Radboud Repository

Radboud University Nijmegen

PDF hosted at the Radboud Repository of the Radboud University Nijmegen

The following full text is a publisher's version.

For additional information about this publication click this link.
http://hdl.handle.net/2066/166185

Please be advised that this information was generated on 2017-09-04 and may be subject to change.

Human Reinforcement Learning of Sequential Action

George Kachergis[2], **Floris Berends**[1], **Roy de Kleijn**[1], **& Bernhard Hommel**[1]
[1]Institute of Psychology / Leiden Institute for Brain & Cognition, Leiden University
Leiden, the Netherlands
[2]george.kachergis@nyu.edu
Psychology Department
New York University

Abstract

Learning sequential actions is an essential human ability, for most daily activities are sequential. We modify the serial reaction time (SRT) task, originally used to teach people a consistent sequence of button presses by cueing them with the next target response, to record mouse movements, collecting continuous response trajectories. Further, we introduce a reinforcement learning version of the paradigm in which the next target is not cued. Instead, learners must explore response alternatives, and receive a penalty for each incorrect response, as well as a reward for a correct response. Participants are not told that they are to learn a single deterministic sequence of responses, nor that it will repeat (nor how often), nor how long it is. Given the difficulty of the task, it is unsurprising that some learners performed poorly. However, many learners performed remarkably well, and some acquired the full 10-item sequence within 10 repetitions. We compare the high- and low-performers' detailed results in this reinforcement learning (RL) task with a cued trajectory SRT task, finding both similarities and discrepancies. Finally, we note that humans in this task outperform three standard RL models and have different patterns of errors that suggest future modeling directions.

Keywords: Sequence learning; serial reaction time task; sequential action; reinforcement learning; movement trajectory

Introduction

Traditionally, the bulk of cognitive psychology studies deal with single stimulus-response actions or decisions. However, much of human behavior can better be described as sequential action, consisting of partially-ordered hierarchies of simple actions–from cooking and cleaning to speaking and sports. More recently, several diverse lines of research have considered sequential action learning, stemming from various domains including linguistics (Elman, 1990; Saffran, Newport, & Aslin, 1996), implicit learning (Nissen & Bullemer, 1987; Cleeremans & McClelland, 1991), and everyday actions (Cooper & Shallice, 2000; Botvinick & Plaut, 2004).

Throughout these lines of research, a recurring topic of interest has been the process by which action-sequences are acquired, stored, and executed. Early work by Nissen and Bullemer (1987) focused on the modulatory role of attention, and introduced the Serial Reaction Time (SRT) task. Unbeknownst to the participants, the task utilized a 10-symbol sequence, with each successive symbol indicating which of four corresponding keys was to be pressed next. Participants showed improved reaction times across training, although the knowledge gains seemed to be implicit: when explicitly asked at the end of the experiment, they were unable to reproduce the sequence. Nissen and Bullemer (1987) concluded that attention is critical in developing awareness of learned behavior. The role of attention in the SRT task was further studied in Fu, Fu, and Dienes (2008), finding that reward motivation can improve the development of awareness of the sequence. Fu et al. (2008) reasoned reward motivation regulates the amount of attention paid to the stimuli, which in turn facilitates sequence learning.

Cleeremans and McClelland (1991) demonstrated that associative processes could account for the improvement in performance. They adapted the SRT paradigm to include a sequence derived from a 'noisy' finite-state grammar, and showed that the presence of grammatical structure facilitates sequence acquisition. Cleeremans and McClelland (1991) explained their findings with a simulation of the learning process. A Simple Recurrent Network (SRN; Elman (1990)) was able to produce results similar to human performance, findings that were later confirmed by Boyer, Destrebecqz, and Cleeremans (2005). The SRN demonstrates that associative processes are sensitive to the statistical structure of the training material, implying that rule-like behavior can emerge from networks trained on structured sequences. Indeed, Botvinick and Plaut (2004) showed that the SRN is capable of producing everyday hierarchical actions such as coffee- and tea-making after training on a set of valid examples that varied in order and complexity (e.g., sugar then milk or vice-versa, or only milk). Motivated by a study (Stadler, 1995) showing that introducing a longer delay (2000 ms instead of 400 ms) to a random selection of the response-stimulus intervals reduced sequence learning, Dominey (1998) proposed another recurrent network model that is able to account for both serial structure effects and temporal structure (i.e., patterns of delays).

To better reveal the mechanisms behind human sequential action learning, more information is needed than just the speedup of keypresses across an experiment. To measure real-time dynamics and uncertainty during learning, the SRT task was adapted to require and record movement trajectories in (Kachergis, Berends, de Kleijn, & Hommel, 2014b, 2014a). That is, instead of measuring discrete button-presses, continuous recordings were made during a mouse tracking task that replaced the original SRT's buttons with locations on a computer screen. The trajectory SRT paradigm not only replicated earlier findings, but also found evidence of sequential context effects (e.g., predictive movements towards the next response location), and unveiled changes in the movement dynamics (e.g., pre- and post-stimulus onset).

Paradigms such as artificial language learning tasks and the

SRT task have demonstrated that sequence learning is dependent upon the statistical structure within the training material, and attention towards the stimuli which is modulated by reward motivation. However, rewards do not only provide a source of motivation: in many situations they are an irreplaceable source of information–perhaps even the only feedback. Contrary to the SRT task, everyday human action learning is often not characterized by cued responses but by exploring the environment and learning which actions result in positive effects, and which result in negative effects. We believe that, in order to investigate human sequence learning in an ecologically valid manner, it is necessary to draw in another line of research: reinforcement learning (RL), a well-established paradigm in the field of machine learning (Sutton & Barto, 1998), which of course was originally motivated by much earlier behaviorist stimulus-response learning studies (e.g., Skinner (1950). RL paradigms allow learning agents to interact with a task solely through observations, actions, and rewards. The rewards validate the actions, without the need for explicit cueing or other forms of instruction. Thus, learning is exploratory, and accomplished via trial-and-error.

Although much reinforcement learning research is conducted in computer simulation, the inspiration for the approach and many algorithms is in fact rooted in animal behavior (Sutton & Barto, 1998) and there is evidence that similar processes play a role in human learning. For instance, the error-related negativity (ERN) event-related potential (Falkenstein, Hohnsbein, Hoormann, & Blanke, 1991; Gehring, 1992) has been studied extensively as a component of error processing. The ERN originates in the brain whenever task-relevant errors are committed. Holroyd and Coles (2002) links the ERN to the mesencephalic dopamine system, and proposes it is the result of a negative reinforcement signal which it conveys to the anterior cingulate cortex.

The current study adapts the trajectory SRT task to allow for free movement and limited instruction, allowing learners to explore and learn from trial-and-error. This RL sequence learning paradigm allows us to study the effect of rewards on sequence acquisition in more detail, yielding not only correct response times but also mistakes over time, which may be indicators of distinct mechanisms. For example, committing an error indicates incomplete knowledge (or a lapse in memory), whereas RT may be correlated with overall certainty or fluency at that point in the action sequence. Thus, we investigate the RL paradigm data both in terms of earlier trajectory SRT data and in comparison to three standard RL models.

Experiment

The goal of the current study is to examine sequence learning within the trajectory SRT paradigm, and to compare human performance to basic baseline reinforcement learning models. The trajectory SRT task was adapted to no longer cue participants with the next target position, forcing them to instead explore the response alternatives until the correct one was found. Moving the mouse cursor from the previous target to another response alternative resulted in a reward (+1) or penalty (-1) that was accumulated throughout the experiment and displayed continuously. Upon reaching a valid target, it would change color to green, add to the score by +1, and allow the participant to continue exploring. Reaching for an invalid target caused it to change to red, subtract from the score by 1, while the cursor was relocated to the previously occupied target, effectively resetting the participant's progress. Target validity was determined by a recurrent sequence, taken from the Nissen and Bullemer (1987) study, and adapted to fit the trajectory SRT paradigm. Designating the stimuli as numbers from left to right, top to bottom, the sequence read 4-2-3-1-3-2-4-3-2-1.

Methods

Participants Participants in this experiment were 13 Leiden University students and employees (age: M = 23.9, sd = 6.4) who participated in exchange for 3.5 euros or for course credit.

Procedure Participants were instructed that they would be presented with four target squares in the corners of the screen which they were to explore by moving the mouse, each time resulting in either a gain or loss of one point. Participants were told to try to maximize their score, which was displayed continuously at the top of the screen. Unbeknownst to the participants, only one of the four targets would be valid at any given moment, but all were colored blue, so the target could not be visually distinguished. Upon reaching a valid target, its color would change to green momentarily and the score would increase by one. The participant would be able to continue exploring for the next target. Arriving at an invalid target caused it to change to red momentarily and the score was decreased by one, while the cursor was relocated to the previously occupied target. Thus, although there were no instructions explicitly indicating it, participants likely inferred that they had chosen the incorrect stimulus, and should choose one of the remaining two–if they also assumed the same target was never repeated immediately, which was true. In the absence of a previous target (i.e., at the beginning of the experiment or after a rest break) the cursor was moved back to the middle of the screen.

Unbeknownst to the participants, each trial consisted of a series of 10 targets (labeled 1-4 left-to-right and top-to-bottom: 4-2-3-1-3-2-4-3-2-1) that repeated continuously, with no indication where one trial stopped and the next began. Participants completed eight blocks of 10 such trials, with a short rest break after every 2 blocks (i.e., 200 correct movements). A participant who somehow knew the sequence before entering the experiment and never made a mistake would therefore make 800 movements to valid targets, receiving a theoretical maximum of 800 points. At worst, a participant with no memory of even the previous target they had tried may make an infinite number of mistakes, and may never finish the experiment. Assuming enough memory to not repeat the same invalid target more than once when seeking each target (i.e., an elimination strategy), a participant using this elimination strategy would expect on average to score 0

points, as the expected value of completing one movement successfully is 0.[1] Note that participants were not told that there was a single deterministic sequence, let alone details such as how long the sequence was.

Results

The data from all 13 participants were analyzed. Figure 1 shows a histogram of the final score achieved by each participant. The distribution of scores is non-normal (Shapiro-Wilk's $W = 0.87$, $p < .05$), instead looking bimodal, with four participants collecting less than 300 points and all but one of the rest accumulating more than 500 points each. Given the bimodal score distribution, a median split was used to divide the participants into high-performing ($\geq$ 526; 7 people) and low-performing ($<$ 526; 6 people) groups. In the high-scoring group, participants achieved almost flawless performance after only approximately 30 trials, with a final mean score of 652 (max: 725), while the low-scoring group only gradually increased their score (final mean score: 287). The remaining analyses are carried out for each group in an attempt to understand the great variability in performance—and the impressive success of the high-scoring group.

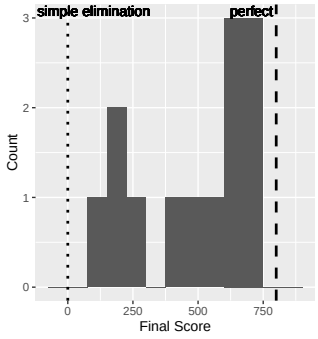

Figure 1: The histogram of participants' final scores after completing 80 sequence repetitions (800 targets) shows a bimodal distribution (lines: elimination strategy EV=0; perfect knowledge EV=800).

Response Times The overall median response time (RT) for all stimulus arrivals was 1,401 ms (sd: 4,980). Of 10,400 correct target arrival times (median: 1,078 ms, sd: 2,216), 317 (3%) were trimmed for being too slow (median + 2 · sd). Of the 4,117 incorrect stimulus arrival times (median: 2,397 ms, sd: 8,401), 100 were trimmed for being too slow (2.4%). Each subject's median RT for correct and incorrect movements was computed for each 10-trial block. Figure 2 shows the mean of subjects' median correct and incorrect RTs over the experiment, split into high- and low-performing group. RTs for correct movements improve in both groups during the first few blocks, but the high-scoring group speeds up more than the low-scoring group. Figure 2 also shows that the rare incorrect RTs for the high-performing group get slower over the course of the experiment, whereas the low-performing group's incorrect RTs only increase a bit. The strikingly slow mistakes of high-performing participants, compared to mistakes that are barely slower than correct movements for the low performers may indicate a different mode of behavior. A possible explanation is that low performers are simply not trying to learn a sequence, or do not expect it to to be deterministic, whereas high performers explicitly learn the sequence, and when they are uncertain they must pause to try to recall the next target.

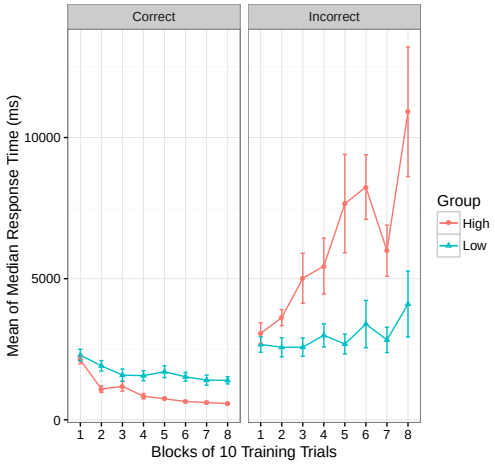

Figure 2: The mean of subjects' median correct RTs by block (left panel) shows that high-performers' RTs improved more than the low-performers' RTs over training. The mean of subjects' median incorrect RTs by block (right panel) shows that the high-performing group's incorrect RTs actually increased, whereas the low-performing group's stayed roughly the same across the experiment. Error bars show +/-1SE.

Accuracy The mean number of mistakes made over the entire experiment was 19.8 (sd: 21.3) for the high-scoring group, and 63.5 (sd: 11.9) for the low-scoring group. Over time, the number of mistakes decreased especially for the high scoring group. Examining the mistakes made by each group of participants according to where they were in the sequence revealed that for both groups the fifth stimulus was particularly challenging. This is reflected in the mean number of mistakes for each group (see Figure 4, as well as in the mean RT to the target by sequence position (see Figure 3).

Comparison to previous research The pattern we observe in the accuracy and response time data bears some resemblance to the pattern observed in a previous trajectory SRT study that used cues (Kachergis et al., 2014b). Although the task in this study (RL) was fundamentally different from the previous study (cued SRT), the same sequence was used in both experiments which enables us to compare the scaled response times from the former and the accuracy from high- and low-performers in the latter experiment. Shown in Figure 5, we see a similar pattern across experiments, and the overall mistakes per position in the RL experiment and the correct RTs in the cued experiment are significantly correlated ($r = 0.64$, t(8)=2.36, $p < .05$), with detailed comparisons below. We also compared the data to the Simple Condensator Model (SCM) proposed by Boyer et al. (2005), which implements a negative recency bias: expectation (activation) for every response builds at each step until a given response occurs,

[1] 33% of chance success in one try (+1), 33% chance of success in two tries (-1+1), and 33% chance of success in three tries (-1-1+1).

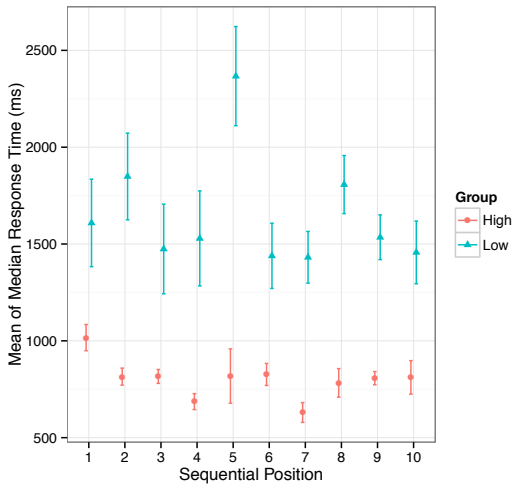

Figure 3: Mean of subjects' median correct response times by median split and sequential position. The correct RTs for the two performance groups were not significantly correlated ($r = 0.17$, t(8)=0.48, $p = 0.65$). Low-scorers were slowest at position 5, followed by 2 and 8, whereas high-scorers were worst at position 1, and almost consistently fast besides that. Error bars reflect +/-1SE.

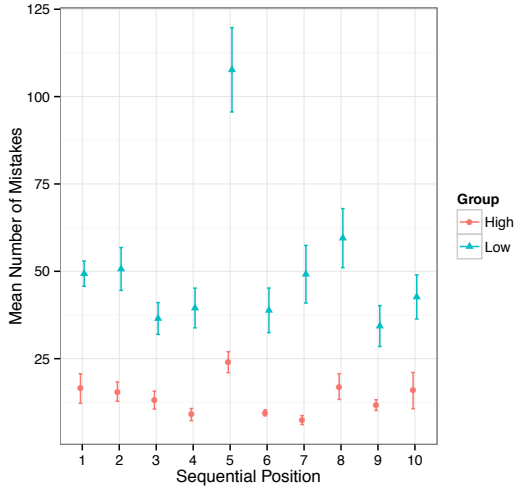

Figure 4: The mean number of mistakes made per block by position in the sequence split by performance group. The errors are highly correlated ($r = .79$, t(8)=3.68, $p < .01$), though note how much worse sequence position 5 was for the low-performing group relative to the next-worst position (8). Low-performers showed twice as many errors in position 5 as in 8, while the high-performing group showed only a 25% increase in errors. Error bars reflect +/-1SE.

and activation resets. Thus, stimuli that have been used least recently have highest activation—and fastest RTs. The SCM has previously been shown to correspond closely to human cued SRT responses, and Figure 5 shows it mirrors mistakes in both groups of the RL experiment quite well.

We examined mistakes and correct response times by their sequential position, and compared these to RTs from the previous cued SRT study. Overall, there is a significant correlation ($r = .88$, t(8)=5.37, $p < .001$) between correct RTs from the RL study and RTs from the cued SRT study. Comparing the cued RTs to the high- and low-scoring groups separately revealed a difference between the groups. The cued SRT RTs do not correlate significantly with the high-scoring group's RTs ($r = .51$, t(8)=1.68, $p = .13$), but do correlate significantly with the number of mistakes made in the RL study ($r = .83$, t(8)=4.18, $p < .01$). The low-scoring group shows the opposite pattern. The cued SRT RTs correlated significantly with the RL correct RTs ($r = .80$, t(8)=3.79, $p < .01$) but not with the RL mistakes ($r = .57$, t(8)=1.96, $p = .09$). Comparing the two groups with each other revealed a significant correlation in mistakes ($r = .79$, t(8)=3.68, $p < .01$), but no significant correlation in RT ($r = .17$, t(8)=0.48, $p > .05$).

The SCM was strongly correlated with the low-scoring RL group's RTs ($r = .93$, $t(8) = 6.93$, $p < .001$), but not with the high-scoring RL group's RTs ($r = .16$, $t(8) = 0.45$, $p = .67$). The low-scorers, failing to discern the repeating sequence, may have mostly used a negative recency bias. The SCM was correlated with the mistakes of both RL groups (low: $r = .89$, $t(8) = 5.52$, $p < .001$; high: $r = .74$, $t(8) = 3.12$, $p = .01$).

Models of sequence learning To compare human sequence acquisition with existing reinforcement learning models, we implemented the models using PyBrain (Schaul et al., 2010; see Figure 6 for an overview of the modeling experiment's

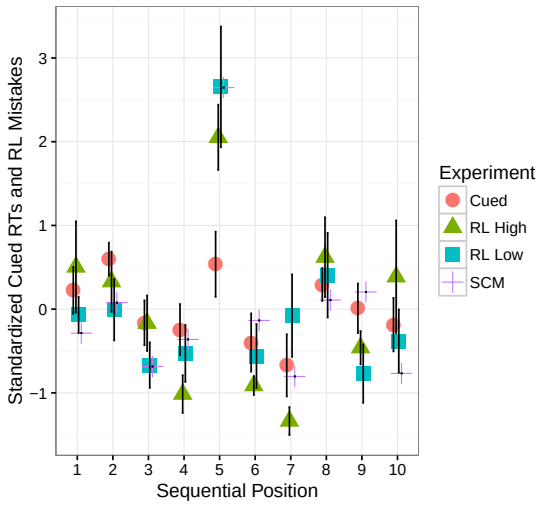

Figure 5: Scaled mean number of mistakes in the current experiment against scaled correct RTs from a cued SRT study Kachergis et al. (2014b) by sequence position. Error bars show +/-1SE.

setup). The environment contains all data regarding the targets, which it passes to the task, which in turn passes the current state of the environment to the agent, which selects the relevant action. The action is evaluated by the environment, which updates itself and passes a reward to the agent. The reward is used to update the agent's strategy, and the model continues with the next step. We defined the reinforcement learning SRT task in this framework for our simulations.

As in the human experiment, the data regarding the targets was only partially-visible to the agent. The task acted as a veil through which a certain state would be observable. To a human participant, the current position in the sequence would

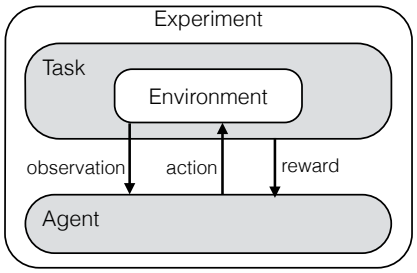

Figure 6: Overview of the experimental setup for the RL models. Each plated component is a PyBrain class, which interact with each other according to the arrows to simulate the same trial-and-error learning process that humans undergo.

be obvious, as it was colored differently from the other stimuli. At a minimum, the immediately prior occupied position was probably obvious as well, readily available in memory. Positions preceding that, however, might not be reliably accessible in memory. In the sequence we used (4231324321), following (Nissen & Bullemer, 1987), each position's identity is fully determined by the previous two positions. That is, one could perfectly predict the next position given only the two prior to it–assuming one has determined that there is a deterministic, periodically-repeating sequence. The RL models we use rely on a set of third-order observations, assuming that the models know their current position and the two prior positions.

The models differ in their learning component, which is contained within the agent and maintains a mapping between states and action-values. For each given input-state there are three action-values, corresponding to the number of movements that can be made by the agent. Upon receiving a reward, the agent updates the action-values using its learning algorithm. We tested three learning algorithms: SARSA (Rummery & Niranjan, 1994), standard Q-learning, and Q(λ)–Q-learning with eligibility traces (Watkins, 1989). Q-learning is an off-policy algorithm, learning about the greedy policy, updating old action-values using the maximum of all action-values for the current state, while it stochastically selects actions, sometimes exploring. SARSA is on-policy: instead of the maximum, it also takes into account the action it has selected for the current state. The eligibility traces in Q(λ) are temporary records of an event (e.g., an action or state) that help with temporal credit assignment by adding a trace to events that are eligible for learning updates. Theoretically, eligibility traces link RL temporal difference methods (like Q-learning and SARSA) to Monte Carlo methods.

These algorithms were chosen as simple baselines that differ somewhat in exploratory behavior and learning speed, and thus may be suitable to compare to human behavior which varied widely. As with the human participants, the simulated SARSA and Q-learners were tasked with iterating over the repeated sequence until the successful completion of 800 movements. For each model, a grid search over the parameters (learning rate α and discounting factor for future rewards γ) was used to find optimal values.

The best parameters found for the SARSA model ($\alpha = .01$, $\gamma = .98$) achieved a mean final score of 200 (sd=218). The best parameters found for Q-learning ($\alpha = .38$, $\gamma = .98$) yielded a mean final score of 290 (sd=116), while Q(λ) reached a mean final score of 451 (sd=34, parameters: $\alpha = .001$, $\gamma = .95$, $\lambda = .99$). However, despite considerable learning by the end of the experiment, none of the models performed as well as the high-performing human learners, who averaged a final score of 652. Even the *maximum* scores achieved by the models were below the high-scoring humans average or maximum (human= 725; Q-learning= 518, Q(λ)=557; SARSA= 546).

Although these common RL models were unable to reach human-level performance, we thought it worthwhile to examine whether their error patterns resemble those of people. The mistakes made by the SARSA and Q-learning algorithms did not vary much by sequence position, and while Q(λ) made more mistakes in the middle of the sequence (vaguely like humans), none of the models' error patterns were significantly correlated with humans.

Discussion

We adapted the trajectory SRT paradigm to be a reinforcement learning task. The task proved to be more challenging for some than for others, as indicated by a bimodal distribution of scores, and differences in the high- and low-performing groups' response times (RTs) and mistakes by sequence position. These data may suggest that participants adopt different strategies, discussed in greater detail below. Overall, the findings of the reinforcement learning paradigm are similar to a previous trajectory SRT experiment with cued targets: RT and accuracy were correlated across experiments. In particular, data from the high-performing participants compared remarkably well to the previous trajectory SRT study, despite the task differences. The most notable similarity was the difficulty participants experienced with the fifth stimulus position.

The better-performing half of participants made very few errors after as little as 10 repetitions of the length 10 sequence. Block-by-block analysis of the RTs by performance group showed a difference in speed-up across the experiment: the high-performing group already made faster correct responses in the second block of ten repetitions, and maintained this advantage. The difference in response times to incorrect targets suggests the two groups might have used different strategies. The rare but increasingly-slow mistakes in the high-performing group may indicate retrieval attempts and an awareness of their uncertainty as to the next step. In contrast, the persistent and relatively fast mistakes of the low-performing group suggest these participants may have adopted a probabilistic view of the task, randomly trying options instead of trying to learn a deterministic pattern. This was corroborated by the particular success of the parameter-free Simple Condensator Model (SCM) in matching the low-scoring group's RTs using a simple negative recency bias. A

distinction of stimulus-based and plan-based control (Tubau, Hommel, & López-Moliner, 2007) may capture the apparent differences between the low- and high-performing groups.

The results by sequence position showed that participants in both groups had more trouble with the target in sequence position five than any other target. This is similar to the pattern observed in previous studies (Nissen & Bullemer, 1987; Kachergis et al., 2014b), and has previously been taken to indicate that participants chunk the sequence into two parts: the initial 4-2-3-1 and the final 4-3-2-1, with positions 5 and 6 bridging the two chunks. We note that the only repeated transition in the sequence (3-2, at positions 6 and 9)–which might be expected to be worse due to the higher transition probability–shows neither slow correct responses nor more mistakes for either position it occurs in, somewhat unlike (Nissen & Bullemer, 1987). Models make the clearest predictions, but in the absence of a winning theory, a comparison to other paradigms can also be illuminating.

Despite the major difference of no cueing of the next response, performance in the RL experiment was quite comparable to performance in the cued SRT study. The correlations of mistakes and RTs by sequence position indicated a difference between the low- and high-performing groups that was not immediately obvious. Overall, the cued SRT response times are correlated to RTs and accuracy in the RL experiment, whereas this is not true for both the low- and high-performing groups separately. The low-scoring group closely matched the negative recency bias of the SCM (in mistakes, but especially in RT), but the pattern and strategy of the high-scoring group is less clear.

Hoping to better understand especially the high-performing group, we developed a simple reinforcement learning model and tested three different learning algorithms. High-performing humans were still far better than the models, which on average scored roughly as well as the low-performing humans. SARSA had quite variable performance, but was lowest on average, while Q-learning with eligibility traces fared the best. Examining the models' performance by sequence position showed they did not correspond well with errors in either group of humans. This suggests that simple model-free reinforcement algorithms do not capture the process by which humans learn action sequences, even though they eventually converge on a proper solution. One explanation for this is the fact that the task and models used in studies like this do not fully capture the essence of human action learning, which is goal-directed by nature. Future studies could shed light on the role of goals in the acquisition of such action sequences, as has been shown to exist for single-step action (see, for example, Hommel, Müsseler, Aschersleben, and Prinz (2001) for one proposed mechanism of goal-directed action). The process by which humans acquire action sequences is subtle, can yield quite variable performance, and is not easily captured by simple learning algorithms. However, studying it is important, as most of human behavior is essentially sequential in nature.

Acknowledgments

The preparation of this work was supported by the European Commission (EU Cognitive Systems project ROBOHOW.COG; FP7-ICT-2011).

References

Botvinick, M., & Plaut, D. C. (2004). Doing without schema hierarchies: A recurrent connectionist approach to routine sequential action and its pathologies. *Psychological Review*, *111*, 395–429.

Boyer, M., Destrebecqz, A., & Cleeremans, A. (2005). Processing abstract sequence structure: Learning without knowing, or knowing without learning? *Psychological Research*, *69*, 383–398.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*, 235–253.

Cooper, R. P., & Shallice, T. (2000). Contention scheduling and the control contention scheduling and the control of routine activities. *Cognitive Neuropsychology*, *17*(4), 2987–338.

Dominey, P. F. (1998). A shared system for learning serial and temporal sturcture of sensori-motor sequences? evidence from simulation and human experiments. *Cognitive Brain Research*, *6*(163), 163–172.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Falkenstein, M., Hohnsbein, J., Hoormann, J., & Blanke, L. (1991). Effects of crossmodal divided attention on late {ERP} components. ii. error processing in choice reaction tasks. *Electroencephalography and Clinical Neurophysiology*, *78*(6), 447–455.

Fu, Q., Fu, X., & Dienes, Z. (2008). Implicit sequence learning and conscious awareness. *Consciousness and Cognition*, *17*, 185–202.

Gehring, W. J. (1992). *The error-related negativity: Evidence for a neural mechanism for error-related processing* (Unpublished doctoral dissertation). University of Illinois at Urbana-Champaign.

Holroyd, C. B., & Coles, M. G. (2002). The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity. *Psychological Review*, *109*(4), 679 - 709.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences*, *24*, 849–937.

Kachergis, G., Berends, F., de Kleijn, R., & Hommel, B. (2014a). Reward effects on sequential action learning in a trajectory serial reaction time task. *IEEE Conference on Development and Learning / EpiRob 2014*.

Kachergis, G., Berends, F., de Kleijn, R., & Hommel, B. (2014b). Trajectory effects in a novel serial reaction time task. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*.

Nissen, M. J., & Bullemer, P. (1987). Attentional requirements of learning: evidence from performance measures. *Cognitive Psychology*, *19*, 1–32.

Rummery, G. A., & Niranjan, M. (1994). *On-line q-learning using connectionist systems* (Tech. Rep. No. CUED/F-INFENG/TR 166). Cambridge University.

Saffran, J., Newport, E., & Aslin, R. (1996). Word Segmentation: The Role of Distributional Cues. *Journal of Memory and Language*.

Schaul, T., Bayer, J., Wierstra, D., Sun, Y., Felder, M., Sehnke, F., ... Schmidhuber, J. (2010). PyBrain. *Journal of Machine Learning Research*, *11*, 743–746.

Skinner, B. F. (1950). Are theories of learning necessary? *Psychological Review*, *57*(4), 193–216.

Stadler, M. A. (1995). The role of attention in implicity learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, *21*, 674–685.

Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.

Tubau, E., Hommel, B., & López-Moliner, J. (2007). Modes of executive control in sequence learning: From stimulus-based to plan-based control. *Journal of Experimental Psychology: General*, *136*, 43–63.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards* (Unpublished doctoral dissertation). Cambridge University.