$$

This can be verified by using

$$P_t^h + N_t^h + Z_t^h + F_t^h = 1$$

for each $t \geq 1$ and the transition matrix to establish it as an identity in P_t^h, N_t^h, Z_t^h and F_t^h. The explicit solution to the above difference equation is

$$X_t^h = \beta_\emptyset(B^h)^{t-1} + \frac{\beta_c(1-(B^h)^{t-1})}{1-B^h} + \frac{G^h((\Lambda^h)^{t-1} - (B^h)^{t-1})}{\Lambda^h - B^h}.$$

Then, from

$$W^h(\beta;\gamma) = \sum_{t=1}^{\infty} \delta^t X_t^h$$

we can complete the geometric sums to get

$$W^h(\beta;\gamma) = \frac{\delta}{1-\delta B^h}\left(\beta_\emptyset + \frac{\delta\beta_c}{1-\delta} + \frac{\delta G^h}{1-\delta\Lambda^h}\right).$$

This reduces to $V^h(\beta)$ of section 2 when there is no memory manipulation.

Symmetrically, defining

$$W^l(\beta;\gamma) = \sum_{t=1}^{\infty} \delta^t X_t^l$$

and deriving X_t^l in the same way as for X_t^h, we have

$$W^l(\beta;\gamma) = \frac{\delta}{1-\delta B^l}\left(\beta_\emptyset + \frac{\delta\beta_c}{1-\delta} + \frac{\delta G^l}{1-\delta\Lambda^l}\right),$$

where

$$\Lambda^l = (1-\beta_\emptyset)(1-\gamma_c) + \beta_\emptyset((1-q)(1-\gamma_+) + q(1-\gamma_-)),$$

and

$$G^l = (\beta_\emptyset - \beta_+)\beta_\emptyset(1-q)(1-\gamma_+) + (\beta_\emptyset - \beta_-)\beta_\emptyset q(1-\gamma_-) + (\beta_\emptyset - \beta_c)(1-\beta_\emptyset)(1-\gamma_c).$$

Finally, we can write

$$W(\beta;\gamma) = (2q-1)(\eta W^h(\beta;\gamma) - (1-\eta)W^l(\beta;\gamma)).$$

We have the following result:

PROPOSITION 4.1. *For all $\eta \in (\eta_*, \eta_1)$, $\max_\beta W(\beta;1) < W(\beta';\gamma)$ for some β' and $\gamma \neq 1$.*

By definition, $\max_\beta W(\beta;1)$ is the optimal value of the period 0 discounted sum of utilities when there is no memory manipulation. From the characterization of Proposition 2.1, this optimal value can be attained by using random and periodic experimentation, with $\beta_\emptyset = \beta_+ = 1$ and $(1-\beta_-)/(1-\delta(1-\beta_c)) = K(\eta)$. The claim of Proposition 4.1 is established by showing at any such optimal β with no manipulation, there exists $\gamma \neq 1$ such that $W(\beta;\gamma) > W(\beta;1)$.

The rough intuition behind Proposition 4.1 may be understood as follows. Without memory manipulation, there are effectively only three memory states, $+$, $-$ and c, because the initial memory state of $\emptyset$ exists only for the first period.[8] Unlike those for $+$, $-$ and c,

[8] By assumption, the decision maker does not recall calendar time but is able to distinguish the first period from the rest of decision nodes.

the experimental decision corresponding to $\emptyset$ is one time only. By the characterization of the optimal learning strategy in Proposition 2.1, $\beta_\emptyset$ is equal to 1 if the value of objective function under an optimal learning strategy is positive, 0 otherwise. In contrast, memory manipulation allows the decision maker to make the memory state $\emptyset$ a recurring state. This can help improve the decision maker's ex ante welfare because an additional memory state can be used to enrich the state space and allow the strategy to better respond to new information.

The above intuition can be made more precise by following the steps of the proof of Proposition 4.1. We first observe that with $\beta_\emptyset = \beta_+ = 1$, the decision maker attains the same ex ante payoff by setting $\gamma_+ = 0$ and $\gamma_- = 1$ as no manipulation (setting $\gamma_+ = \gamma_- = 1$). The path of decisions is identical in these two scenarios if the payoff from the risky arm in the first period is negative because $\gamma_- = 1$, while the same decisions are made following a positive payoff from the risky arm in the first period even though $\gamma_+ = 0$, as $\beta_\emptyset = \beta_+$. We ask if the decision maker can improve his ex ante payoff by reducing γ_- while maintaining $\gamma_+ = 0$. The key is to note that under $\gamma_+ = 0$ and $\gamma_- = 1$ the memory state of $\emptyset$ carries distinct information from the memory state of $+$: the state of $\emptyset$ occurs only after a string of positive payoffs from the risky arm, whereas the state of $+$ occurs only after getting at least one negative payoff in the past. The former suggests a more favorable belief about the risky arm and thus should lead to a greater probability of experimentation than the latter, but such distinction can not be made when there is no memory manipulation by the decision maker. With memory manipulation, this can be exploited by the decision maker by reducing γ_- to just below 1. Then, the decision maker has a positive probability of ignoring a negative payoff when the current memory state is $\emptyset$. For small reductions in γ_-, the benefit of increasing experimentation when the state is likely to be h outweighs the potential cost of repeatedly ignoring the unfavorable information of negative payoffs.

Proposition 4.1 is proved by changing γ while maintaining the same optimal β under no manipulation. This raises the question of whether the decision maker not only wants to make γ less 1 but also wishes to deviate from the optimal β with no manipulation. The answer is yes. To see this, for any β and γ such that $\beta_\emptyset = \beta_+ = 1$, we can write $W(\beta; \gamma)$

as

$$\frac{\delta(2q-1)\eta}{1+\delta(1-q)K}\left(\frac{1}{1-\delta}+\frac{\delta(1-q)(1-\gamma_-)K}{1-\delta\Lambda^h}\right)-\frac{\delta(2q-1)(1-\eta)}{1+\delta qK}\left(\frac{1}{1-\delta}+\frac{\delta q(1-\gamma_-)K}{1-\delta\Lambda^l}\right),$$

where

$$K=\frac{1-\beta_-}{1-\delta(1-\beta_c)},$$

and

$$\Lambda^h=q(1-\gamma_+)+(1-q)(1-\gamma_-);$$
$$\Lambda^l=(1-q)(1-\gamma_+)+q(1-\gamma_-).$$

Thus, as in section 2, β_- and β_c matter only through K. The derivative of $W(\beta;\gamma)$ with respect to γ_+ has the same sign as

$$-\frac{\eta(1-q)}{1+\delta(1-q)K}\frac{q}{(1-\delta\Lambda^h)^2}+\frac{(1-\eta)q}{1+\delta qK}\frac{(1-q)}{(1-\delta\Lambda^l)^2}.$$

It is straightforward to verify that the second derivative of $W(\beta;\gamma)$ with respect to γ_+ is strictly positive when the first derivative is zero. Similarly, the derivative of $W(\beta;\gamma)$ with respect to γ_- has the same sign as

$$-\frac{\eta(1-q)}{1+\delta(1-q)K}\frac{1-\delta q(1-\gamma_+)}{(1-\delta\Lambda^h)^2}+\frac{(1-\eta)q}{1+\delta qK}\frac{1-\delta(1-q)(1-\gamma_+)}{(1-\delta\Lambda^l)^2},$$

with a strictly negative sign for the second derivative when the first derivative is zero. Further, one can easily check that $\partial W/\partial\beta_+\geq 0$ implies that $\partial W/\partial\beta_+>0$. It follows that the optimal value for γ_+ is either 0 or 1, and $\gamma_-\geq\gamma_+$ at an optimum. The derivative of $W(\beta;\gamma)$ with respect to K has the same sign as

$$-\frac{\eta(1-q)}{(1+\delta(1-q)K)^2}A^h+\frac{(1-\eta)q}{(1+\delta qK)^2}A^l,$$

where

$$A^h=1-\frac{(1-\delta)(1-\gamma_-)}{1-\delta+\delta q\gamma_++\delta(1-q)\gamma_-};$$
$$A^l=1-\frac{(1-\delta)(1-\gamma_-)}{1-\delta+\delta(1-q)\gamma_++\delta q\gamma_-}.$$

If $\gamma_+<\gamma_-$, then $A^h<A^l$ and therefore $\partial W(\beta;\gamma)/\partial K>0$ at $K=K(\eta)$. We already know from Proposition 4.1 that for any $\eta\in(\eta_*,\eta_1)$ in the *random and periodic experiment*

region, the decision maker can improve his ex ante welfare by memory manipulation without changing the optimal learning strategy β under no manipulation. At any such optimal manipulation we must have $\gamma_+ < \gamma_-$, which then implies that the decision maker could further increase his ex ante payoff with changes in the learning strategy β by increasing K.[9]

5. Open Questions

This paper is a simple example of dynamic decisions with short term memories. We have looked at a two armed bandit problem, and the hope is that it is suggestive of the time inconsistency and memory manipulation issues we want to study in generational learning. Similarly, the short term memory constraint takes a simple form in our model. It will be worthwhile to pursue more general forms of such constraint, for example by allowing the decision maker to recall the past experience of more than a single period. In particular, we have shown that optimal learning strategies are necessarily time inconsistent if they are responsive to new information. Whether this is true with more general dynamic decision problems and more general short term memory constraints remain to be seen. Further, the memory manipulations considered in this paper are one of many ways available to the decision maker. Whether, and how, other kinds of manipulations can improve the ex ante welfare of the decision maker are interesting topics that we plan to pursue in future research. Finally, we have treated the issues of time inconsistency and memory manipulation separately. Is there link between these two issues? In particular, does manipulation make the optimal policy more likely to be time inconsistent?

[9] Given the interpretation of K as a measure of the frequency of playing the safe arm, an increase in K compensates the increase in the probability of experimentation that comes with memory manipulation (i.e. a decrease in γ_- to below 1).

Appendix

A.1. Proof of Proposition 2.1

PROOF. The derivatives of $V(\beta)$ with respect to β_-, β_c and β_+ are given by:

$$\frac{\partial V}{\partial \beta_-} = \delta^2(2q-1)\left(\beta_\emptyset + \frac{\delta\beta_c}{1-\delta}\right)\left(\frac{\eta(1-q)}{(1-\delta B^h)^2} - \frac{(1-\eta)q}{(1-\delta B^l)^2}\right),$$
$$\frac{\partial V}{\partial \beta_c} = \delta^2(2q-1)\left(\frac{(1-\delta)(B^h-\beta_\emptyset)-\delta\beta_c}{(1-\delta)(1-\delta B^h)^2} - \frac{(1-\delta)(B^l-\beta_\emptyset)-\delta\beta_c}{(1-\delta)(1-\delta B^l)^2}\right),$$
$$\frac{\partial V}{\partial \beta_+} = \delta^2(2q-1)\left(\beta_\emptyset + \frac{\delta\beta_c}{1-\delta}\right)\left(\frac{\eta q}{(1-\delta B^h)^2} - \frac{(1-\eta)(1-q)}{(1-\delta B^l)^2}\right),$$

Since $V(\beta)$ is linear in $\beta_\emptyset$, we have $\beta_\emptyset = 1$ if $V(\beta) > 0$ at any optimal β, and $\beta_\emptyset = 0$ otherwise. The 0 payoff can be implemented by setting $\beta_\emptyset = \beta_c = 0$, regardless of β_+ and β_-. We have:

$$V(\beta_\emptyset = \beta_c = 0) = 0.$$

For the remainder of the proof, we assume that $\beta_\emptyset = 1$.

It is straightforward to verify that

$$\frac{1-q}{q} \leq \frac{1-(q\beta_+ + (1-q)\beta_-)}{1-((1-q)\beta_+ + q\beta_-)} \leq \frac{q}{1-q},$$

with the first as an equality if and only if $\beta_+ = 1$, and the second as an equality if and only if $\beta_- = 1$. It follows that the signs of the derivatives of $V(\beta)$ with respect to β_-, β_c and β_+ are ordered: $\partial V/\partial\beta_- \geq 0$ implies that $\partial V/\partial\beta_c > 0$ if $\beta_+ < 1$, and the two have the same signs if $\beta_+ = 1$; while $\partial V/\partial\beta_c \geq 0$ implies that $\partial V/\partial\beta_+ > 0$ if $\beta_- < 1$, and the two have the same signs if $\beta_- = 1$. We distinguish the following three cases.

(1) If $\beta_+ = 0$, then $\partial V/\partial\beta_+ \leq 0$ at the optimum. We have $\partial V/\partial\beta_c, \partial V/\partial\beta_- < 0$, and therefore $\beta_c = \beta_- = 0$. In this case

$$V(\beta_\emptyset = 1, \beta_+ = \beta_- = \beta_c = 0) = \delta(2q-1)(2\eta-1).$$

(2) If β_+ is in the interior, then $\beta_c = \beta_- = 0$ as in case (1). We have:

$$\frac{\partial V}{\partial \beta_+} = \frac{\delta^2(2q-1)}{1-\delta}\left(\frac{\eta q}{(1-\delta q\beta_+)^2} - \frac{(1-\eta)(1-q)}{(1-\delta(1-q)\beta_+)^2}\right).$$

Thus, there can be at most one critical point at which $\partial V/\partial\beta_+ = 0$. Evaluating the second derivative at this point, we find that $\partial^2 V/\partial\beta_+^2$ has the same sign as

$$\frac{q}{1-\delta q\beta_+} - \frac{1-q}{1-\delta(1-q)\beta_+},$$

which is positive because $q > \frac{1}{2}$. It follows that an interior β_+ can not be optimal.

(3) If $\beta_+ = 1$, then $\partial V/\partial\beta_+ \geq 0$ at the optimum. This case allows for interior solutions in β_c and β_-. Since $\beta_+ = 1$, the signs of $\partial V/\partial\beta_c$ and $\partial V/\partial\beta_-$ are the same, and so both β_c and β_- can be interior at the same time. Indeed, with $\beta_\emptyset = \beta_+ = 1$, we can rewrite V as follows

$$V = \frac{\delta(2q-1)}{1-\delta}\left(\frac{\eta}{1+\delta(1-q)K} - \frac{1-\eta}{1+\delta qK}\right),$$

where

$$K = \frac{1-\beta_-}{1-\delta(1-\beta_c)}.$$

By definition, we have $0 \leq K \leq 1/(1-\delta)$. Since V depends on β only through K, we can take derivatives with respect to K and get the following first order condition:

$$-\frac{\eta(1-q)}{(1+\delta(1-q)K)^2} + \frac{(1-\eta)q}{(1+\delta qK)^2} = 0.$$

Define $K(\eta)$ as the point that satisfies the above first order condition. It is straightforward to verify that the second order condition is satisfied at $K = K(\eta)$. Thus, if $K(\eta) \in [0, 1/(1-\delta)]$, the maximal payoff with $\beta_\emptyset = \beta_+ = 1$ is reached when β_c and β_- satisfy

$$\frac{1-\beta_-}{1-\delta(1-\beta_c)} = K(\eta).$$

Using the definitions of $K(\eta)$, η_1 and η_*, we have $K(\eta) \geq 0$ if and only if $\eta \leq \eta_1$, while $K(\eta) \leq 1/(1-\delta)$ if and only if $\eta \geq \eta_*$. The maximal payoff with $\beta_\emptyset = \beta_+ = 1$ for $\eta \in [\eta_*, \eta_1]$ is thus given by

$$\begin{aligned} V(\beta_\emptyset = \beta_+ = 1, &(1-\beta_-)/(1-\delta(1-\beta_c)) = K(\eta)) \\ &= \frac{\delta(2q-1)}{1-\delta}\left(\frac{\eta}{1+\delta(1-q)K(\eta)} - \frac{1-\eta}{1+\delta qK(\eta)}\right). \end{aligned}$$

For all $\eta > \eta_1$, one can verify that $\partial V/\partial K < 0$ for all $\eta > \eta_1$, implying that $K = 0$ at the optimum and thus $\beta_- = 1$ (β_c is unrestricted). The maximal payoff with $\beta_\emptyset = \beta_+ = 1$ for $\eta \geq \eta_1$ is then given by

$$V(\beta_\emptyset = \beta_+ = \beta_- = 1) = \frac{\delta(2q-1)(2\eta-1)}{1-\delta}.$$

For all $\eta < \eta_*$, we have $\partial V/\partial K > 0$, implying that $K = 1/(1-\delta)$ at the optimum and thus $\beta_- = \beta_c = 0$. The maximal payoff with $\beta_\emptyset = \beta_+ = 1$ for $\eta \leq \eta_*$ is then given by

$$V(\beta_\emptyset = \beta_+ = 1, \beta_c = \beta_- = 0) = \delta(2q-1)\left(\frac{\eta}{1-\delta q} - \frac{1-\eta}{1-\delta(1-q)}\right).$$

Comparing the last scenario of case (3) with $\eta \leq \eta_*$ to case (1), we find that

$$V(\beta_\emptyset = \beta_+ = 1, \beta_c = \beta_- = 0) > V(\beta_\emptyset = 1, \beta_+ = \beta_- = \beta_c = 0),$$

whenever the latter is positive, which is when $\eta > \frac{1}{2}$. Thus, case (1) can not occur at the optimum. Finally, by the definition of η_0, we have

$$V(\beta_\emptyset = \beta_+ = 1, \beta_c = \beta_- = 0) > 0$$

if and only if $\eta > \eta_0$. The characterization of optimal strategies in Proposition 2.1 then follows immediately. *Q.E.D.*

A.2. Proof of Proposition 3.1

PROOF. We check time consistency for each of the four cases in Proposition 2.1 separately.

Case (i). The only memory state that happens with positive probability after the initial period is c. To calculate $\Pr[h|c]$, we need to make assumptions on the values of β_+ and β_-, which are unrestricted in this case. We choose $\beta_+ = \beta_- = 0$, implying that

$$\Pr[h|c] = \frac{\eta}{\eta + (1-\eta)} = \eta.$$

Since the updated belief stays at η, the optimal strategy is time consistent in this case.

Case (ii). Here we have

$$\Pr[h|-] = \frac{\eta(1-q)q}{\eta(1-q)q + (1-\eta)q(1-q)} = \eta,$$

which is greater than η_0 by assumption. Thus, the optimal strategy is time inconsistent in this case.

Case (iii). Note that the optimal $\beta_\emptyset$ is either 0 or 1, except when $\eta = \eta_0$, in which case an interior $\beta_\emptyset$ can be optimal because the decision maker is indifferent. Since either β_c or β_-, or both, must be interior in any optimal strategy in case (iii), the only candidate β for time consistent optimal strategy requires

$$\Pr[h|c] = \Pr[h|-] = \eta_0.$$

One can easily verify that

$$\Pr[h|c] \geq \Pr[h|-],$$

with equality if and only if $\beta_- = 0$. Thus, we must have $\beta_- = 0$ for an optimal policy to be time consistent. Then, it follows from $\Pr[h|c] = \Pr[h|-] = \eta_0$ that

$$\frac{q(1-\eta)}{(1-q)\eta} = \frac{1-\delta(1-q)}{1-\delta q}\frac{\beta_c + q}{\beta_c + 1 - q},$$

while it follows from the definition of $K(\eta)$ that

$$\frac{q(1-\eta)}{(1-q)\eta} = \left(\frac{1-\delta(1-\beta_c)+\delta q}{1-\delta(1-\beta_c)+\delta(1-q)}\right)^2.$$

However, the above two equalities can not be satisfied simultaneously, as

$$\frac{1-\delta(1-q)}{1-\delta q} > \frac{1-\delta(1-\beta_c)+\delta q}{1-\delta(1-\beta_c)+\delta(1-q)},$$

and

$$\frac{\beta_c + q}{\beta_c + 1 - q} > \frac{1-\delta(1-\beta_c)+\delta q}{1-\delta(1-\beta_c)+\delta(1-q)}.$$

Therefore, no optimal strategy is time consistent in this case.

Case (iv). We have

$$\Pr[h|+] = \frac{\eta q}{\eta q + (1-\eta)(1-q)} > \eta,$$

which is greater than η_0. Further,

$$\Pr[h|-] = \frac{\eta(1-q)}{\eta(1-q)+(1-\eta)q}.$$

Using the definitions of η_0 and η_1, we can verify that $\Pr[h|-] > \eta_0$ for all $\eta > \eta_1$ because $\eta_0 < \frac{1}{2}$. Thus, the optimal strategy is time consistent in this case. *Q.E.D.*

A.3. Proof of Proposition 4.1

PROOF. For any $\eta \in (\eta_*, \eta_1)$, let β be such that $\beta_\emptyset = \beta_+ = 1$, $\beta_- = 0$ and β_c satisfies

$$\frac{1}{1-\delta(1-\beta_c)} = K(\eta).$$

Then, we can write the difference between $W(\beta;\gamma)$ for any γ and $W(\beta;\gamma_+ = \gamma_- = 1)$ as

$$\delta^2(2q-1)\left(\frac{(1-q)(1-\gamma_-)\eta K}{(1+\delta(1-q)K)(1-\delta\Lambda^h)} - \frac{q(1-\gamma_-)(1-\eta)K}{(1+\delta qK)(1-\delta\Lambda^l)}\right),$$

where

$$\Lambda^h = q(1-\gamma_+) + (1-q)(1-\gamma_-);$$
$$\Lambda^l = (1-q)(1-\gamma_+) + q(1-\gamma_-).$$

Note that

$$W(\beta;\gamma_+ = 0, \gamma_- = 1) - W(\beta;\gamma_+ = \gamma_- = 1) = 0.$$

The derivative of $W(\beta;\gamma) - W(\beta;\gamma_+ = \gamma_- = 1)$ with respect to γ_-, evaluated at $\gamma_- = 1$, has the same sign as

$$-\frac{(1-q)\eta K}{(1+\delta(1-q)K)(1-\delta q(1-\gamma_+)} + \frac{q(1-\eta)K}{(1+\delta qK)(1-\delta(1-q)(1-\gamma_+))}.$$

Since

$$\frac{(1-q)\eta}{(1+\delta(1-q)K(\eta))^2} = \frac{q(1-\eta)}{(1+\delta qK(\eta))^2},$$

the sign of the derivative evaluated at $K = K(\eta)$ is the same as

$$\frac{\delta(2q-1)K(\eta)}{1-\delta(1-\gamma_+)}\left(K(\eta) - \frac{1-\gamma_+}{1-\delta(1-\gamma_+)}\right).$$

For $\gamma_+ = 0$, we have $K(\eta) > 0$ and $K(\eta) < 1/(1-\delta)$ for all $\eta \in (\eta_*, \eta_1)$. Thus, the derivative of $W(\beta;\gamma) - W(\beta;\gamma_+ = \gamma_- = 1)$ with respect to γ_-, evaluated at $\gamma_- = 1$, $\gamma_+ = 0$ and $K = K(\eta)$ is strictly positive for all $\eta \in (\eta_*, \eta_1)$. The proposition follows immediately. *Q.E.D.*

References

Börgers, T. and A. Morales, 2004, "Complexity constraints in two-armed bandit problems: an example," University College London working paper.

Gittins, J.C., 1989, *Multi-armed Bandit Allocation Indices*, New York: John Wiley & Sons.

Kalai, E. and E. Solan, 2003, "Randomization and simplification in dynamic decision-making," *Journal of Economic Theory* 111, pp 251–264.

Kuhn, H.W., 1953, "Extensive games and the problem of information," in *Contributions to the Theory of Games III*, pp 79–96, Princeton, NJ: Princeton University Press.

Meyer, M., 1991, "Learning from coarse information: biased contests and career profiles," *Review of Economic Studies* 58, pp 15–41.

Lipman, B., 1995, "Information processing and bounded rationality: a survey," *Canadian Journal of Economics* 28, pp 42–67.

Piccione, M. and A. Rubinstein, 1997, "On the interpretation of decision problems with imperfect recall," *Games and Economic Behavior* 20, pp 2–35.

Rubinstein, A., 1986, "Finite automata play the repeated prisoners' dilemma," *Journal of Economic Theory* 39, pp 83–96.

Wilson, A., 2004, "Bounded memory and biases in information processing," Princeton University working paper.