Optimal habits can develop spontaneously through sensitivity to local cost

Theresa M. Desrochers[a,b], Dezhe Z. Jin[c], Noah D. Goodman[a], and Ann M. Graybiel[a,b,1]

[a]Department of Brain and Cognitive Sciences, [b]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139; and [c]Department of Physics, Pennsylvania State University, University Park, PA 16802

Contributed by Ann M. Graybiel, September 8, 2010 (sent for review August 16, 2010)

Habits and rituals are expressed universally across animal species. These behaviors are advantageous in allowing sequential behaviors to be performed without cognitive overload, and appear to rely on neural circuits that are relatively benign but vulnerable to takeover by extreme contexts, neuropsychiatric sequelae, and processes leading to addiction. Reinforcement learning (RL) is thought to underlie the formation of optimal habits. However, this theoretic formulation has principally been tested experimentally in simple stimulus-response tasks with relatively few available responses. We asked whether RL could also account for the emergence of habitual action sequences in realistically complex situations in which no repetitive stimulus-response links were present and in which many response options were present. We exposed naïve macaque monkeys to such experimental conditions by introducing a unique free saccade scan task. Despite the highly uncertain conditions and no instruction, the monkeys developed a succession of stereotypical, self-chosen saccade sequence patterns. Remarkably, these continued to morph for months, long after session-averaged reward and cost (eye movement distance) reached asymptote. Prima facie, these continued behavioral changes appeared to challenge RL. However, trial-by-trial analysis showed that pattern changes on adjacent trials were predicted by lowered cost, and RL simulations that reduced the cost reproduced the monkeys' behavior. Ultimately, the patterns settled into stereotypical saccade sequences that minimized the cost of obtaining the reward on average. These findings suggest that brain mechanisms underlying the emergence of habits, and perhaps unwanted repetitive behaviors in clinical disorders, could follow RL algorithms capturing extremely local explore/exploit tradeoffs.

free-viewing | naïve monkey | reinforcement learning | saccade

Reinforcement learning (RL) theory formalizes the process by which rewards and punishments can shape the behaviors of a goal-seeking agent—person, animal, or robot—toward optimality (1). RL algorithms have been widely applied in neuroscience to characterize neural activity in animals and human subjects, most famously for the dopamine-containing systems of the brain and related brain regions (2–5). These ideas have also been influential in the study of habit learning, in which habits are typically thought to arise when behaviors, through repetition, eventually become reinforcement-independent, stimulus-response (S-R) associations that can be executed in a semiautomatic manner (6).

In most learning experiments designed to test these ideas, a small range of relationships between actions and reward is imposed, cost-benefit ratios are explicit, and fixed and usually limited numbers of response choices are available, as for example when human subjects are asked to move a cursor in one direction to receive a monetary reward in a computer game, or when rodents are trained to press one or a small set of levers to receive a food reward. RL algorithms of varying complexity robustly account for decision-making behavior in many such experiments (7–9). But what if the actions needed to receive the reward were sequential, there were an exponentially large number of choices, the rewards were not predictable, and no explicit instructions were given to the agent? Such conditions occur in everyday life (10, 11). In computer science, the challenge of forming optimal sequential behaviors due to the vast number of possibilities has been highlighted by the traveling salesman problem, in which the goal is to minimize the total distance of visiting a given number of cities exactly once; this optimization problem has been difficult to solve (12). It is still unclear whether, and how, animals and humans learn optimal habits facing analogous challenging situations (8, 11).

We designed a task to incorporate such extreme uncertainty, we applied this task to experiments in macaque monkeys, and we then tested whether RL algorithms could account for the observed behaviors. In this free-viewing scan task, experimentally naïve monkeys were free to choose their own saccadic sequences to scan a grid of dots, one of which was randomly baited. In each experimental trial, the monkeys could not predict when or where the target would become baited. Thus, no particular saccade or saccade sequence could be relied upon to produce a reward, meaning that no fixed S-R "habit" could be acquired to solve the task. Moreover, the reward in every trial was identical, so that the only way to improve cost-benefit ratios was to minimize cost by reducing as much as possible the total length of scanning before reward delivery. As such, the task had many similarities to the traveling salesman problem.

Despite this complexity, monkeys, without instruction, developed repetitive scan patterns that were optimal or nearly optimal for solving the task. Moreover, the evolution of their preferred scanning patterns changed markedly through months of task performance, despite the fact that the monkeys maximized total rewards per session and minimized total travel distance (cost) per session relatively early on. With standard session-based analysis, the habitual behaviors appeared to change without reinforcement as a driving force.

However, within session, trial-by-trial analysis of the monkeys' scan-path transition probabilities showed that their behaviors could be accounted for, and closely simulated, by simple RL algorithms emphasizing local, trial-by-trial cost reduction. These findings demonstrate the power of short-time RL analysis and suggest that, even under highly uncertain conditions, relatively simple learning rules can govern the formation and optimization of habits. This deep-structure of habit formation may be critical not only for the emergence of habits and mannerisms in everyday life, but also for the insistent quality of repetitive behaviors occurring in neuropsychiatric disorders.

Results

We recorded the eye movements of two experimentally naïve monkeys as they were exposed to a free-viewing scan task that they

Author contributions: T.M.D. and A.M.G. designed research; T.M.D., D.Z.J., and N.D.G. contributed new reagents/analytic tools; T.M.D., D.Z.J., and A.M.G. analyzed data; T.M.D. and A.M.G. performed surgeries; T.M.D. trained and recorded from the monkeys; D.Z.J. did the simulations; N.D.G. designed the NMF analysis; and T.M.D., D.Z.J., and A.M.G. wrote the paper.

The authors declare no conflict of interest.

See Commentary on page 20151.

[1]To whom correspondence should be addressed. E-mail: graybiel@mit.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1013470107/-/DCSupplemental.

performed during daily sessions across months (Fig. 1). Each monkey was head-fixed and seated in front of a computer screen on which a grid of four or nine green target dots was presented. The monkey was free to move its eyes in any way as long as its gaze remained in the space occupied by the green grid. After a variable 1- to 2-s Delay Scan period, which prevented the monkey from receiving reward immediately, one of the dots was chosen to be the baited target according to a pseudorandom schedule. The start of this Reward Scan period was not signaled to the monkey. Capture of the baited target when the monkey's gaze entered the target window immediately extinguished the green targets (processing delay, mean 61 ± 61 ms SD). The trial then proceeded through the reward delay, reward, and the intertrial interval (ITI), as illustrated in Fig. 1. Because the monkeys were naïve, to improve performance during initial sessions, behavioral shaping such as gradually lengthening the Delay Scan periods was employed (*Methods* and Fig. S1), but the monkeys were never given instruction. Both monkeys learned the four- and nine-target tasks, averaging ~70% rewarded trials overall (Fig. S1*A*). The Total Scan Time was typically 1.5–4 s and contained 7–20 fixations (Fig. S2 *A–D*).

Despite the absence of explicit training on how or whether to move their eyes in the task, the monkeys developed particular repeated patterns of saccades to scan the target grids (e.g., Fig. S2 *A* and *B*). To examine these patterns, we selected the most frequent "loop" sequences: sequences that began and ended on the same target compiled from a pool of the top 20 most frequent five-fixation sequences across all rewarded trials and sessions (*SI Methods*). We calculated the percentage of trials in each session that contained each loop sequence. The percentages for the loop sequences were not constant across sessions. Instead, the loop sequences were acquired and dropped throughout the months of task performance (Fig. 2).

To test this conclusion without assuming either fixed-length sequences or deterministic scan patterns that did not account for all saccades (Fig. S2*E*), we turned to a probabilistic analysis of the scanning behaviors. We compiled for each session the transition probabilities of saccades between all pairs of targets. These transition probabilities were decomposed using nonnegative matrix factorization (NMF) into positive mixtures of transition components (13). These transition components represent the most explanatory parts of the overall transition patterns (deterministic or not) that could be seen at the level of saccades between adjacent targets.

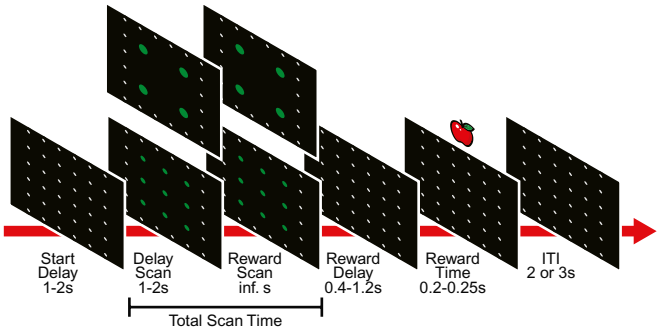

Fig. 1. Schematic of the free-viewing scan task. There was no requirement for the monkey's eye position when the gray grid was displayed. After a variable Start Delay, the green target grid was presented indicating the start of the Scan Time. When the green target grid was displayed, and once the monkey's gaze entered the area defined by the green grid, the only requirement was that the eye position remained in that space. After the variable Delay Scan, the Reward Scan began when a randomly chosen target was baited without any indication to the monkey. There was no time limit on the duration of the Reward Scan. Once the monkey captured the baited target by fixating or saccading through it, the green grid immediately turned off and the trial proceeded through the remaining task periods as illustrated. If the monkey's eye position exited the green grid area before capturing the baited target, the trial was immediately aborted by extinguishing the green target grid, and no reward was delivered.

The results were striking (Fig. 3). The weights of the NMF factors, like the favorite loop sequences, changed over time and were clustered into epochs across sessions rather than being randomly distributed. This clustering into temporally localized epochs exceeded chance, as tested by applying the same NMF analysis to random permutations of the saccade data ($P < 0.01$, dispersion permutations test; *SI Methods*). The NMF analysis thus also suggested that the monkeys' repeated scan patterns systematically shifted until eventually they settled into particular habitual scan patterns.

We next asked whether reward and cost, key drivers in RL models, could account for the shifts of scan patterns. We calculated reward rate as the number of rewards earned in each session divided by the total amount of time spent in the Reward Scan for rewarded trials (Fig. 4*A*). Because the energy required by the extraocular muscles to maintain fixation is relatively little in comparison to that required by saccades and these muscles contract in proportion to saccade amplitude (14), we used the mean total distance that the monkeys' eyes traveled during the Reward Scan per trial for each session as an estimate of the cost of the scanning (Fig. 4*B*). We also examined other possible measures of cost (e.g., no. of saccades; Fig. S2*F*) and found all measures to have similar trends (see *SI Text* for further discussion). Despite the fact that the scan patterns continued to change across the entire experimental time, both reward rate and cost per session reached asymptote early in the sessions. For the four-target task, asymptotic reward and distance were reached by session 9 for G4 in both measures, and by sessions 10 and 20, respectively, in Y4. In the nine-target task, G9 had a relatively steady rate, and Y9 reached asymptote for the final 18 sessions (*SI Methods*).

These findings appeared to challenge RL models. Both reward and cost evolved to asymptotic levels, yet the scanning patterns of the monkeys continued to change long past the time that these values reached steady state. A further observation, however, suggested that the session averages might not have the resolution to show the effects of changing reward or cost. The scan patterns became more repetitive and habitual through the months of task performance, as evidenced by the steady decline of the entropy of the monkeys' eye movements (*Materials and Methods*; Eq. **1** and Fig. 4*C*). These entropy measurements suggested that, if the habitual patterns became optimal or nearly so, the pattern shifts might still conform to RL theory. In the face of the large variability intrinsic to the task, local changes in reinforcement might take over the learning if shifts in the sequences of saccades performed were "local" in time.

To investigate this possibility, we first asked whether the monkeys achieved optimal habitual scanning patterns. We computed the optimal scan path, defined as that which minimized the total distance to obtain the reward in a single trial on average. One simple component of an optimal solution to this task would involve the monkeys not making saccades during the Delay Scan period, when no target was baited. This strategy, however, was apparently not used by either monkey: both performed a nearly equal average number of saccades in the first second and the last second of each trial (Fig. S2 *G* and *H*). A second possible optimality would be for the monkey to maximize the number of saccades that passed through, rather than fixated on, targets. The monkeys did not use this strategy either; the percentage of saccades that passed through targets tended to decrease across sessions, not increase (Fig. S2*I*). Therefore, the optimal realized solution to this task must involve the eyes moving all through the Delay and Reward Scan periods.

During the Reward Scan period, revisiting a previously scanned target did not lead to reward. Thus, the optimal strategy would be to scan each target exactly once. Given that the onset of the Reward Scan was unpredictable and that there was no evidence of different behaviors in the Delay Scan and Reward Scan periods, we assumed that the monkeys treated the entire period as the reward period. This assumption necessarily means that the optimal scanning pattern was a loop pattern in which each target was visited once before returning to the original target. Among such

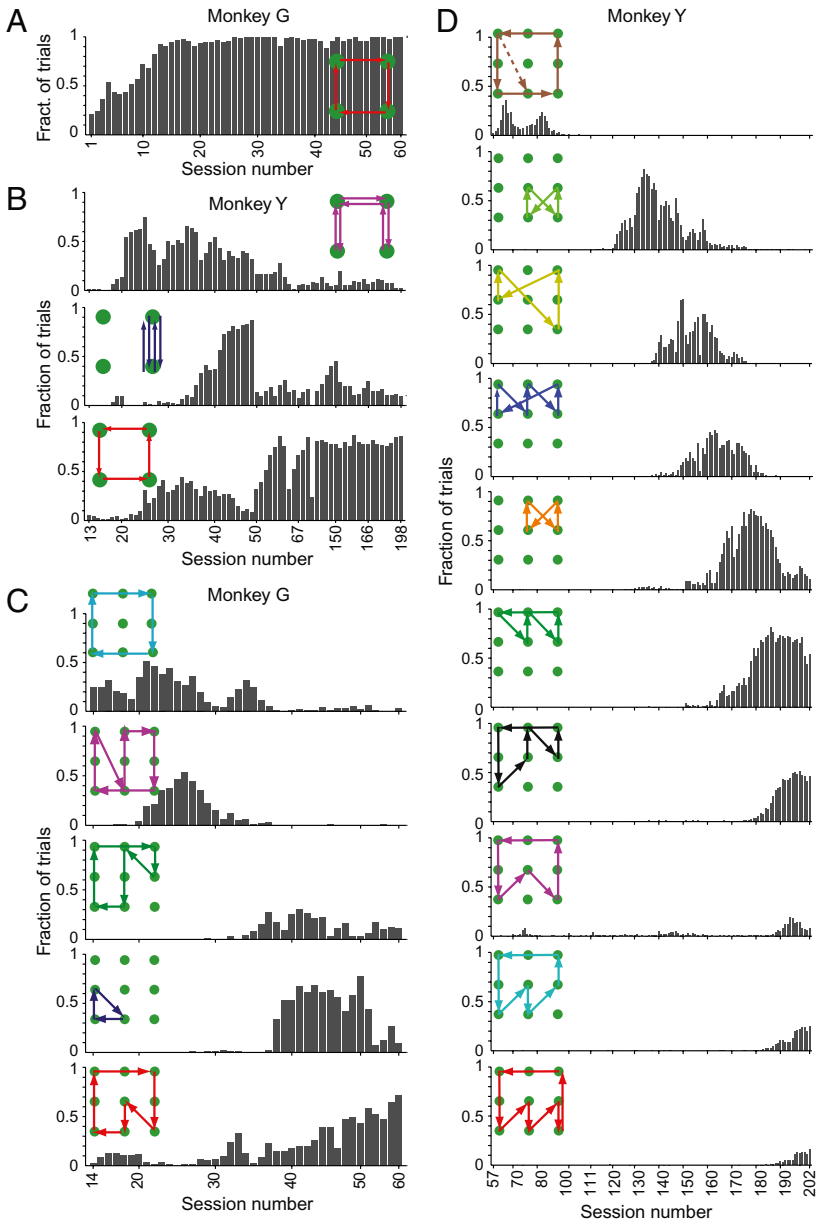

Fig. 2. Loop sequences emerge and shift during prolonged task performance. Each plot shows the fraction of rewarded trials per session containing the most frequent saccade paths that form a closed loop (start and stop on the same target), regardless of start or stop position during Total Scan Time. (*A*) Monkey G, four-target task (G4). (*B*) Monkey Y, four-target task (Y4). (*C*) Monkey G, nine-target task (G9). (*D*) Monkey Y, nine-target task (Y9). Dashed line in first panel indicates slight variation from main pattern included in the fraction.

patterns, the one with the minimum total distance should be the optimal policy.

This formulation of the scan task renders it similar to the traveling salesman problem, except that in the scan task, some targets could be passed through instead of fixated on, and no explicit instructions about the task were given. Given that there is no efficient computer algorithm for solving the traveling salesman problem, we predicted that the monkeys would not solve the task. Remarkably, they did.

We determined the optimal loop scan patterns that covered all targets once (Fig. S3) by using an exhaustive search algorithm (*SI Methods*), and then we compared them to the habitual patterns reached by the monkeys. Both monkeys reached the optimal pattern in the four-target task, and monkey G did so in the nine-target task (Fig. 2 *A*–*C*). Moreover, G9 reached the optimal path after transitioning through the sixth most optimal pattern (Fig. 2*C* and Fig. S3*B*, purple path). Y9 gradually approached the optimum by progressively covering more of the targets in the loop sequences, and her final pattern was near optimal, as it was the fifth most optimal (Fig. 2*D* and Fig. S3*B*). The monkeys thus not only generated habitual behaviors de novo and without explicit instruction, but also were able eventually to "solve" this task in the most optimal manner.

To determine whether RL was the driving force behind the evolution toward the optimal pattern, despite initial indications from session-averaged data, we performed trial-by-trial analysis of the effects of cost on the changes of the scan patterns. We excluded only initial trials on which behavioral shaping occurred (*SI Methods*). The analysis was based on the explore-exploit principle central to RL algorithms (1, 15). A fluctuation of the scan pattern from the previous trial was taken to represent exploration. If the monkey was "rewarded" with a shorter scan distance in the current trial (k) than the previous trial ($k-1$), the monkey could attribute the success to the differences between the scan patterns in the two trials ($k-1$ to k; *Materials and Methods* Eq. **2**). The monkey would then strengthen the differences between the scan patterns, and similar differences would likely occur again in the next trial (k to $k+1$). In contrast, if the distance were increased, the changes in the scan patterns should be extinguished so that similar changes would not likely occur in the next trial. This pattern of changes represents exploitation. Therefore, a positive correlation should exist between the reduction in distance from trial $k-1$ to trial k and the

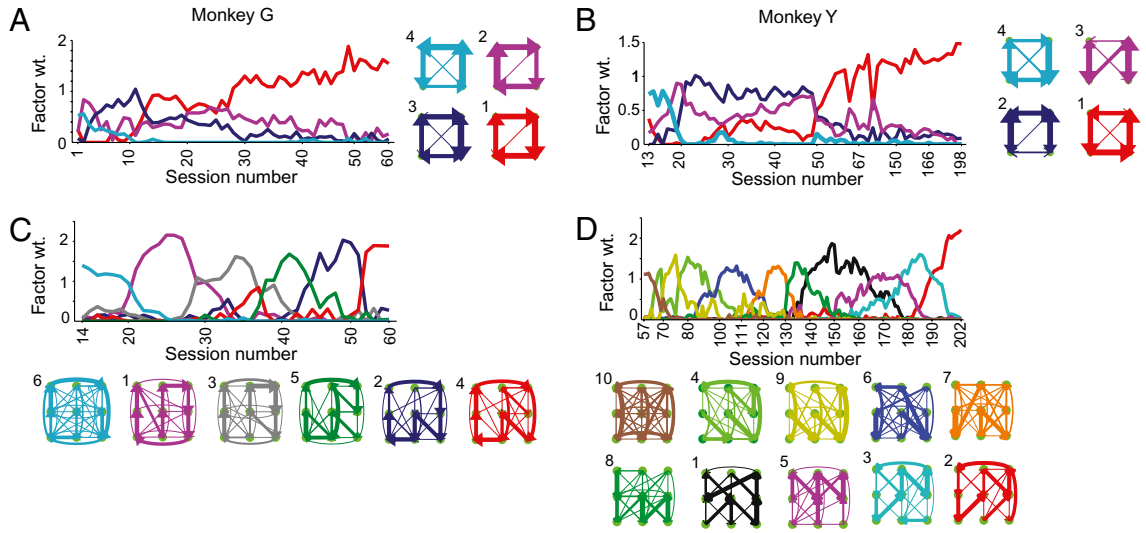

Fig. 3. Nonnegative matrix factorization (NMF) shows successive appearance of factors resembling loop sequences. Each panel displays the weight of each factor during Total Scan Time on all rewarded trials through task performance. Factors are diagrammed in colors to show similarity to the loop sequences in Fig. 2. Numbers on upper corner of the factors indicate their rank order by total magnitude (i.e., sum of the weight across sessions). (*A*) G4, rms error of factorization = 0.02673. (*B*) Y4, rms error = 0.02452. (*C*) G9, rms error = 0.0225. (*D*) Y9, rms error = 0.01728.

similarity in the changes of the scan patterns between trials $k-1$ to k and k to $k+1$ (*Materials and Methods* Eq. **3**).

Taking the transition probabilities to represent a given trial's scan pattern, we computed the correlation between the change in distance (cost) and the similarity of changes in scan patterns for each successive set of three contiguous trials $(k-1, k, k+1)$. With this "reinforcement test," we found that the cost vs. similarity correlations were small but highly significant for both monkeys in both conditions ($P < 0.002$, shuffle test; Fig. 5 and Fig. S4). These local changes in distance were also positively correlated with the number of trials that occurred before a particular loop sequence was repeated (*SI Text* and Fig. S5 *A–D*). These results from trial-by-trial analyses of the data strongly suggested that, contrary to the results based on session-averaged data, RL analysis could account for the changes in scanning behavior of the monkeys.

To demonstrate directly that RL could lead to the shifts of the monkeys' repetitive scanning patterns, we simulated an agent that performed the scan task and changed the agent's scan patterns according to a RL algorithm. The agent generated saccades according to transition probabilities that were determined by the action values of making the transitions (*SI Methods*). If an action value was large, the corresponding transition probability was also large. We implemented the REINFORCE algorithm developed by Williams (15) to perform RL (*SI Methods*). At each trial, the action values shifted randomly around the mean to produce explorations. The transition probabilities were used to generate a sequence of saccades. The baited target was randomly chosen and the reward was obtained as in the experiments. At the end of the trial, the mean action values were shifted toward the values at the beginning of the trial if the total scan distance was smaller than the previous trial, and otherwise, the values were shifted away. This protocol represented the exploitation stage.

The simulations closely resembled the performance of the monkeys on this task (Fig. 6). The reward rate, the distance the eyes traveled, and the entropy of the paths (Fig. 6 *A–C*) all paralleled those measured for the monkeys' behavior (Fig. 4). Reward rate and distance showed large changes only in the earlier sessions run, whereas entropy continued to decrease through task performance. Remarkably, the final most-probable loops reached by the simulations (Fig. 6 *D* and *E*) were equivalent to the final loop sequences in the monkeys (Fig. 2). Different runs of the simulation had different convergence rates, but they nevertheless

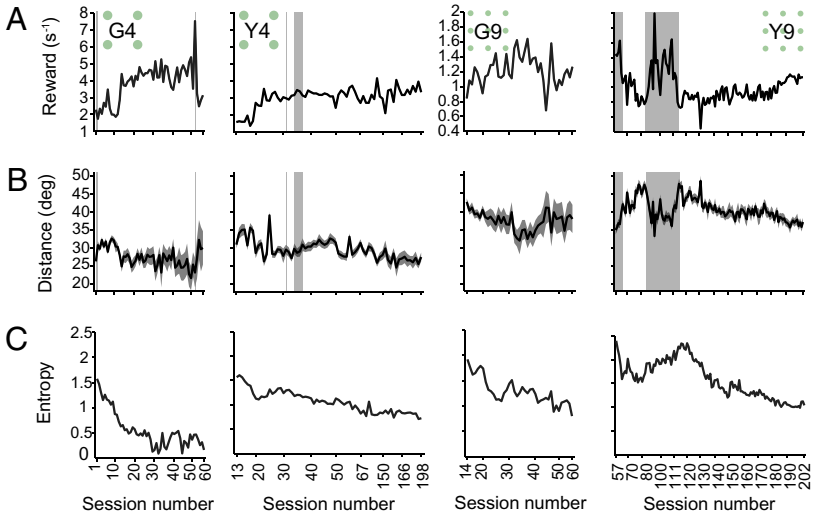

Fig. 4. Session-averaged behavioral measures. All rows show monkeys and conditions in the order depicted in *A* (G4, Y4, G9, and Y9). (*A*) Reward rate measured as number of rewards per total Reward Scan time in each session. (*B*) Mean saccade distance during Reward Scan per session with shading indicating approximate confidence limits (±1.96 × SEM). Gray vertical bars in *A* and *B* indicate sessions containing shaping periods when the task was made easier for the monkey (see *Materials and Methods* and Fig. S1). (*C*) Entropy of transition probabilities during Total Scan Time.

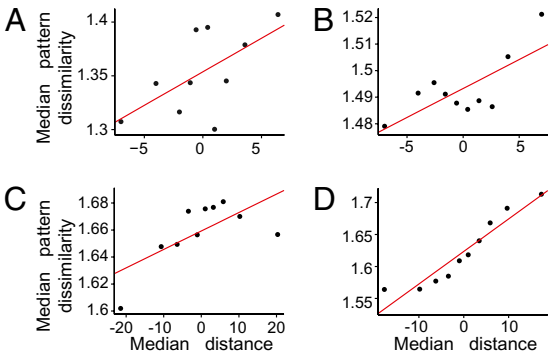

Fig. 5. Trial-by-trial reinforcement test shows correlation between cost and change in pattern. Distances were simplified to be the geometric distance: one unit is the horizontal or vertical distance between adjacent targets. The change in total scan distance and the pattern dissimilarity (one minus pattern similarity; *SI Methods*) for each trial was computed. Trials were then binned into 10 equal bins. The median of each of the 10 bins was plotted and used to compute the correlation (red line) between the change in distance and the pattern dissimilarity. The total number of trials (n), correlation coefficients (R), and correlation p values are listed below. Note this p value is different from the P value reported to indicate significance resulting from the shuffle test in the text. (*A*) G4: n = 6,109 trials; R = 0.613, p = 0.060; slope = 0.006. (*B*) Y4: n = 25,113; R = 0.737, p = 0.015; slope = 0.002. (*C*) G9: n = 5,912; R = 0.672, p = 0.033; slope = 0.001. (*D*) Y9: n = 54,214; R = 0.951, $p < 0.0001$; slope = 0.005.

converged on the same paths (Fig. S6). In addition, the session-by-session structure of the shifts from one pattern to another was strikingly similar between monkeys (Fig. 3) and the simulations (Fig. 6 *F* and *G*). The statistical dispersion of the weight of all but one Sim4 factor (no. 3) and one Sim9 factor (no. 9) was significantly lower than expected by chance ($P < 0.0001$, 10,000-session permutations test), much the same as found for the NMF factors in the monkeys. These results demonstrate that the relatively simple RL model we used captured the essence of the monkeys' complex behavior in this task. The simulations further validated the reinforcement test used for the monkey data to detect evidence of RL (*SI Text* and Figs. S7–S9).

Discussion

Strong associations between stimulus and response characterize many habitual behaviors. Through repeated trials and reinforcement learning, the associations are formed and strengthened. The S-R paradigm has been the focus of most experimental studies of habit learning, in which the tasks performed involved a limited number of stimuli and responses. But habits can consist of extensive sequences of behaviors, not just single responses, and the number of possible sequences can be enormous. The learning mechanisms underlying such habitual action sequences have rarely been studied experimentally. RL algorithms have been employed in computer simulations to learn sequential actions, such as in playing games (7, 9), and in a hippocampus-dependent water-maze learning in rats (16). Here we asked whether RL could account for learning such sequential habits and individualized rituals under conditions of extreme uncertainty.

Our task was designed to study the formation of such sequential habits. The stimulus was a simple grid of dots, but because the responses were self-generated saccade sequences, the number of possible sequences was virtually infinite. Given that the baited target was chosen at random in each trial, simply saccading to a particular target or a particular set of targets in sequence did not lead to the reward each time. Despite the uncertainty intrinsic to the task, the monkeys formed repetitive scan patterns without instructions, and the patterns evolved through months of performing the task. Eventually, the monkeys settled into highly stereotypical scanning patterns with low entropy, characteristic to habitual behaviors.

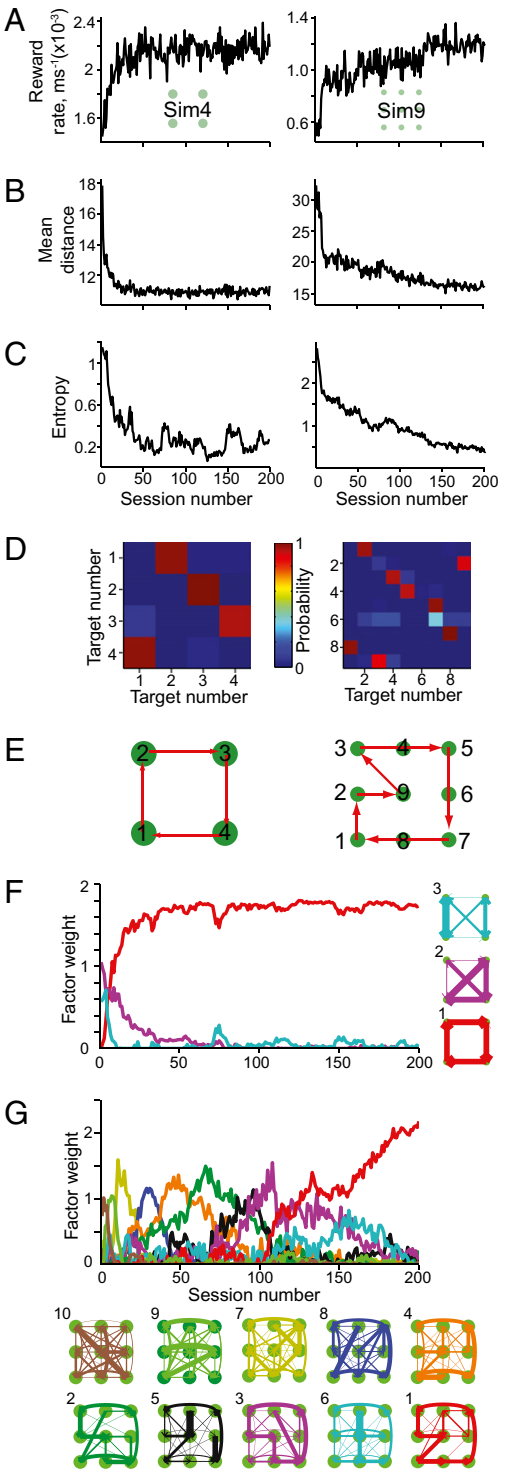

Fig. 6. The REINFORCE algorithm simulation performs similarly to the monkeys. (*A*–*E*) For each row, columns represent conditions depicted in *A*: REINFORCE simulation of the four-target task (Sim4) and REINFORCE simulation of the nine-target task (Sim9). Reward rate measured as no. of rewards per total simulated trial time in each session (*A*); mean geometric distance per session (*B*); entropy of transition probabilities per session (*C*); final transition probabilities (*D*); and resulting most probable pattern (*E*). (*F* and *G*) NMF of simulations as in Fig. 3 for Sim4 (*F*, rms = 0.02314) and Sim9 (*G*, rms = 0.03655).

We did not test whether reward devaluation affected the patterns, but the pervasiveness of these repetitive scanning patterns, especially in the many sessions of the four-target task, clearly

suggests that the patterns were habitual. The monkeys achieved optimal or nearly optimal scan patterns near the end of the experimental sessions. Despite there being many equivalent loop patterns for each level of optimality (rotations and reflections of Fig. S3), the monkeys chose to perform a single instance of each pattern. This fact again suggests that the monkeys have formed habitual sequences of eye movements without instruction.

It is remarkable that the monkeys solved this task at all, given their lack of instruction. This finding may mean that there is an innate drive toward optimality. Our experiments indicate that the monkeys could not have been using session-averaged reward amounts or cost to shape their behavior successfully—these mostly reached asymptote early in the experiments. But trial-by-trial analysis demonstrated that the changes in saccade patterns were correlated with the fluctuations of the total distances to capture the baited target and earn reward, itself the same for each trial. A simple RL algorithm simulating the task captured the richness of a naïve monkey's performance: the total cost (saccade distance) to obtain the reward in a single trial was, on average, minimized. Our findings thus suggest that even for sequences of actions made to receive a singular, identical reward in a highly uncertain context, nonhuman primates can use RL to acquire habitual optimal sequences of behavior.

There were large individual differences in the rates at which the monkeys learned the optimal patterns, both across monkeys and across task versions. Our RL simulations showed that the scan task has complex cost structure in the space of scan patterns. An unfortunate choice of actions during the learning could lead to long detours from the optimal patterns due to local minima. Consequently, the rates of convergence to the optimal patterns varied greatly from run to run in the simulations, even with exactly the same parameters (Fig. S6). The choices of the simulation parameters, especially the degrees of exploration and exploitation in each trial, also influenced the convergence rates, perhaps mirroring the effects of innate individual differences.

The reinforcement test we devised demonstrated that cost reduction is positively correlated with the similarity of the pattern changes on a trial-by-trial basis, as expected from the exploration-exploitation tradeoff inherent to RL learning. The correlation tended to be weak for at least three main reasons: (*i*) the shift in the patterns was the consequence of both learning and random exploration in each trial; (*ii*) the estimates of the transition probabilities in each trial, used to compute the similarities between the shifts of the scan patterns in consecutive trials, could be inaccurate due to the limited number of saccades; and (*iii*) the distance to capture the baited target fluctuated over a wide range due to its random assignment in each trial. Consequently, a large number of trials was required to show that the correlations were significant (see *SI Text* for further discussion). It is also possible that the simplifying assumptions made in the corresponding simulations, such as the creation of saccade sequences using the limited trial-by-trial transition probabilities, were not entirely adequate to mimic the monkey behaviors in great detail (*SI Text* and Fig. S5*E*). Nevertheless, this analysis and relatively simple algorithm captured much of the richness of the monkeys' behaviors.

Habits, mannerisms, routines, and repetitive behaviors of any sort can share the feature of semiautomaticity. We found that monkeys can acquire such habitual behaviors without training and without a simple or explicit S-R environment, and that the habitual behaviors acquired were nearly optimal in minimizing cost. Repetitive behaviors are not always advantageous, however. They can be dangerous when they allow predators to predict behavior. Further, repetitive behaviors frequently occur in disorders ranging from the stereotypies expressed in some neuropsychiatric conditions to the repetitive behaviors triggered as responses to anxiety-provoking situations (17, 18). Our findings suggesting that there may be a spontaneous drive to repetitiveness should thus have implications for a broad range of both normal and abnormal behaviors.

Materials and Methods

Monkeys' eye position was monitored using infrared eye tracking (500 Hz; SR Research Ltd.) and recorded using a Cheetah Data Acquisition system (2 KHz; Neuralynx, Inc.). Initial calibration of the eye movement signal required the use of a handheld treat; subsequent sessions used the four-target grid. There were three parameters used to shape the monkeys' behavior during task acquisition: (*i*) rewarding any target captured after the Delay Scan; (*ii*) adjusting the size of the window around each target that would trigger capture; and (*iii*) the duration of the Delay Scan (*SI Methods* and Fig. S1 *B–D*).

Sessions with poor monkey performance were not included in analyses (~4% overall). All eye movement analysis was done in Matlab. NMF was completed using a number of components that corresponded to a drop in residuals and the algorithm that produced the smallest error. The entropy of the transition probabilities for each session (q) was defined as:

$$E = -\sum_i q_i \sum_j q_{ij} \log_2 q_{ij}, \quad [1]$$

where q_i is the probability of observing target *i*, and q_{ij} is the probability of observing target *j* followed by target *i*.

The difference in the transition probabilities from trial $k-1$ to trial k was

$$\Delta(k, k-1) = P(k) - P(k-1), \quad [2]$$

where $P(k)$ is a vector whose components are the transition probabilities between all pairs of targets. The similarity S_k between the changes $\Delta(k+1, k)$ and $\Delta(k, k-1)$ was computed as the cosine distance measure

$$S_k = \frac{\Delta(k, k-1) \cdot \Delta(k+1, k)}{|\Delta(k, k-1)||\Delta(k+1, k)|}. \quad [3]$$

Details of analyses and the REINFORCE algorithm are given in *SI Methods*.

ACKNOWLEDGMENTS. We thank D. Wolpert for discussing these findings, and D. Gibson, J. Feingold, M. Cantor, and other members of the A.M.G. laboratory for help and discussions. This work was supported by National Eye Institute Grant EY012848 (to A.M.G.), by Office of Naval Research Grant N000014-07-10903 (to A.M.G.), and by a National Defense Science and Engineering Graduate Fellowship (to T.M.D.), a Friends of the McGovern Fellowship (to T.M.D.), and a Sloan Research Fellowship (to D.Z.J.).

1. Sutton RS, Barto AG (1998) *Reinforcement Learning: An Introduction* (MIT Press, Cambridge, MA).
2. Bayer HM, Glimcher PW (2005) Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron* 47:129–141.
3. Schultz W, Dayan P, Montague PR (1997) A neural substrate of prediction and reward. *Science* 275:1593–1599.
4. Samejima K, Ueda Y, Doya K, Kimura M (2005) Representation of action-specific reward values in the striatum. *Science* 310:1337–1340.
5. Morris G, Nevet A, Arkadir D, Vaadia E, Bergman H (2006) Midbrain dopamine neurons encode decisions for future action. *Nat Neurosci* 9:1057–1063.
6. Dickinson A (1985) Actions and habits: The development of behavioural autonomy. *Philos Trans R Soc Lond B Biol Sci* 308:67–78.
7. Barraclough DJ, Conroy ML, Lee D (2004) Prefrontal cortex and decision making in a mixed-strategy game. *Nat Neurosci* 7:404–410.
8. Montague PR, Dayan P, Person C, Sejnowski TJ (1995) Bee foraging in uncertain environments using predictive hebbian learning. *Nature* 377:725–728.
9. Daw ND, O'Doherty JP, Dayan P, Seymour B, Dolan RJ (2006) Cortical substrates for exploratory decisions in humans. *Nature* 441:876–879.
10. Song C, Qu Z, Blumm N, Barabási AL (2010) Limits of predictability in human mobility. *Science* 327:1018–1021.
11. Noser R, Byrne RW (2010) How do wild baboons (Papio ursinus) plan their routes? Travel among multiple high-quality food sources with inter-group competition. *Anim Cogn* 13:145–155.
12. Applegate DL (2006) *The Traveling Salesman Problem: A Computational Study* (Princeton Univ Press, Princeton, NJ).
13. Lee DD, Seung HS (1999) Learning the parts of objects by non-negative matrix factorization. *Nature* 401:788–791.
14. Robinson DA (1964) The mechanics of human saccadic eye movement. *J Physiol* 174: 245–264.
15. Williams RJ (1992) Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Mach Learn* 8:229–256.
16. Foster DJ, Morris RG, Dayan P (2000) A model of hippocampally dependent navigation, using the temporal difference learning rule. *Hippocampus* 10:1–16.
17. Ridley RM (1994) The psychology of perserverative and stereotyped behaviour. *Prog Neurobiol* 44:221–231.
18. Graybiel AM (2008) Habits, rituals, and the evaluative brain. *Annu Rev Neurosci* 31: 359–387.