A Dynamic Behavioral Model of Sequential Decisions[*]

by

Dale O. Stahl

Emeritus Professor
Department of Economics
University of Texas at Austin

April 18, 2020

ABSTRACT

Most important economic decision problems are sequential, and thus naturally represented as Markov Decision Problems (MDP). After reviewing the theory of MDPs, the applicability of MDPs to real-life sequential decisions appears impractical. The central question addressed in this essay is how ordinary humans behave in the real-life sequential decision problems they face. A formal behavioral approach is presented, and it also appears impractical for all but toy MDPs. After engaging in introspection, a key insight is the enormous extent to which information and reinforcement provided by parents, teachers and others shapes our behavior in real-life. With these insights, an integrated framework of dynamic behavior emerges in which genetic evolution, short-term reinforcement learning and long-term acquisition of knowledge via institutions are seen as important aspects.

[*] This essay was made possible by a Faculty Research leave funded by the University of Texas for the fall semester of 2017, which was spent at the University of California at Santa Barbara. The author is indebted to the many members of the faculty there for their hospitality and wide-ranging discussions on human behavior, especially to Gary Charness, Ignacio Esponda, and Ryan Opera.

1. Introduction.

Most important economic decision problems are sequential (e.g. intertemporal pricing, advertising, trading, R&D, capital investment, quality signaling, bargaining, etc.), and thus naturally represented as Markov Decision Problems (MDP). Moreover, since life has no definite terminal date, every decision is imbedded in a much larger infinite-horizon MDP with an uncertain terminal date. Unfortunately, infinite-horizon MDPs are generally intractable. The strategy sets are infinite and often uncountable, and (except for textbook MDPs) well beyond the conceptual and computational abilities of humans. Even if someone of genius stature were to solve such a MDP, the solution would have little bearing on how ordinary humans would behave. Solving a finite snippet of an infinite MDP is much easier, but that partial solution is not necessarily better than heuristics for the full MDP considering the time and resource constraints on the decision-making process. The central question addressed in this essay is how ordinary humans behave in the real-life sequential decision problems they face.

My approach is first to present in Section 2 a formal model for a general MDP to establish terms and notation. Given this model, I will review the extant results stemming from Bellman (1954), and then critique the practicality of finding a solution. I will argue that it is unreasonable to think that humans solve real-life MDPs. Section 3 presents a behavioral framework for addressing how an ordinary human might approach the real-life MDP he/she faces. This behavioral framework entails behavioral rules and reinforcement learning instead of optimization. I will argue that for all but toy MDPs, one-period and two-period MDPs, reinforcement learning based on the individual's own experience and/or model-based simulations is not likely to lead to behavior that is close to being optimal.

With such a negative conclusion, it seems appropriate to consider approaches other than the MDP framework. Accordingly, in Section 4, I turn to introspection about my own sequential decision making for clues. A key insight is the enormous extent to which information and reinforcement provided by parents, teachers and others shapes our behavior in real-life. Naturally, Darwinian evolution plays a significant role, but the availability and use of advice from others is the key to our success.

With these insights, an integrated framework of dynamic behavior emerges in which genetic evolution, short-term reinforcement learning and long-term acquisition of knowledge are

seen as aspects of an underlying dynamic. This framework is developed and elaborated in Section 5. While this essay focuses on the individual, it may provide a fruitful framework for thinking about organizational behavior. This and other extensions are discussed in Section 6.

2. The Markov Decision Problem Framework.

For simplicity, assume decisions occur at discrete times, indexed by $t \in \{1, 2, \ldots\}$. Let ω_t denote the state (or information available) at time t. Let Ω denote the set of all possible (information) states for all t. Part of the information available at time t is the set of feasible actions, $\mathcal{A}(\omega_t)$, where $\mathcal{A}(\,)$ denotes the projection onto the action subspace of Ω. Let $A \equiv \mathcal{A}(\Omega)$ denote the subspace of all possible actions for all t. For expositional convenience I will assume A and Ω are finite.[1]

Let $r(a, \omega)$ denote the immediate reward from action $a \in \mathcal{A}(\omega)$ in state ω. We assume this real-valued reward function is bounded and continuous. Let $f(\omega' \mid a, \omega)$ denote the probability of transitioning to state ω' in the next period conditional on current action and state (a, ω). In other words, f() is a dynamic *model of the world*.

A *decision rule* is a mapping $\alpha: \Omega \to A$. Since a decision rule specifies an action for every possible state, it can also be considered a plan of action. It is assumed that the Decision Maker (DM) chooses a decision rule α that maximizes the expected discounted present value of immediate rewards generated by α given initial state ω_1:

$$V(\alpha, \omega_1) \equiv r[\alpha(\omega_1), \omega_1] + \sum_{t>1} \delta^{t-1} \sum_{\omega_t \in \Omega} r[\alpha(\omega_t), \omega_t] f[\omega_t \mid \alpha(\omega_1), \omega_1], \qquad (1)$$

where $\delta \in [0, 1)$ is the discount factor[2], and where for $t \geq 2$,

$$f(\omega_t \mid \alpha, \omega_1) \equiv \sum_{\omega_1 \in \Omega} \ldots \sum_{\omega_t \in \Omega} \left(\prod_{s=2}^{t} f[\omega_s \mid \alpha(\omega_{s-1}), \omega_{s-1}] \right) \qquad (2)$$

[1] If not, then you may interpret A and Ω as finite approximations of the non-finite spaces. Since humans have a finite number of neurons, the representation of A and Ω in a human brain must be finite.

[2] When the time periods are heterogeneous, we would replace β by $e^{-r\Delta t}$, where Δt is the period duration.

is the probability of transitioning from ω_1 to ω_t. Thus, a Markov Decision Problem (MDP) is defined by a triplet (r, f, δ) and eqs(1-2).

It is insightful to rewrite eq(1) as

$$V(\alpha, \omega_1) = r[\alpha(\omega_1), \omega_1] + \delta \sum_{\omega_2 \in \Omega} V(\alpha, \omega_2) f[\omega_2 \mid \alpha(\omega_1), \omega_1] . \quad (3)$$

In other words, the discounted present value of rewards can be expressed as the immediate reward plus the expected discounted value of future rewards given the decision rule and the probability distribution of the state in period 2 conditional on the current action and state.

Define

$$V^*(\omega_1) \equiv max_{\alpha} V(\alpha, \omega_1) \quad (4)$$

Then from eq(3-4):

$$V^*(\omega_1) = max_{a} [r(a, \omega_1) + \delta \sum_{\omega_2 \in \Omega} V^*(\omega_2) f(\omega_2 \mid a, \omega_1) d\omega_2] . \quad (5)$$

This is the well-known Bellman equation for dynamic programming [Bellman, 1954; Bertsekas, 2000]. It is also well-known that, provided r() is bounded, and A is compact, there is a unique function V*( ) satisfying eq(5). Associated with V*( ) is an *optimal decision rule* α*( ) such that $V[\alpha^*(\omega), \omega] = V^*(\omega)$ for all $\omega \in \Omega$. Although V*( ) is unique, there can be multiple optimal decision rules, so a complete solution to the MDP requires the addition of a selection rule for the cases of multiple optimal decision rules. Also, when there are multiple optimal solutions, then any probabilistic rule whose support is confined to $arg\ max_{\alpha} V(\alpha, \omega_1)$ is also an optimal solution.

Moreover substituting $v_i()$ for V*() on the r.h.s. of eq(5), and $v_{i+1}()$ for V*() on the l.h.s., eq(5) becomes a contraction mapping; therefore the sequence $\{ v_i(), i = 0, 1, ...\}$ will converge uniformly to V*(). In practice, one typically imposes a finite grid on Ω,[3] which introduces computational error into the result: i.e. $|v_i() - V^*()| > 0$ for all finite i. For large dimensional

[3] Sparce grid methods can improve computational efficiency, as can function approximation methods.

MDPs[4], this error can be substantial. Machine learning algorithms have been developed; in particular "deep learning" uses a neural network on the front end to essentially reduce the dimensionality of the problem and then uses supervised learning to tune the network [e.g. Arulkumaran, et al. (2017)]. For finite horizon MDPs, the standard solution method is *backward induction*, but again this method is not implementable for large dimensional MDPs.[5] Thus, we are left with the original question: how do ordinary humans behave when facing a MDP?

3. A Behavioral Extension.

Instead of solving for an optimal decision rule, suppose humans use a behavioral rule which is a mapping from the available information into a probability measure on the available actions. Let $\rho: \Omega \rightarrow \Delta(A)$ denote a *behavioral rule*, such that $\rho(a \mid \omega)$ is the probability of action $a \in \mathcal{A}(\omega)$.

It is convenient to work with the log of $\rho(a \mid \omega)$:

$$u(a, \omega) \equiv ln[\rho(a \mid \omega)], \tag{6}$$

which is not confined to [0, 1]. I will call $u(a, \omega)$ the *attraction* of action a in state ω. Clearly, the probability of action a in state ω can be expressed as the logistic function of $u(\ , \omega)$;

$$\rho(a \mid \omega) = exp[u(a, \omega)]/\sum_{b \in A(\omega)} exp[u(b, \omega)]. \tag{7}$$

The expected discounted present value of following rule ρ in state ω_1 is:

$$V(\rho, \omega_1) \equiv r(\rho, \omega_1) + \sum_{t>1} \delta^{t-1} \sum_{\omega_t \in \Omega} r(\rho, \omega_t) f(\omega_t \mid \rho, \omega_1), \tag{8}$$

where $r(\rho, \omega) \equiv \sum_{a \in A(\omega)} r(a, \omega)\, \rho(a \mid \omega)$,

$$f(\omega_t \mid \rho, \omega_{t-1}) \equiv \sum_{a \in A(\omega)} f(\omega_t \mid a, \omega_{t-1})\, \rho(a \mid \omega_{t-1}) \text{ for } t > 1, \text{ and}$$

[4] The dimension of a finite MDP is the dimension of Ω.

[5] Chess, for example, is a finite perfect information game solvable in principle by backward induction, and yet our most powerful computers cannot solve for the optimal strategy.

$$f(\omega_t \mid \rho, \omega_1) \equiv \sum_{\omega_{t-1} \in \Omega} \cdots \sum_{\omega_2 \in \Omega} \left(\prod_{s=2}^{t} f(\omega_s \mid \rho, \omega_{s-1}) \right) . \quad (9)$$

The task of computing $V(\rho, \omega_1)$ is at least as difficult as solving the Bellman eq(3), and therefore it is unreasonable to assume that humans do this. Moreover, in many real life MDPs, there may be uncertainty about the reward function r and the transition function f, further complicating the problem. Hence, we turn to methods for learning instead of optimization.

3.1 Reinforcement Learning.

Reinforcement learning applied to behavioral rules is an iterative process.[6] Let $u_t(a, \omega)$ denote the attraction function for period t. Then, reinforcement learning dictates that

$$u_{t+1}(a, \omega) = (1-\beta)u_t(a, \omega) + \beta\, g_t(a, \omega), \quad (10)$$

where $\beta \in (0, 1)$ and $g_t(a, \omega)$ is the *reinforcement function.* Obviously the key ingredient in this process is the reinforcement function. For example, if it is equal to a fixed function $g(a, \omega)$, then $u_t(a, \omega)$ will converge to $g(a, \omega)$ as $t \to \infty$ provided (a, ω) occurs infinitely often.

The appropriate reward function depends on the particular MDP. It is convenient to focus on T-period problems, where T is a finite number. Adopting the convention that the current period is t = 1, for an MDP to be a T-period problem, it is sufficient that T is the smallest integer such that $f(\ \mid a_T, \omega_T)$ is the same for all $a_T \in \mathcal{A}(\omega_T)$ and all $\omega_T \in \Omega$. In other words, the future beyond period T is not affected by events from period 1 to period T. Let Ω^T denote the set of states which correspond to T-period problems; that is: if $\omega_1 \in \Omega^T$, then the DM faces a T-period problem now when t = 1, and next period the DM will face a T-1 period problem, etc.

For one-period problems, $\omega_1 \in \Omega^1$, clearly $r(a, \omega_1)$ is the most reasonable candidate[7] for the reinforcement function because $u_t(a, \omega_1)$ will converge to $r(a, \omega_1)$ provided (a, ω_1) occurs infinitely often, and the behavioral rule will converge to $\rho(a \mid \omega_1) = exp[r(a, \omega_1)]/\sum_{b \in A(\omega_1)}$

[6] The origins date back to Thorndike (1898), Bush and Mosteller (1955), Luce (1959) and Taylor and Jonker (1978). Most machine learning algorithms use reinforcement learning.

[7] Any monotonic transformation of $r(a, \omega)$ would also work.

exp[r(b, ω_1)]. Thus, reinforcement learning will lead to the optimal action (in state ω_1) with high probability and the worst action with low probability.

For T > 1, the appropriate reinforcement function is not straightforward. Consider two-period problems: i.e. $\omega_1 \in \Omega^2$. The reward in the next period will depend on the state in the next period (ω_2) and the action the DM takes (a_2) in that state. Therefore, the function r(a, ω_1) will be inadequate as the reinforcement function since it does not take account of my plan for period 2.

To accommodate the need for a plan, let $\hat{\rho}$ denote the behavioral rule the DM expects to follow in period 2. Then, the expected discounted present value of taking action a_1 in state ω_1 and following rule $\hat{\rho}$ in whatever state the DM reaches in the next period is

$$U(a_1, \omega_1; \hat{\rho}) \equiv r(a_1, \omega_1) + \delta \sum_{\omega_2 \in \Omega^1} r(\hat{\rho}, \omega_2) f(\omega_2 \mid a_1, \omega_1) . \qquad (11)$$

Notice that if $\omega_2 \notin \Omega^1$, then $f(\omega_2 \mid a, \omega_1) = 0$, so the summation in eq(11) is confined to states in Ω^1. That is, if in period 1, the DM believes he/she faces a two-period problem, then the DM also believes that in period 2 he/she will face a one-period problem.

One might suggest using $U(a_1, \omega_1; \hat{\rho})$ as the reinforcement function, but this would not necessarily lead to behavior anywhere near optimal. The reason is that the anticipated second period behavior $\hat{\rho}$ is arbitrary and thus unlikely to be anywhere near optimal. For $U(a_1, \omega_1; \hat{\rho})$ to be desirable as the reinforcement function, we need $\hat{\rho}$ to be near optimal. The obvious way to achieve this is to use reinforcement learning for the one-period problem that constitutes the second period of the two-period MDP. After each experience with $\omega_2 \in \Omega^1$, the DM updates his/her expectation $\hat{\rho}$ about what he/she will do in period 2, so in finite time, $\hat{\rho}$ will be essentially the same as the optimal behavioral rule $\rho(a \mid \omega_2)$. With this result, it is now reasonable to use $U(a, \omega_1; \hat{\rho})$ as the reinforcement function for period 1. Reinforcement learning will then cause $u_t(a, \omega_1)$ to converge to $U(a, \omega_1; \hat{\rho})$, and hence the best action in state ω_1 and contingency plan for period 2 will be taken with high probability and the worst with low probability. Note, however, that for near convergence, we need a sequence twice as long as for one-period problems, because we need to solve two one-period problems, one after the other, each requiring a convergent sequence.

One can easily extend this backward induction method to any T-period problem. However, the time for near convergence and the computational burden can be beyond the capability of a human brain. To illustrate, let K be the number of actions, and let M be the number of states. Then a complete deterministic contingency plan for a T-period problem is a (MT+1)-tuple, in which each element can have one of K values; therefore, there are K^{MT+1} different plans. For instance, if K = 10 and M = 5, and T = 2, there are 10^{11} distinct plans: (α_1, α_2). As we increase the number of periods (T), the number of distinct plans grows as $(K^M)^T$ - hardly a task a human can solve.[8] The backward-induction method illustrated involves only K^M one-period plans, but T convergent sequences, one after the other. Real-life time constraints are not likely to allow anything close to the required time. While a computer may be able to solve simple problems in finite computer time, humans do not have the luxury of stopping real time until a convergence threshold is met. Thus, evaluating complete contingency plans is practical only for toy MDPs with a very short horizon.

3.2 A Model-Based Simulation Approach.

Recall that we define Ω^T to be the states that correspond to the beginning of a T-period MDP. This setup allows us to express a model f as a vector of models (f_2, f_3, ...,), where f_T is a mapping from $A(\Omega^T)\times\Omega^T$ to a probability distribution on Ω^{T-1}. In other words, f_T is the component of a *model of the world* that applies to the first period of a T-period MDP and gives the probability distribution over the second-period of that T-period MDP (which is the first period of a (T-1) period problem.

The first component of the simulation approach is a partial adjustment process for learning the immediate reward function. Specifically, let $\hat{r}_t()$ denote expected reward function *prior to* period t. Given the state in period t is ω_t and action a_t is taken, the anticipated reward function is updated as follows:

[8] Recent advances in machine learning have been able to use "deep reinforcement learning" to outperform humans in simple video games such as Atari's pinball, pong, and space invaders by training a neural net to use the pixel-level video display of the game to drastically reduce the dimensionality (K^{TM}) of the problem (Mnih, 2015). In March 2016, DeepMind's AlphaGo defeated the world champion Go player. These accomplishments of machine learning leave no doubt that humans are constrained by the computational limitations of our brains.

$$\hat{r}_{t+1}(a_t, \omega_t) = \theta \hat{r}_t(a_t, \omega_t) + (1-\theta) r(a_t, \omega_t). \quad (12)$$

where $\theta \in (0, 1)$ and is the "inertia" of the updating process. The expected reward function $\hat{r}_{t+1}()$ for other values of (a, ω) are not updated.[9] Using this process, only $\hat{r}_{t+1}()$ need be carried forward in the state variable ω_{t+1}. The initial condition $\hat{r}_t()$ is part of ω_1.[10] Convergence of $\hat{r}_t(a, \omega)$ to the $r(a, \omega)$ will happen if (and only if) all $a \in \mathcal{A}(\omega)$ and $\omega \in \Omega$ occur infinitely often[11]. Let us assume that a human brain can carry out this updating/learning process. Henceforth, I will use $\hat{r}(a, \omega)$ in lieu of $\hat{r}_t(a, \omega)$ to simplify the notation.

The second component of the simulation approach is an algorithm to estimate the value function for T-period MDPs. Obviously, for one-period problems, simulations are not required so the algorithm uses simple reinforcement learning: i.e. $\hat{r}(a_1, \omega)$ is the reinforcement function. Let

$$\rho_1(a_1 \mid \omega) \equiv exp[\hat{r}(a_1, \omega)] / \sum_{b \in A(\omega)} exp[\hat{r}(b, \omega)]$$

denote the resulting logistic behavioral rule. [Note that the subscript "1" on ρ_1 denotes that it is a rule for a *one*-period problem; i.e. $\omega \in \Omega^1$.]

For T >1, assume that a human brain can run simulations of the MDP based on a model of the world. Specifically, given model f_2 (for a two-period problem) the DM's brain can simulate random draws ω_{2j} from $f_2(\mid a_1, \omega_1)$ and random draws a_{2j} from $\rho_1(\mid \omega_{2j})$, and compute expected rewards $\hat{r}(a_{2j}, \omega_{2j})$, j = 1, ..., J.[12] Let $y_1(a_1, \omega_1; \rho_1, f_2)$ denote the average of these simulated rewards: i.e.

---

[9] This direct dynamic could be extended to allow reinforcement of similar action-state pairs.

[10] It is reasonable to assume that since any action $a \in \mathcal{A}(\omega)$ that results in my immediate death cannot occur more than once, the initial expected reward for such actions is below all other possible realizations $\hat{r}_t$. Equivalently, my model of the world entails setting the initial expected values in this manner, since not doing so would be disadvantageous to my reproductive success.

[11] This process can be generalized to the case in which rewards are generated by a stationary stochastic process with mean $r(a, \omega)$.

[12] Note that since the second period of a two-period MDP is the first period of a one-period MDP, $\omega_{2j} \in \Omega^1$, the behavior rule $\rho_1(\mid \omega_{2j})$ is the one derived for one-period problems.

$$y_1(a_1, \omega_1; \rho_1, f_2) \equiv (1/J) \sum_j \hat{r}(a_{2j}, \omega_{2j}).^{13}$$

Then the discounted present value of the simulated rewards over the two periods is

$$EU_2(a_1, \omega_1; \rho_1, f_2) \equiv \hat{r}(a_1, \omega_1) + \delta\, y_1(a_1, \omega_1; \rho_1, f_2). \tag{13}$$

Let $\rho_2(\ |\ \omega_1, f_2)$ denote the resulting logistic behavioral rule according to eq(7) when substituting $EU_2(a_1, \omega_1; \rho_1, f_2)$ for $u_t(a_1, \omega_1)$; this is the behavioral rule for two-period MDPs. While many simulation runs are required for $y_1(a_1, \omega_1; \rho_1, f_2)$ to converge, presumably a human brain can run thousands of simulations in the real time of one period, whereas reinforcement learning is restricted to one update per realization of the specific two-period MDP.

For a three-period MDP, given model f_3, assume a DM's brain can simulate random draws ω_{2j} from $f_3(\ |\ a_1, \omega_1)$ and random draws a_{2j} from $\rho_2(\ |\ \omega_{2j}, f_3)$. If ω_{2j} has been experienced before, then $EU_2(a_{2j}, \omega_{2j}; \rho_1, f_2)$ has been simulated, so it can be used as the expected discounted present value of the simulated rewards over periods 2 and 3. However, if ω_{2j} has not been experienced before, then a sub-simulation must be conducted to compute the expected discounted present value of the simulated rewards over periods 2 and 3. Specifically, for *each* top-level simulated (a_{2j}, ω_{2j}), the DM's brain would need to simulate $EU_2(a_{2j}, \omega_{2j}; \rho_1, f_2)$ by making random draws ω_{3j} from $f_2(\ |\ a_{2j}, \omega_{2j})$ and random draws a_{3j} from $\rho_1(\ |\ \omega_{3j}, f_2)$, and compute the average of $\hat{r}(a_{2j}, \omega_{2j}) + \delta\, y_1(a_{2j}, \omega_{2,}; \rho_1, f_2)$ over these simulations, and then return to the top-level simulation. In any case, given these simulations, the expected discounted present value of the simulated rewards over all three periods is

$$EU_3(a_1, \omega_1; \rho_2, f_3) \equiv \hat{r}(a_1, \omega_1) + \delta\, EU_2(a_1, \omega_1; \rho_2, f_3). \tag{13}$$

Let $\rho_3(\ |\ \omega_1, f_3)$ denote the resulting logistic behavioral rule according to eq(7) when substituting $EU_3(a_1, \omega_1; \rho_2, f_3)$ for $u_t(a_1, \omega_1)$; this is the behavioral rule for three-period MDPs.

However, the computation burden of simulating $EU_3(a_1, \omega_1; \rho_2, f_3)$ is likely to be

[13] One could easily specify a process whereby each time ω_1 occurs, additional simulations are run and merged into $y_1(a_1, \omega_1; \rho_1, f_2)$, so $y_1()$ should be so interpreted; similarly for $y_2(a_1, \omega_1; \rho_2, f_3)$ below.

enormously larger than for two-period problems due to the likelihood that for some ω_{2j}, $EU_2(a_{2j}, \omega_{2j}; \rho_1, f_2)$ has not been previously simulated. Roughly, if N simulations are needed for a two-period MDP, then N^2 simulations are needed for a three-period MDP, and N^{T-1} simulations are needed for a T-period simulation.

While this simulation procedure can be applied in principle for periods 3, ... , T, a human brain will run out of the capacity to run a sufficient number of simulations for some $T \geq 3$, so the performance will surely deteriorate; i.e. $\rho_T(\mid \omega_1, f_T)$ will be far from optimal. Note that in period T of a T-period MDP, the DM faces a one-period problem, so no simulations are needed. Thus, the behavior in the last period of a multi-period MDP is likely to be closer to optimal than in earlier periods, especially period 1. On the bright side, the greater randomness in the early periods ensures a greater sampling of the later states and therefore smaller errors in the last period.

The third component of the model-based simulation approach is *model selection*. Up to this point, I have described how models can be used to produce behavioral rules for T-period MDPs, but I have not addressed the question of which model to use. No human has a single model of the world - not even theoretical physicists. Instead, we have many specialized models for different situations. Even if one believes that in principle biology can be reduced to particle physics, the function that maps particle physics to biological phenomena is beyond all conceivable computation ability. In lieu of a single theory of everything, we have specific models like biological models that relate biological phenomena to biological phenomena and some models that relate chemical phenomena to biological phenomena, etc.

An advantage of having different models is that a model can reduce the dimensionality of the problem. Specifically, a model could have as its domain a coarsening of the action and state spaces. A behavioral rule associated with a coarse model would be a mapping from a coarse state space to a probability distribution over a coarse action space. To extend this rule to the fine spaces, (i) any state belonging to the coarse set of states would produce the same probability distribution over the coarse action space, and (ii) whenever the model dictates that one of the coarse actions is to be taken, any action belonging to that coarse action would be taken with a pre-determined probability distribution over the elements of that coarse action space (perhaps uniform or concentrated about a representative action). With this two-step process in mind, I

shall use "model" to mean the model-specific action space and state space and the associated transition function.

We have many models that could be applied to any given sequential decision problem. This fact raises two pertinent questions: (i) what model do we use for specific decision problems, and (ii) how do we acquire, evaluate and improve these models? To be precise, let $\mathcal{F}$ denote the set of models at the DM's disposal (a space of models). Assume that the DM selects a model (and its behavioral rule) based on its performance.

Given model f_T, after L instances of using f_T and the behavioral rule, $\rho_T(\ |\ \omega, f_T)$, generated by the process described above, the average realized discounted present value of following f_T is defined as:

$$AU(f_T) \equiv (1/L) \sum_{i\geq 1}\sum_{t\geq 1} \delta^{t-1}\, r(a_{ti}, \omega_{ti} \mid f_T), \tag{15}$$

where $r(a_{ti}, \omega_{ti} \mid f_T)$ is the ith realization of the reward in the tth period of a T-period MDP using model f_T. $AU(f_T)$ is the natural measure of performance of model f_T, and thus it should have a major influence on whether to use model f_T, or some other model $f'_T \in \mathcal{F}$. Accordingly, it is reasonable to specify the probability of using model f_T as a logistic function of AU( ):

$$\psi(f_T) = exp[AU(f_T)] / \sum_{f'_T \in \mathcal{F}} exp[AU(f'_T)]. \tag{16}$$

Notice that for $T \geq 3$, $AU(f_T)$ entails the models $f_2, \ldots, f_{T-1}$, but only $AU(f_T)$ affects the probability of using f_T. $\psi(f_T)$ is specified this way because f_k ($k < T$) undergoes its own selection via $\psi(f_k)$, and we should not use data multiple times as if each use is an independent sample. For instance, $\psi(f_2)$ uses data from two-period problems. If we used this same data again to influence the selection of f_2 when considering three-period problems, we will have used it twice for the same purpose leading to a false sense of confidence. Thus, the selection of $f = (f_2, f_3, ...)$ happens component by component.

If $\psi(f_T)$ becomes negligible, then f_T could be deleted from $\mathcal{F}$. If the DM becomes aware of a new model f'_T for a T-period MDP, then he/she can add that new model to $\mathcal{F}$, simulate its

expected reward $EU_T(\ ;\rho_1, f'_T)$ and use that simulated reward as an initial value for $AU(f'_T)$ and $\psi(f'_T)$. Thus, the space of models $\mathcal{F}$ can change over time.

The expected behavior in state $\omega \in \Omega^T$ is

$$\tilde{\rho}_T(\ |\ \omega) \equiv \sum_{f_T \in \mathcal{F}} \psi(f_T)\, \rho_T(\ |\ \omega). \tag{17}$$

Obviously, the probability of using model f_T and behavioral rule $\rho_T(\ |\ \omega, f_T)$ will evolve as experience is accumulated. However, the change in $\psi(f_T)$ could be very slow since it requires experience with multiple models and T-period MDPs. On the other hand, as mentioned above, for two-period MDPs, $\psi(f_2)$ would adjust much faster than $\psi(f_T)$ for $T > 2$.

This model-based simulation approach is one theory of how humans behave and adjust their behavior with experience. It is not claimed that this process will converge to optimal behavior (clearly not in one human lifetime), only that it is responsive to rewards and has the potential to improve outcomes over time.

3.3 A Value-Function Approach.

Another approach that could be pursued is simply to suppose that the DM has an expectation over the expected value of future rewards given action $a \in \mathcal{A}(\omega_1)$; call it $\tilde{V}(a, \omega_1)$. Then, substituting $\tilde{V}(a, \omega_1)$ for the expression $\sum_{\omega_2 \in \Omega} V(\alpha, \omega_2)\, f[\omega_2 \mid \alpha(\omega_1), \omega_1]$ in eq(3), we define the expected discounted value function:

$$V(a, \omega_1) \equiv r(a, \omega_1) + \delta \tilde{V}(a, \omega_1). \tag{18}$$

Then, using $V(a, \omega_1)$ as the reinforcement function in eq(7), actions with higher values of $V(a, \omega_1)$ will be more likely than actions with lower values, and that a learning and/or evolutionary process sensitive to $V(a, \omega_1)$ would tend to drive out the poorer performing decision rules.

However, there is no free lunch; the function $\tilde{V}(a, \omega_1)$ sweeps all the previously mentioned computational problems under the rug. Where does it come from? How does it change over time? Why would our brains use this untethered function to steer behavior? How

could it become tethered to the true discounted present value of future rewards?

3.4 Summary.

Exact solutions to most real-life MDPs are beyond human computational abilities. The behavioral approach drops the focus on optimal solutions and instead considers probabilistic behavioral rules. Reinforcement learning can be effective for one-period problems. A backward induction version of reinforcement learning can in principle lead to optimal behavior, but due to the multiply infinite time needed for convergence it is also impractical for all but toy MDPs with very short horizons. The model-based simulation approach produces behavioral rules from simulations of the discounted present value of rewards generated by a model, but there can be significant errors due to practical constraints on the number of simulations. Selection among models depends on actual realizations of rewards from the model-based behavioral rules, but the number of trials needed for accurate assessment of model performance is not likely to be feasible in practice.

Where do we go from here? Surely it is reasonable to consider the path of physical science in seeking a simplifications of models of the world that can yield results with practical implications. Since we are addressing the internal working of our mind, introspection is a natural place to start.

4. Introspection about My Behavior.

If I am honest with myself, I will admit that much of my behavior seems automatic. Obviously there are the autonomic body functions (breathing, heart rate, blood pressure regulation, hormone levels, digestion, blinking, etc.) that are performed perfectly well by my *autonomic nervous system* without any conscious direction. In addition to body functions, instinctive behaviors such as fight or flight, motherly nurturing, blushing (and other reflexes), etc. are autonomic.

4.1 Routine Behavior.

While not instinctive, learned behaviors that become habits or routines also happen

without any conscious micro-management once initiated. These include getting out of bed, brushing my teeth, putting on clothes, rubbing my beard, twiddling my thumbs, pouring a glass of water, writing my name, walking, routine speech (such as "hello, how are you?").

The language of computer programming is useful for describing these routine behaviors. A computer program contains a main program and typically subprograms (a.k.a. subroutines), which in turn may contain subprograms. The purpose of a program is to compute a function of its inputs, and the result of that computation is its output. In abstract terms, this is what a behavioral rule does, so it can be useful to think of a behavioral rule as a computer program. For example, a behavioral rule for eating soup may first locate the edges of the bowl and the current positions of my spoon and my mouth; a subprogram may map out a physical path of my spoon to the bowl and then to my mouth; and a another subprogram may cause a sensory-motor response that brings about that motion, and another subprogram may cause a sensory-motor response of opening my mouth and inserting the spoon into my mouth and swallowing, all without spilling any soup.

Note that the computer program results in not only an immediate action but also a *plan* of actions. Indeed my behavioral rules generally produce plans of actions, in which future actions are contingent on updated sensory input about the current situation (as the current location of my spoon relative to the bowl and my mouth). The plan is carried out (i.e. the subprograms are executed without conscious effort) unless interrupted by a danger alert, an unexpected change in the planned path (my arm gets bumped), or a more pressing call by a higher-level program (such as being asked a question).

As routine behaviors, I seldom consciously activate a particular program; they just seem to happen. A major difference between my brain and a common digital computer is that my brain can be running many programs at the same time in contrast to the computer I am typing on at this moment which runs one program at a time. However, there have been many advancements in designing computers (or networks of processors) that can run multiple programs at the same time, and these parallel-processing computers are a much better analogy to my brain. With parallel processing the results of one program can be sent to another program that is already running, and so the direction of the latter program can be changed while it is running (even aborted). A program is activated when it is "called", and it is called when another

program is in a particular set of states and ultimately when my sensory system is in a particular set of states. For example, if my sensory system detects an immediate and serious danger, programs can be aborted and an emergency response program initiated. For another example, the detection of a question from my wife at dinner can interrupt the program for eating my soup, and initiate a program to answer her question. My reaction seems automatic unlike other situations in which I contemplate what action to take.

Given the limited capacity of my brain to process new information, and given the regularity of my world, it is more efficient to initiate programs that run more or less autonomously than to analyze every feasible action in real time. Indeed, I am not conscious of making decisions continuously, but instead sense that there is a default program running until such time that some input interrupts the default program or initiates a *contemplative* program. Even during contemplation, a very large proportion of the actions available to me are still generated by a default subprogram, and my contemplation focuses only on a very small number of non-default actions. By utilizing default programs, I free up brain capacity to consider other problems current or forthcoming.

I cannot enumerate all the programs stored in my brain, but I have no doubt there are billions if not trillions. I acquired these programs in various ways: genetic inheritance, imitation of observed behavior, lessons taught by my parents and teachers, and modification of existing programs. In the next section I will venture to model the process by which these programs are evaluated, and modified, retained or discarded.

4.2 Non-Routine Behavior.

In contrast to routines, non-autonomic non-instinctive behavior entails conscious contemplation about the available actions and potential consequences. Since contemplation takes time and requires substantial computational resources, it would not be efficient to engage those scarce resources for situations with negligible or minimal consequences. Somehow I anticipate upcoming situations where the potential future paths of states will diverge in ways that will have significant consequences for me. The time until the anticipated fork in those paths obviously limits how much time there is for contemplation. When the time is short (minutes), contemplation is more or less a cursory examination of the feasible actions and the immediate consequences. For example, I face a choice between flavors of ice cream and the situation

requires me to make a choice within the next minute. I quickly scan the choices and find the flavor that is most appealing (or produced the most pleasure in my memory of similar choices). With regularity in my experiences, such behavior can become routine and automatic, thus reserving deeper contemplation for more serious situations.

When the possible consequences of an upcoming action includes potential loss of life, wealth or status, it behooves me to anticipate that situation early enough to afford sufficient time for contemplation of possible plans of action and their long-term consequences. When I was a child I had no ability to anticipate such situations. The responsibility fell on my parents. For example, "What Halloween costume do you want this year?" "What summer camp looks interesting?" "Whom do you want to invite to play this Saturday?" As I grew up, teachers and peers prompted me to contemplate future decisions and future consequences, such as going to college and/or choosing a career. Moreover, I acquired many models that were recommended for various situations. As my models of my world improved, they incorporated the regular daily and seasonal patterns, and past experience and memory enabled me to anticipate scenarios that could occur. The point is that I receive significant help and cues from my environment that prompt timely contemplation.

As an adult, my typical mode of contemplating the future consequences of current actions is to consider several plausible plans of action (due to similarities with past situations or conventions) and to generate possible *scenarios* of future consequences for each plan using a model of my world. I say "a" model, since I have many models that are tailored to different kinds of situations. The appropriate model for the current situation seems to pop up automatically, but on rare occasions I may consider several possible models as part of my contemplation.

Besides imagining possible future actions and consequences, contemplation involves an *evaluation* of those consequences. Often choosing a plan comes down to a gut feeling for what is "best" for me in some ineffable sense - in contrast to the concept of conscious rational decision making. At other times, the evaluation process entails anticipated rewards (feelings) along the paths that may be realized.[14] When my behavior can be observed by family or someone

[14] Indeed, this process is what I formalized in eq(11) above.

important to me, my future rewards can include how I think my action will affect their opinion of me and their future reaction to me. As a child, anticipated parental rewards often dominated all other potential rewards. As a teenager, anticipated peer rewards had considerable weight. As an adult, anticipated rewards from my spouse, employers, and professional peers have considerable weight. I also can anticipate self-reward from successfully implementing a previous plan that I designated as important.

As an aside, of particular interest to me as a researcher of human behavior are laboratory experiments in which a subject is presented with a choice problem. Often the problem is described in abstract or context-free terms and is unlike any situation the subject has encountered before. Moreover, it is assumed that the subject's behavior in the experiment will not affect the subject's future path through life. In many experiments the connection between the subject's behavior and the outcome of the experiment is complicated and unlikely to be fully understood by the subject, and thus the subject's cognitive ability can have a non-negligible effect on the observed behavior. As a researcher I hope (perhaps in vain) that humans have generalized contemplative programs that can be adopted to the experiment task I present them with, and that the behavior I observe and the inferences I make about those programs have relevance beyond the laboratory situation.

In summary, introspection leads me to categorize my behavior into (i) autonomic and routine behavior, and (ii) contemplative consideration of potential plans of action, the possible future consequences of each plan, and an evaluation process that leads to one plan being initiated. In both cases, the language of computer programs is useful for describing these behavioral rules and processes. It is clear to me that many of these behavioral rules (programs) have been hard-wired by genetics, and many have been suggested by parents, teachers and peers.

What I haven't addressed so far is how I learn which programs are best for which situations. In other words, how do I evaluate and prioritize these programs (whether consciously or unconsciously)? That is the topic of the next section.

5. The Evolution/Learning of Behavior in Real-Life Problems.

Since reproductive success is the ultimate engine of Darwinian evolution, DNA that

encodes instincts and reflexes which increase reproductive success have become more common in the human species, and those that decrease reproductive success have become significantly less common. The dynamics of evolution are analogous to reinforcement learning, but on a time scale of millions of years.

Behavior that is not directly determined by DNA must be learned during the lifetime of an individual human and usually early in life. Reinforcement learning is one possible mechanism. The reinforcement function is the crucial ingredient. Where does it come from? The evolutionary viewpoint implies that actions in particular states are positively (negatively) reinforced to the extent they increase (decrease) the probability of reproductive success: denote it by $p_s(a, \omega)$. To operationalize this principle, our brain must have a model (encoded via DNA) of how the actions in particular states affect reproductive success. Without being specific, we can simply assume that the output of such a model is a reinforcement function $r(a, \omega)$ that is a monotonic transformation of $p_s(a, \omega)$. Obviously the evolutionary advantage of reinforcement learning is the ability to learn behavior on a human time scale, and we can presume that natural selection has given us a model that is better than none.

A shortcoming of this evolutionary viewpoint is that for many states, reproductive success, $p_s(\, , \omega)$, will be the same for almost all feasible actions. Therefore, almost all feasible actions will be equally likely to be taken. But we observe that they are not equally likely, so this evolutionary viewpoint is incomplete. One way to resolve this dilemma would be to exaggerate the effects: i.e. the reinforcement function $r(a, \omega)$ outputted by the model is a monotonic transformation of reproductive success such that $\partial r(a, \omega)/\partial a = \kappa \, \partial p_s(a, \omega)/\partial a$, where κ is a large positive scalar.[15] This transformation approximates the logic that if option A has a higher likelihood of a Win versus a Loss than option B, then the optimal decision is option A; i.e. never choose option B. With this modification, reinforcement learning would produce a non-uniform distribution over state-contingent actions and change within a lifetime.

On the other hand, if we want to predict behavior, we need a fairly good proxy for $r(a, \omega)$. Economists typically assume that monetary outcomes are the dominant consideration, and so

[15] For discrete actions, let a_0 denote a default action, and then let $r(a_0, \omega) = 0$, and $r(a, \omega) = \kappa[p_s(a, \omega) - p_s(a_0, \omega)]$. For even more exaggeration, we could have $r(a, \omega) = \kappa \, ln[p_s(a, \omega)/p_s(a_0, \omega)]$.

the reinforcement function is taken to be the monetary outcome. To the extent that reproductive success is correlated with wealth, the economist's assumption can be a reasonable approximation of the true reinforcement function for many states. Other potentially dominant measures that approximate $r(a, \omega)$ for some states are social status and know-how. For example, "do favors for powerful people" and "listen to your teacher".

The potential application of the MDP framework begins with T-period problems with $T = 2$. Section 3.1 on reinforcement learning shed doubt on the efficacy of using the experienced rewards to learn due to the much larger number of samples required than for one period problems. Section 3.2 explored how model-based simulations could help; but for $T \geq 3$, the computational burden is likely to exceed the capacity of our brains, so the performance of the resulting behavioral rules may be far from optimal.

Where does this leave us? Nature provides a clue. Newborn mammals seem instinctively to follow and imitate one or both parents. Thus, imitation provides the newborn with behavioral rules that have been selected for over thousands of generations. These observed and imitated behaviors as well as instincts, plus variations of these, are the initial collection of rules for the newborn. Further, the parent rewards "good" behavior and discourages "bad" behavior, and this reinforcement can induce the creation of neural circuits in the offspring that internalize this pattern of reinforcement as part of the offspring's model of the world. Such a process leads to learned routines that tend to increase reproductive success of the offspring.

Obviously, a similar process also operates in the human species. However, human children appear to challenge their parents far more than non-human children. Perhaps this is due to the larger, more imaginative and exploratory human brain. On the other hand, humans live in groups and in a culture. Typically we have multiple role models that provide behaviors to imitate and give reinforcement. Humans seem to have some innate ability to detect whom to trust, and those who are trusted provide even more examples, advice and reinforcement than parents of non-humans. Coupled with this ability, humans are somewhat skeptical,[16] and this skepticism often motivates us to test advice by running our own internal simulations. Indeed, this extension beyond the immediate family could be a significant driver in human

[16] Perhaps such skepticism has a survival advantage by preventing us from getting stuck in a local maximum.

advancement.[17]

It is not unreasonable to suppose that a similar process applies to multi-period problems that require contemplation. First, our parents, family, teachers and other trusted people provide us with a toolbox full of models of the world. Many of these models pertain directly to survival and reproductive success. Many more models pertain to how our behavior affects short-term and long-term wealth, status and know-how which are highly correlated with reproductive success. Indeed, modern society has institutionalized this process through formal schooling, and science provides a reasoned empirical approach to improving those models of the world. Second, these same adults teach us "meta-programs" designed to alert us to significant upcoming situations that we should contemplate.

The "advice" from parents, family, teachers and other trusted people fits well into the "value function" framework of Section 3.3. In essence it can be summarized as $\widetilde{V}(a, \omega_1)$, the expected discounted value of future rewards from taking action $a \in \mathcal{A}(\omega_1)$, that appears on the r.h.s. of eq(18). Moreover, advice often comes with an explanation, which can be considered a model of the world f^i, where the superscript identifies the source of the advice and model. When do I follow this advice rather than use my default program/model?

This advice selection problem can be addressed in the same way that selection of models was addressed in Section 3.2, eq(16). Let $AU(f^i)$ denote my assessment of the advice/model f^i. This assessment could be part of the advice itself, come from my own simulations, or come from actual experience if I have followed this advice in the past. Then adding f^i to my space of models $\mathcal{F}$, eq(16) defines the probability $\psi(f^i)$ of following this advice/model. In contrast to model-based simulation, which is not computationally feasible for most real-life MDPs, advice is often a ready source for the expected performance $AU(f^i)$. Indeed, advice liberates us from having to learn only from our own experience or simulations.

On the other hand, with multiple sources of advice there are bound to be conflicts. Without our own experience or the ability to run reliable simulations, how do we weigh conflicting advice? I suggest that we weigh advice based on the trustworthiness and/or prestige

[17] See R. Heinrich, (2016). *The Secret of Our Success*.

of the advisor. Formally, let τ^i denote the trustworthiness of source i, and modify eq(16) by multiplying each $AU(f^i)$ by τ^i; then $\psi(f^i)$ will incorporate our assessment of trustworthiness when selecting a model to use.

$$\psi(f^i) = exp[\tau^i AU(f^i)] / \sum_{f^h \in \mathcal{F}} exp[\tau^h AU(f^h)]. \quad (19)$$

Naturally, τ^i would adjust over time, as our experience confirms or conflicts with the advice of source i. In other words, we learn during our life whom to trust and how much weight to give their advice. When we are children, our parent's advice typically outweighs any other advice. As we age, extended family and teachers are given more weight. When we are taught chemistry and physics, the perceived authority of our professors and the evidence supporting these sciences induces overwhelming weight relative to superstitions, folk lore and prior models. Similarly, since our own experience with multi-period problems is severely limited, when we read and hear stories about human interactions from many others, these inputs often out-weigh our limited experience because they accumulate the experience of our culture. Our inclination to give considerable weight to these inputs can be derived from the positive correlation of these inputs with reproductive success.

Of course there could be many cases in which the received community wisdom is compatible with several models and rules, so there is a significant probability on more than one model and rule. Nonetheless, without this influence of societal experience on our selection of models and rules for multi-period problems it would be very much more difficult to learn behavioral rules that improve our survivability. Moreover, in the modern society, the survival and evolution of our models and behavioral rules may well be more important than the survival of individuals.

6. **Extensions.**

The application of model-based rule learning is not limited to MDPs. Games (such as

Chess and Go) that involve other human beings has already been addressed quite successfully by deep reinforcement learning. I introduced the concept of rule learning in static normal-form games [Stahl, 1996, 2000; and Stahl and Haruvy, 2012]. Unlike formal game theory, instead of seeking a model of behavior common to all players and consistent with rationality, archetypal models of other players define a space of models, and reinforcement comes from the payoffs experienced (and counterfactually simulated). The complication in modelling other players is that they also can have a model of other players, so your model of them should contain their model of you, and so on. How deep such models go is an empirical question, and research suggests a very limited depth in reality.

The central question addressed in this essay has been how ordinary humans behave in real-life sequential decision problems they face. Can we extend any insights to the case of agency behavior (i.e. firms, governments, and organizations)? Formally, we could simply interpret the decision-maker in a MDP as an agency. However, our assumptions about the computational abilities of the DM need to be adjusted significantly upward. For instance, NASA has clearly solved the problem of how to send a human to the moon and back. Pharmaceutical firms use sophisticated chemical and biological models in developing new drugs.

While genetic evolution is not the mechanism for passing along learned behavior among agencies, as a culture we have huge stores of information such as success/failure stories, best-practices, engineering know-how and science-based models. In this realm, it is not the number of copies of specific information that matters most for progress[18], but rather it is the likelihood that it will be used. It is this likelihood of use that is the output variable in the model-based rule learning developed in this essay, so model-based rule learning may be quite applicable to firms and organizations.

7. **Conclusions**

I have used Markov Decision Problems as a framework to think about how humans behave in the real life sequential decision problems they face. One advantage of this framework

[18] Although one might point out that the destruction of the Alexandrian library had a significant negative effect on human progress that lasted for a thousand years.

is that there is a well-developed theory establishing the existence of an optimal solution, and algorithms for computing an optimal solution. This foundation has fostered huge advances in machine learning, especially deep reinforcement learning. It also presents a possible foundation for human behavior and learning.

It is not a coincidence that reinforcement learning is derived from the theory of Darwinian evolution (a.k.a. replicator dynamics). Indeed, evolution provides a plausible explanation of why our DNA codes for autonomic and instinctive behavior that is vital to survival. Darwinian reinforcement has a simple binary form: survival or death.

A defining characteristic of instinctive behavior is that it is a fast response to stimulus that signals danger, in which the immediate consequence is critical to survival. In the MDP framework, these are one-period problems. One-period problems for the most dangerous situations have been "solved" and encoded in our DNA, but other one-period problems have not. For these one-period problems, reinforcement learning is a feasible and practical model of how humans can learn and act in a manner that is nearly optimal. It has been used successfully to explain animal behavior and some human behavior. The reinforcement is presumably an internal feeling (e.g. pain or pleasure) that has a genetic basis having evolved due to its correlation with reproductive success.

However, for most multi-period problems, the computational requirements become exponentially larger and beyond the ability of human brains. Nonetheless, it may be that human behavior is still at least partially driven by reinforcement learning. The concept of model-based rule learning is one way humans might simulate potential plans for multi-period problems and the potential internal feelings those plans may produce. But the more important source of reinforcement in these cases comes from parents, teachers and culture.

Clearly DNA that would encode parental approval/disapproval as pleasure/pain, especially for children would have survival value, so it is reasonable to assume DNA does such encoding. It is a small step to suppose that this encoding extends to grandparents and other trusted adults. Indeed, the categorization of others as trustworthy or not trustworthy definitely affects one's survival and reproductive success, and hence it likely to have a genetic basis.

This external source of reinforcement greatly widens the applicability of reinforcement learning. Parents, teachers and other trusted adults can not only reinforce behavior, they can suggest multi-period plans that may not occur to the child, and provide their evaluation of the consequences of those plans. Since the child will have little or no personal experience, this external reinforcement will play a dominant role in learning.

The term *model-based* rule learning emphasizes the central role of the model of the world that underlies forecasts of future consequences. Passing from one generation to the next the accumulation of know-how and the models that work best has undoubtedly had a major role in the development of human society. We have formalized this process as the scientific method and created institutions to perform this function systematically. Reinforcement in this process comes from teachers and compliments the internal reinforcement of discovery.

Despite the successes of human society, it is important to recognize the astronomical computational requirements of real-life MDPs and the impossibility of convergence in finite time. Therefore, we must admit that all our models are wrong, all our rules are suboptimal, and that we are at best making incremental improvements.

The most significant improvements in our models of the world have come through the institutions we have created to produce and propagate those improvements. These institutions contain astronomically more information than can be encoded in our DNA. Perhaps our future will depend more on the survival of that information via those institutions than on the survival of individual humans.

8. **Bibliography.**

Arulkumaran, K., Deisenroth, M., Brundage, M., and Bharath, A. (2017). "A Brief Survey of Deep Reinforcement Learning," *IEEE Signal Processing Magazine*, arXiv:1708.05866v2.

Bellman, R. (1954). "The Theory of Dynamic Programming," *Bulletin of the American Mathematical Society,* **60**, 503–516.

Bertsekas, D. P. (2000). *Dynamic Programming and Optimal Control*, (2nd ed.), Athena Scientific.

Bush, R., and F. Mosteller. (1951). "A Mathematical Model for Simple Learning", *Psychological Review,* **58**, 313-323.

Henrich, J. (2016). *The Secret of Our Success,* Princeton University Press.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis.* Wiley & Sons, New York.

Mnih, V., et al. (2015). "Human-level Control Through Deep Reinforcement Learning," *Nature*, doi:10.1038/nature14236.

Stahl, D. (1996). "Boundedly Rational Rule Learning in a Guessing Game," *Games and Economic Behavior*, **16**, 303-330.

_____ (2000). "Rule Learning in Symmetric Normal-Form Games: Theory and Evidence," *Games and Economic Behavior*, **32**, 105-138.

Stahl, D. and E. Haruvy. (2012). "Between-Game Rule Learning in Dissimilar Symmetric Normal-Form Games," *Games and Economic Behavior*, **74**, 208-221.

Taylor, P., and L. Jonker. (1978). "Evolutionary Stable Strategies and Game Dynamics", *Mathematical Biosciences,* **40**, 145-156.

Thorndike, E. (1898). "Animal Intelligence: an Experimental Study of the Associative Processes in Animals," *Psychological Monographs* #8.