# Control of complex actions in humans and robots

Kleijn, R.E. de

CHAPTER 4

# Predictive movements and human reinforcement learning of sequential action

Most daily human behaviors can be seen as learned sequential actions: from walking, cooking, and cleaning to speaking and writing. Consequently, sequence learning has been studied in different contexts ranging from implicit sequence learning [19, 29, 107, 149] to language acquisition [41, 136], typing [46, 52], and manual everyday actions [18, 32]. In implicit learning research, an important paradigm has been the *serial reaction time* (SRT) task, which requires participants to press one of four buttons when cued by a corresponding light, in a sequence that repeats—unbeknownst to learners—every 10 presses [107]. Subjects trained on this repeating sequence developed faster reaction times (RTs) over the course of training, as compared to a control group responding to a random sequence of stimuli. The SRT paradigm has been cited as evidence for implicit learning, as subjects experiencing the re-

This chapter is an adaptation of the article *de Kleijn, R., Kachergis, G., & Hommel, B. (under revision). Predictive movements and human reinforcement learning of sequential action.*

peating sequence, despite showing faster RTs over time, report no explicit knowledge of the sequence when debriefed afterwards. However, performance does suffer somewhat when participants must simultaneously perform a second task [107], suggesting that learning in the SRT task does require some attentional resources or effort. The role of attention in the SRT task was further studied by Fu et al. [50], who demonstrated that reward motivation can improve the development of awareness of the sequence. They reasoned that reward motivation regulates the amount of attention paid towards the stimuli, which in turn facilitates sequence learning. Additionally, Willingham et al. [170] found that some participants achieved a degree of declarative knowledge after a fixed training period in the SRT task, and that additional training resulted in more explicit knowledge for many subjects, if not all. On balance, it seems that the SRT task is neither wholly implicit nor wholly explicit.

The dissociation of implicit and explicit processes facilitating sequence learning remains a topic of debate, yet learning remains robust under high degrees of noise and complex structure in the sequences [29]. Complex action sequences are not mere stimulus–response chains, but rather require representing sequential context in order to learn [87]. Moreover, human behavior is often thought of as *predictive*—indeed, many models of sequential learning operate on a prediction-based error signal [18, 76]. As such, it is problematic that the discrete button presses in the SRT paradigm cannot distinguish an *anticipatory* response due to correctly predicting the stimulus (or a slow response due to an incorrect prediction) from *reactive* (though perhaps pre-potentiated) responses based on the cue. Truly predictive responses—that is, those made in the interstimulus interval before the next response is cued—are not valid responses in the SRT paradigm.

In this paper we introduce two modifications of the SRT paradigm that allow us to naturally investigate both predictive and reactive responding in human sequence learning. In Experiment 1, recognizing that actions are continuous movements that can reveal the underlying dynamics of

the cognitive processes driving them [147], we used a mouse-tracking adaptation of the SRT task in which spatial locations are both stimuli and response options [74, 75]. By tracking their movement before and after the next target is cued, we investigated changes in predictive versus cued responding over the course of the experiment [160]. Using this trajectory SRT paradigm, we replicated the overall Nissen and Bullemer [107] RT results, and moreover show sequential context effects—predictive bends in response trajectories—along with different movement dynamics pre- and post-cue.

In many implicit learning tasks such as artificial language learning and the SRT paradigm, learning is dependent on recognizing some statistically reliable sequential structure in stimuli not under the learner's control. However, everyday human action learning is often not characterized by processing a steady stream of stimuli, but by exploring the environment (i.e. choosing actions) and receiving positive and negative feedback. Prediction is thus an essential element of reinforcement learning (RL), which is a well-established paradigm in the field of machine learning [153] that was originally motivated by much earlier behaviorist stimulus–response learning studies [144]. 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. In Experiment 2, we further modified the trajectory SRT paradigm by not cueing responses at all: participants had to explore response alternatives until the correct one was found, receiving feedback (negative or positive points) at each response. We investigated sequence learning in this RL SRT paradigm that required prediction rather than reaction, and found correspondences between successful learners in this paradigm and in the reactive SRT paradigm in Experiment 1. Using the RL paradigm allowed us to study the effect of rewards on sequence acquisition in more detail, yielding not only response times but also errors over time. Thus, the current study adapted the trajectory SRT task to allow for free movement and limited instruction, allowing

learners to explore and learn from trial-and-error.

In addition, we attempted to capture human performance and error patterns using reinforcement learning models. Due to the relatively simple nature of the task, we investigated if simple (i.e. model-free) RL models were sufficient to learn the repeating sequence by trial-and-error. We assessed the RL data both in terms of earlier SRT data and in comparison to three standard RL models. Overall, this study provides insights into prediction error-driven learning of sequential action learning.

## 4.2 Experiment 1

The purpose of the first experiment was to replicate earlier findings by Nissen and Bullemer [107] using the trajectory SRT paradigm. This study used four stimuli in a recurring sequence of length 10, horizontally displayed on a screen. Designating the stimulus positions from left to right as numbers, the original sequence read 4–2–3–1–3–2–4–3–2–1. To fit the trajectory paradigm the sequence was mapped to a square, left-to-right and top-to-bottom (i.e. 1 = top left, 2 = top right, 3 = bottom left, and 4 = bottom right). Participants moved the mouse from one stimulus position to the next, corresponding to the sequence. We tested two groups of participants, one trained on the recurring sequence and the other trained on a random sequence. After ten blocks of training participants completed a generating task. This task consisted of the same basic test conditions, except participants were asked to predict the sequence instead of following it.

Nissen and Bullemer [107] originally found that participants showed improved performance within the first block of training. Performance suffered under dual-task conditions and varied as a function of serial position in a pattern suggesting that learners were chunking the sequence into two pieces. In total, the study's results suggest that attention to the sequence is crucial for both implicit and explicit sequence learning, but that improved performance is not critically dependent on awareness of the sequence. For the purpose of Experiment 1 only the initial experi-

ment was replicated. We expected to replicate the basic improvement of performance, as well as the chunking pattern that was observed. Like Willingham et al. [170], we included a final generation task, in which participants were asked to reproduce any action sequence they felt they had learned during training.

### 4.2.1 Methods

#### Participants

Participants in this experiment were 22 Leiden University undergraduate students who participated in exchange for 3.50 euros or course credit.

#### Apparatus and materials

The experiment was performed on a computer with a 21-inch monitor with 60 Hz refresh rate and a resolution of 1024x768 pixels. Participants used a mouse to move the cursor. The experiment was programmed in Python with the PyGame library, and cursor position was sampled at every screen refresh.

#### Procedure

Participants were alternately assigned to one of the two between-subjects conditions according to the order they signed up. In the NB87 sequence condition, participants were given a repeating sequence of 10 locations corresponding to the Nissen and Bullemer [107] sequence (4–2–3–1–3–2–4–3–2–1). In the random sequence condition, participants followed a randomly generated movement sequence without repetitions (i.e. staying at the same location).

Participants were told to quickly and accurately move the mouse cursor to whichever square turned green. After arriving at the highlighted stimulus, another stimulus was highlighted after a 500 ms ISI. Participants completed 80 training trials, each of which contained a series of 10 locations. Participants were given a rest break every 20 training trials. Fol-

lowing the training phase, participants were asked to try to reproduce any sequence they had learned.

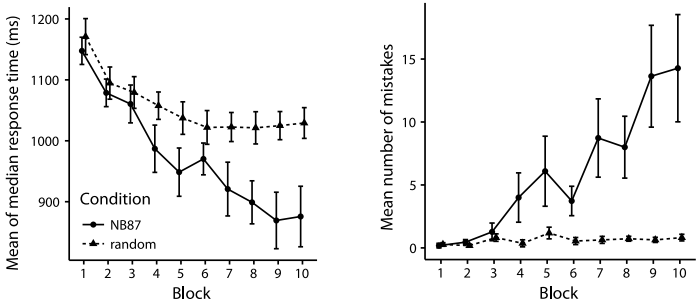
Each block contained a series of 80 location stimuli (i.e. 10 repetitions of the NB87 sequence) which participants had to track with the cursor. The stimulus display consisted of four red squares (location 1 = upper left, 2 = upper right, 3 = lower left, 4 = lower right), displayed continuously. Each stimulus was an 80 × 80 pixel square, separated by 440 pixels of white space. As a participant's cursor arrived at the green square, the square's color would change to red, like the other stimuli. The next target stimulus in the sequence would change color after a 500 ms ISI.

After training, participants were given a generating task similar to the training task. In the generating task, participants were asked to predict where they thought the stimulus would appear and move the mouse to that square. In other words, they were asked to complete the sequence without being cued. A correct prediction would cause no color change while an error would cause the correct continuation of the sequence to appear in green, and participants were to move to the next location.

### 4.2.2 Results

#### Response times

Data were analyzed from the 22 participants (11 per condition) that completed the experiment. Median movement time to a target was 1,040 ms (*SD*: 1,776). Of 17,578 target arrival times, 84 were removed for being slower than 2,816 ms (median + *SD*). Each subject's median RT for correct movements on each block was computed. Figure 4.1a shows the mean of median RTs by block for the two conditions. Participants in both conditions got faster over the course of the experiment, but participants in the NB87 sequence condition improved more than those in the random condition, replicating the Nissen and Bullemer [107] speedup. There was a 25% reduction in reaction time over the course of training. These data were analyzed by a two-way analysis of variance, which indicated significant main effects of condition ($F(1, 20) = 31.3$, $p < .001$) and block ($F(7,$




**(a)** Mean of median RTs by block show that both conditions sped up over the course of Experiment 1, but that NB87 improved more.

**(b)** Mean number of errors by block shows only the NB87 participants made an increasing number of errors.

**Figure 4.1** | Experiment 1 RTs and error rates by block. Error bars show +/–1 SE.

168) = 6.3, $p < .05$), and a significant interaction effect ($F(7, 210) = 14.7$, $p < .01$) between the two.

The accuracy data is shown in Figure 4.1b. Accuracy was high across training blocks although it dropped over time in the NB87 group, particularly after the first three blocks of training. A two-way analysis of variance confirmed a significant main effect of group ($F(1, 20) = 36.7$, $p < .001$) and a significant interaction effect ($F(9, 210) = 14.1$, $p < .001$). These results are evidence of sequence learning, replicating the Nissen and Bullemer [107] keypress-based results. However, there was a speed-accuracy tradeoff in the NB87 condition: both accuracy and RT dropped over time. This was not present in the Nissen and Bullemer [107] results, but can be explained through the difference in response execution. Key-presses are intermittent and can only be made in response to a stimulus (prestimulus responses were not recorded), while mouse movements are continuous and made constantly. Indeed, in the NB87 condition faster median hit RTs on a training block had a significant negative correlation with the number of errors in that block (for the 67 of 110 blocks containing errors; $r = -.56$, $t(65) = -5.48$, $p < .001$), showing a speed-accuracy

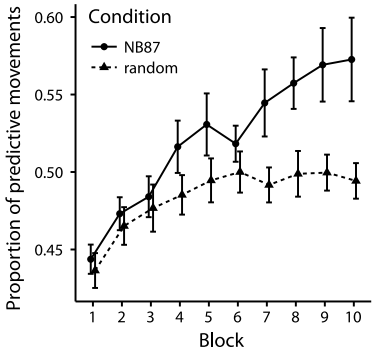



**Figure 4.2 |** Proportion of predictive movements (i.e. movements made during the ITI) by block in each condition. Random condition participants re-center, whereas NB87 participants move towards other stimuli. By block 4, NB87 participants were making more than half of their movement predictively, and continued to move more predictively: up to 57% by the end of the experiment. Error bars show +/–1 SE.

tradeoff. This is likely due to the trajectory SRT paradigm encouraging prediction, allowing participants to move freely while performing the experiment.

Indeed, an analysis of the proportion of distance traveled before arriving at the next target during the 500 ms interval before the cue appeared (i.e. predictive movement), shown in Figure 4.2, shows that participants in the random condition level off at making half of their movement, on average, during the pre-cue interval, whereas by block 10, participants in the NB87 condition predictively completed over 57% of their movement in the 500 ms interval before the next location is highlighted. This shows that participants in the NB87 are predicting the next target location and already moving towards–getting over halfway there–before the next cue appears.

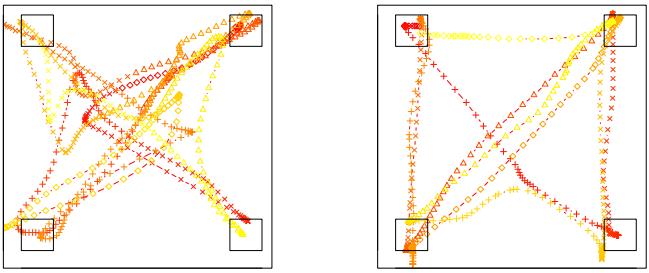
A two-way ANOVA with block as between-subject and serial position as within-subject factors showed significant main effects for block ($F(9, 210) = 32.3$, $p < .001$ and serial position ($F(9, 100) = 10.2$, $p < .01$). To de-

termine whether participants became faster at the entire sequence or rather learned some chunks better than others, mean RT was calculated for each serial position. Similar to the Nissen and Bullemer [107] results, RTs on the second, fifth and eighth serial positions are slow, which may indicate that participants chunk the full sequence into two small, well-learned pieces.

Performance on the generating task was poor, as participants on average did not manage to reproduce the sequence without making many errors ($M = 5.77$ errors). This indicates that, although training performance showed evidence of sequence learning, participants were not explicitly aware of the sequence. It is possible that participants would eventually be able to reproduce the sequence if training were extended, as in Willingham et al. [170]. Nissen and Bullemer [107] originally found that participants were able to score around 80% correct on the generating task after two blocks of ten trials. Although the current study only required participants to complete one block of ten trials during the generating task, participants did not show any improvement during the task.

### Trajectory results

Figure 4.3 shows an example of mouse movements during a characteristic trial from each condition. Participants in the random condition (e.g. Figure 4.3a) tended to re-center the cursor after hitting a target, during the 500 ms ISI. This strategy is not unreasonable under conditions of uncertainty, as it minimizes the distance to potential targets, and the next target cannot be predicted in the random condition. Centering behavior is shown in Figure 4.4a. Centering behavior is defined as the proportion of time spent in the center $100 \times 100$ pixels of the screen between reaching the previous target and current target reached. We deemed the distinction between reactive and predictive movements (as made by Dale et al. [34]) unsuitable for the current analyses due to the random condition used to compare. As the experiment progressed, participants in the random condition adopted a centering strategy that minimized distance to potential targets, while participants in the predictable NB87 condition

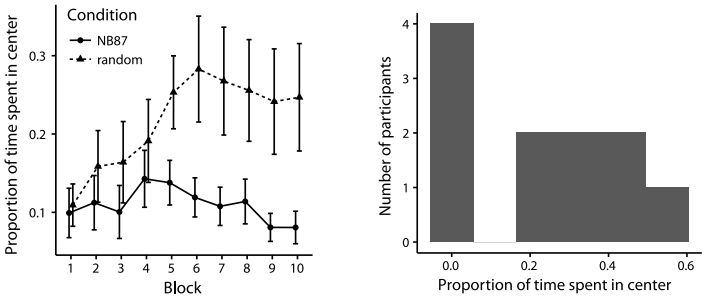


**(a)** A trial from the random condition, in which the next location was chosen at random, without repeats. All 11 random participants adopted a similar strategy of re-centering the cursor after each response. This is optimal in the sense that it was impossible to know which location will be highlighted next.

**(b)** A characteristic trial of a participant's movements during the NB87 sequence, beginning at location 4 (lower right) and ending at location 1 (upper left). These isomorphic trajectories can be compared for context effects. Only 4 NB87 participants showed centering movements in the last half of training.

**Figure 4.3** | Characteristic movements in one trial from the random condition (a) and the NB87 condition (b). $t_0$ = red, $t_{end}$ = yellow.

did not show this behavior. Participants in the random condition spent an increasingly larger proportion of time in the center of the screen compared to NB87 participants, $F(9, 180) = 2.51$, $p = .010$ for the interaction between block and condition. Similar centering behavior has been reported, but not quantified in the current context by Duran and Dale [38], and Dale et al. [34]. Interestingly, not all participants in the random condition displayed this centering strategy, as evidenced by the large standard errors, especially in the final half of the experiment. Instead, participants seemed to employ either a non-centering strategy or a centering strategy in which they spent almost 25% of the ISI in the center of the screen.



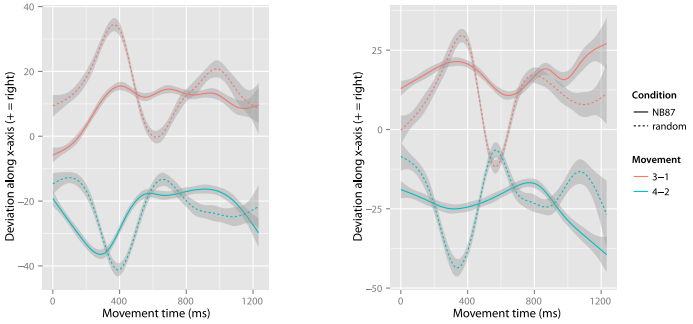

**(a)** Proportion of time spent in the center of the screen, defined as a $100 \times 100$ pixel square in the center of the screen. Centering behavior in the random condition is clearly visible. Error bars show +/–1 SE.

**(b)** Distribution of centering behavior for the last half of the experiment for the random condition. Two groups of participants can be identified: those who center during the ISI and those who do not.

**Figure 4.4** | Centering behavior during the ISI.

With learning, targets are predictable in the NB87 sequence condition, thus participants are expected to show faster reaction times (RTs) as training proceeds.

The NB87 sequence, 4–2–3–1–3–2–4–3–2–1, contains only one identical transition (3–2, a diagonal movement), although other movements are isomorphic (e.g. 4–2 and 3–1). We examined the development of sequential context effects–deflections in response trajectory caused by the prior or subsequent location–by plotting the average trajectories for the isomorphic movements: 4–2 vs. 3–1. In the experiment, these movements are vertical, and we were interested in investigating the average deflections from the direct path from one stimulus center to another. We averaged position across subjects for these movements and plotted their deviation from the direct path (y-axis) over time (x-axis) in Figure 4.5, split by condition, and for each half of training. Early in training, some centering behavior is apparent in both conditions, most notably in the 4–2 movement. This movement also clearly shows the absence of center-




**(a)** Horizontal deviation during movement (i.e. over time) in early training. Both conditions' trajectories show some centering behavior, bending towards the middle (i.e. up for 3–1, down for 4–2). NB87 trajectories show less deviation.

**(b)** Horizontal deviation during movement in late training. The random condition shows more centering behavior, while the NB87 trajectories show little variation except at the end of the movements when they diverge, showing prediction of the subsequent stimulus.

**Figure 4.5** | Averaged trajectories for vertical movements 4–2 and 3–1.

ing behavior late in training for the NB87 condition. The 4–2 movement also shows participants tended to move towards the left after completing the movement. As the next target in the sequence is 3, which is situated to the bottom left of the current target, this indicates they were beginning to move towards the subsequent target. These trajectory analyses corroborate that NB87 participants were making increasingly predictive movements, bending towards the next stimulus position based on their contextual knowledge.

### 4.2.3 Discussion

In summary, Experiment 1 replicated the results from the Nissen and Bullemer [107] serial button-pressing task with a mouse-trajectory version of the task, showing that participants learn regularities in the stim-

ulus stream and exhibit speeded responding, even though they are bad at explicitly reproducing the sequence. We have also demonstrated the advantage of the trajectory-tracking SRT task: because participants can move the mouse cursor during the interstimulus interval—before the next cue has appeared—we can distinguish predictive movements (towards the correct next stimulus) from post-cue speed-ups. Indeed, we found that participants in the NB87 sequence condition made an increasingly large proportion of their movement during the 500 ms pre-cue interval. Also, we found centering behavior similar to Dale et al. [34]. However, in addition to their findings we compared centering behavior between the random and NB87 condition, showing that participants in the random condition show significantly more centering behavior, which can be explained by uncertainty in prediction. Having established that prediction plays a role in the speed-up seen in the SRT-trajectory paradigm, in Experiment 2 we made prediction the essential goal of the task, requiring learners to move to the next location without a cue, and only giving feedback upon making a response.

## 4.3 Experiment 2

The results of Experiment 1 show that spatial sequences can be learned through cued learning, replicating a huge body of literature on the SRT task introduced by Nissen and Bullemer [107]. However, sequence learning in everyday action can hardly be considered cued. Instead, humans are in constant interaction with their environment, exploring it and receiving positive or negative feedback on their taken actions. In Experiment 2, we adapted the paradigm of the trajectory SRT into an exploration paradigm in which participants actively try out the alternative options and receive feedback (reinforcement or punishment). More specifically, the goal of Experiment 2 was to examine reinforcement learning within the trajectory SRT paradigm, and to compare human performance to basic baseline 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 [107] 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.

### 4.3.1 Methods

#### Participants

Participants in this experiment were 13 Leiden University students and employees (aged $M = 23.9$, $SD = 6.4$) who participated in exchange for 3.50 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 then 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 two blocks (i.e. 200 correct movements). A participant who somehow knew the sequence before entering the experiment and never made an error would therefore make 800 movements to valid targets, receiving the 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 errors, 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 (EV) 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.

### 4.3.2 Results

The data from all 13 participants were analyzed. The distribution was 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 ($<$

[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).

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 were carried out for each group in an attempt to understand the great variability in performance—and the impressive success of the high-scoring group.

**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 4.6 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 4.6 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 errors of high-performing participants, compared to errors 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.

**Accuracy**

The mean number of errors 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-

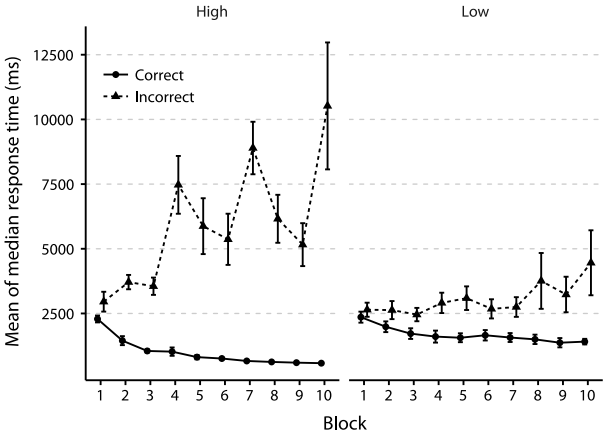
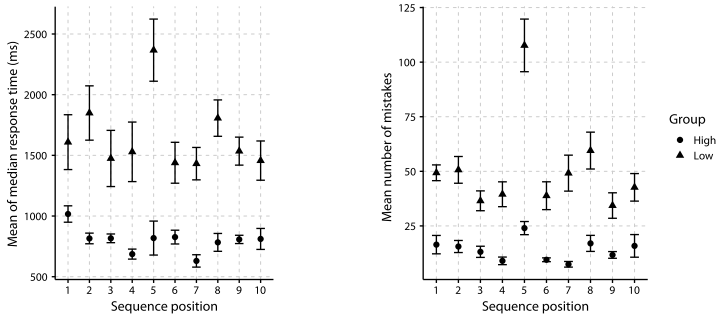



**Figure 4.6 |** The mean of subjects' median correct RTs by block shows that high-performers' (left panel) RTs improved more than the low-performers' (right panel) RTs over training. The mean of subjects' median incorrect RTs by block shows that the high-performing group's incorrect RTs actually increased, whereas the low-performing group's stayed roughly the same across the experiment. Error bars show +/−1 SE.

scoring group. Over time, the number of errors decreased especially for the high scoring group. Examining the errors 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 errors for each group (see Figure 4.7b, as well as in the mean RT to the target by sequence position (see Figure 4.7a).

### Comparison to Experiment 1

The pattern we observe in the accuracy and response time data bears some resemblance to the pattern observed in Experiment 1, despite the use of cues in that experiment. Although the RL SRT task in Experiment 2 was fundamentally different from the cued SRT task in Experiment 2, the same sequence was used in both experiments. We can therefore




**(a)** 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, $r(8) = .17$, $p = .65$.

**(b)** The mean number of errors made at each position in the sequence split by performance group. The errors are highly correlated, $r(8) = .79$, $p < .01$.

**Figure 4.7 |** RTs and error rates by median split and sequential position. 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 +/−1 SE.

compare the scaled response time and accuracy data from the two experiments in Figure 4.8, which shows a similar pattern across experiments.

We examined errors and correct response times by their sequential position, and compared these to RTs from Experiment 1. Overall, there is a significant correlation $r(8) = .88$, $p < .001$, between correct RTs from the RL experiment and RTs from the cued SRT experiment. 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, $r(8) = .51$, $p = .13$, but do correlate significantly with the number of errors made in the RL experiment, $r(8) = .83$, $p < .01$. The low-scoring group shows the opposite pattern. The cued SRT RTs correlated significantly with the RL correct RTs, $r(8) = .80$,

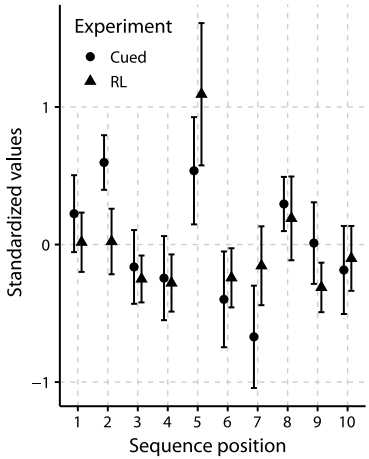



**Figure 4.8 |** Scaled mean number of errors in Experiment 2 (RL) against scaled correct RTs from Experiment 1's cued SRT paradigm (NB87) by sequence position. The number of errors per position and the correct RTs are significantly correlated, $r(8) = .64$, $p < .05$. Error bars show +/−1 SE.

$p < .01$, but not with the RL errors, $r(8) = .57$, $p = .09$. Comparing the two groups with each other revealed a significant correlation in errors, $r(8) = .79$, $p < .01$, but no significant correlation in RT, $r(8) = .17$, $p > .05$.

## 4.4 Models

### Modeling environment

To compare human sequence acquisition with existing reinforcement learning models, we implemented three reinforcement learning models and a simple negative recency biased model (SCM; [19]) using PyBrain [139]. 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

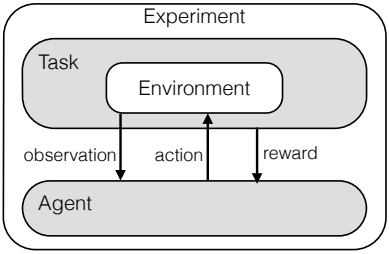



**Figure 4.9 |** Overview of the experimental setup for the reinforcement learning models. Each plated component is a PyBrain class, which interact with each other according to the arrows to simulate the same trial-and-error learning process that humans undergo.

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, see Figure 4.9 for the specific design.

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 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 (4–2–3–1–3–2–4–3–2–1), following Nissen and Bullemer [107], 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.

**On-policy vs. off-policy learners**

The reinforcement learning models differ in their learning component, which is contained within the agent and maintains a mapping between input 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. After receiving a reward, the agent updates the action-values using its learning algorithm. We tested three learning algorithms: SARSA [133], standard Q-learning, and Q($\lambda$)–Q-learning with eligibility traces [168].

Off-policy learners such as Q-learning learn the value of the optimal policy independently of the agent's actions. They learn about the greedy policy, updating old action-values using the maximum of all action-values for the current state, while—depending on the action selection policy—it can stochastically select actions and explore.

The update rule in Q-learning updates Q for any state-action pair $<s, a>$ using an experience tuple $<s, a, s', r>$, with learning rate $\alpha \in [0, 1]$ and discount factor $\gamma \in [0, 1]$:

$$Q'(s, a) = (1 - \alpha)Q(s, a) + \alpha(r + \gamma Q[s', \underset{a'}{\operatorname{argmax}}(Q[s', a'])]) \tag{4.1}$$

In contrast, on-policy learners (e.g. SARSA) learn the value of the policy actually being carried out by the agent: instead of the maximum, they also take into account the action that was selected for the current state. In other words, it does not use the maximum attainable reward in state $s'$ to update the Q-table, but instead chooses $a'$ using the same policy it used to choose $a$. It therefore needs the experience tuple $<s, a, r, s', a'>$:

$$Q'(s, a) = (1 - \alpha)Q(s, a) + \alpha(r + \gamma Q[s', a']) \tag{4.2}$$

The eligibility traces in Q($\lambda$) 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, eli-

gibility traces link RL temporal difference methods (like Q-learning and SARSA) to Monte Carlo methods.

### Simple condensator model

To investigate if perhaps an even more elementary mechanism could be responsible for participants' behavior, we also included a condensator model, introduced by Boyer et al. [19], and inspired by Dominey [37]. In this model, each target is assigned a corresponding unit, with activation ranging from .0 to 1.0. Summed activation across units is always 1.0, and all units were initialized at .25. Each step, the unit with the highest activation is chosen, and its activation is then distributed equally among the other three units.

These reinforcement learning models 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 $\alpha$ and discounting factor for future rewards $\gamma$) was used to find optimal values.

### Modeling results

The best parameters found for the SARSA model ($\alpha = .01$, $\gamma = .98$) achieved a mean final score of 183 ($SD = 292$). The best parameters found for Q-learning ($\alpha = .38$, $\gamma = .98$) yielded a mean final score of 346 ($SD = 75$), while Q($\lambda$) reached a mean final score of 369 ($SD = 53$, 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 = 473, Q($\lambda$) = 440; SARSA = 477).

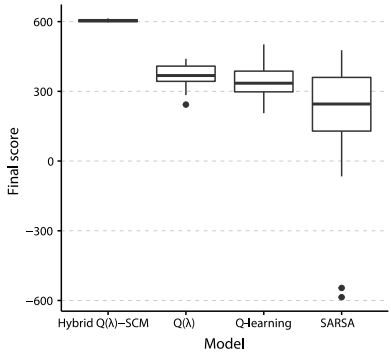



**Figure 4.10 |** RL task scores of the different models tested. A hybrid Q($\lambda$)–SCM model performs better than all of the other RL models, but none of the models reach human performance.

We hypothesized that an RL model combined with a negative recency bias in early learning (with high levels of uncertainty) could perhaps yield better results. Using this technique, humans may be using a recency avoidance strategy in early learning, which would become less necessary after the sequence has been acquired. To investigate, we tested a hybrid model in which the SCM model would choose the next target when certainty (expected action value) of the RL model was low (defined by an optimized parameter: .61). This hybrid Q($\lambda$)–SCM model averaged a final score of 604 ($SD$ = 4). Results from all models are displayed in Figure 4.10.

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 mean number of errors made by each model at each position in the sequence, as was done earlier for humans. The errors made by the SARSA and Q-learning algorithms did not vary much by sequence position. Q($\lambda$) made more errors in the middle of the sequence, but still did not resemble human error patterns.

## 4.5 General discussion

This paper introduced the trajectory serial reaction time task and found that it replicates the results of Experiment 1 of Nissen and Bullemer [107]. Thus, while the trajectory SRT paradigm retains the essence of the original SRT, it also affords the opportunity to measure a variety of more detailed statistics about subjects' continuous motions. Response trajectories can reveal uncertainty, predictive movements, reversals in decision, and other phenomena that may reveal the dynamics of the learning mechanisms at work. The present study examined the average trajectories of two isomorphic vertical movements that appear in the NB87 sequence, as well as in the random condition. The two movements have different subsequent stimuli in the NB87 condition, and were thus expected to show a sequential context effect: as participants learn where the next stimulus will be, they may start to move towards this response even as they finish the previous movement–as a piano player may reach for the next key while the current one is being sustained [145].

We found not only that the expected context effects had developed by late training, but also evidence of possibly strategic adaptive behavior in the random condition. Many participants in the random condition developed a re-centering approach after each response, waiting for the next (unpredictable) stimulus to appear. In a way this behavior is optimal, since the center of the screen is as close as possible to all stimuli. Some participants in both conditions showed this behavior to a limited extent early in training, but those trained on the NB87 sequence lost this behavior over time as they learned to predict the location of the subsequent stimulus–hinted at by the decrease in reaction times in this condition, and confirmed by the deviation in average trajectory towards the subsequent stimuli. Of the participants in the random condition, two groups could be identified: a centering group and a non-centering group. This might reflect differences in strategy similar to Tubau et al.'s [160] stimulus-based vs. plan-based control mode, or Dale et al.'s [34] reactive vs. predictive movements. How these different behavioral strategies are

related could be the focus of future research.

Overall, the behavioral results show a striking similarity to the Nissen and Bullemer [107] results. The pattern of reaction times over sequence position was strikingly similar to the pattern observed in the original study, although the movement reaction times were higher throughout training and participants showed less overall improvement. This can be explained through the mechanics of the paradigm: mouse movements require more time to be executed than single keypresses, and require some fine motor control and error correction. The sensitivity of the mouse can be adjusted to achieve a balance between RT and error; we used a very low sensitivity to reduce overall noise. Participants in the NB87 sequence condition nonetheless showed an increased number of errors during training, indicative of a speed-accuracy trade-off which was not present in the Nissen and Bullemer [107] results. It is possible that extending the training would eventually lead to a reduction of errors, as participants would gradually become aware of the sequence.

In Experiment 2, 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 differences in response times and accuracy. Those data also suggest that participants adopt different strategies, and tried to adapt when they were not learning. These findings are similar to those in Experiment 1: RT and accuracy were correlated across experiments. In particular, data from the high-performing participants compared remarkably well to Experiment 1, despite the task differences. The most notable similarity was the difficulty participants experienced with the fifth stimulus position.

A bimodal distribution of scores showed that half of the participants did really well, as they made very few errors after roughly 10 repetitions of the sequence. Block-by-block analysis of the response times showed a difference in speed-up across the experiment between groups, indicating the high-performing group learned the sequence much better than the low-performing group. The difference in response times to incorrect targets suggests the two groups might have used different strategies. The

rare but increasingly slow errors in the high-performing group suggest more time was spent figuring out the next stimulus, while the persistent and relatively fast errors of the low-performing group suggest participants may have adopted a probabilistic view of the task, randomly trying options instead of trying to learn a deterministic pattern.

Despite the major difference of the absence of cueing of the next response, performance in the RL experiment was quite comparable to performance in the cued SRT experiment. The pattern of correlations 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 data from the RL experiment, whereas this is not true for both the low- and high-performing groups separately. We expect this is due to different strategies among groups, leading to a different pattern of speed and accuracy at different sequence positions.

In addition to our behavioral analyses, we tested three different reinforcement learning models to see if human behavior could be explained by simple, model-free responses to sequential stimuli. 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 human errors in either group. 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. Interestingly, a hybrid model in which a simple negative recency bias guides behavior in early training outperforms all reinforcement learning models. Future studies could shed light on the role of goals in the acquisition of such action sequences, and the way learning shifts from simple to more complex mechanisms, as has been shown to exist for single-step action (see, for example, Hommel

et al. [65] 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.