Control of complex actions in humans and robots

Kleijn, R.E. de

CHAPTER 5

Predicting action plan formation in sequential reactive and reinforcement learning

Almost all types of everyday action can be considered sequential. From making coffee to using the bathroom, these complex actions consist of subactions that are completed one after another. The mechanisms by which we learn such action sequences and execute them has been the subject of investigation for many decades. An early theory by James [69] argued that elementary action units in a sequence are triggered by the sensory effects of the preceding unit. However, Münsterberg [102] noted that such an associative account is insufficient to explain sequential action because a directional element is required to successfully execute subactions in the correct order. Instead, he argued that the learning of action sequences relies on the acquisition of a motor program. Tubau et al. [160] suggested that these two approaches are

This chapter is an adaptation of the article *de Kleijn, R., Kuipers, M., Kachergis, G., & Hommel, B. (in preparation). Predicting action plan formation in sequential reactive and reinforcement learning.*

not mutually exclusive, but in fact reflect two different executive control modes that—under specific circumstances—can be strategically chosen.

5.1.1 Stimulus-based and plan-based control

Tubau et al. [160] compared James's stimulus-driven account of sequential action with the *prepared reflex* concept of Hommel [61], and referred to it as *stimulus-based control*. This type of executive control is characterized by the automaticity by which stimuli are attended to. Due to the highly automatized response to stimuli, the sequence itself is often not learned. Instead, what is learned is a strategy of delegating control to external stimuli [160]. In other words, people learn how to respond quickly to incoming information. *Plan-based control*, on the other hand, is assumed to rely on action plans, which are structured sequences of action effects [62, 96]. In contrast to stimulus-based control, representations in plan-based control are *internally* generated.

There is evidence to suggest that sequence learning does not rely on the prediction of sequences of external stimuli, but the prediction of the motor action to be performed. In other words, participants do not learn stimulus–event sequences, but in fact learn sequences of *responses*. As such, it is thought that sequence learning involves a shift from stimulus-based control to plan-based control, implying the generation of action plans by which participants can predict a sequence of responses even in the absence of stimuli [60, 104].

Tubau et al. [160] investigated this shift and its modulators in a comprehensive study consisting of five experiments. In a serial reaction time paradigm in which participants had to respond to the letter X appearing on the left or right side of the screen and responding with the appropriate hand, they presented participants with a repeating sequence of stimuli. In this sequence, location switches occurred four times more often than location repetitions, but stimuli were equally often presented to the left or right. They found that participants' control mode was influenced by instruction type, where intentional instruction (i.e. telling participants that the shown sequence is deterministic, and is to be learned explicitly) induced plan-based control. Participants' control mode was assessed by the size of the frequency effect, which should be smaller under plan-based control. Participants having received intentional instructions showed a smaller frequency effect, which was attributed to the formation of an action plan. Also, these participants were more likely to have acquired explicit knowledge of the sequence, as they were able to verbally report the correct sequence at the end of the experiment[1].

However, plan-based control is not just a strategy that participants employ at their own choosing—task structure and demands have a large influence. For example, removing stimulus–response compatibility by using symbolic stimuli instead of spatially compatible stimuli seems to lead to plan-based control, as is evidenced by the elimination of the frequency effect. Also, playing irrelevant sounds that hamper symbolic encoding of the sequence prevents the successful formation of an action plan, leaving stimulus-based control the only viable mode of executive control [160]. In some circumstances (for example the exploratory paradigm discussed later), stimulus-based control is not a feasible strategy due to the lack of stimuli.

5.1.2 Studying sequence learning

The acquisition of action sequences has been the subject of study in domains ranging from linguistics [41, 136] to everyday action [18, 32], with perhaps the *serial response time task* (SRT, [107]) being the most popular paradigm.

In the SRT task, a visual stimulus appears in one of four locations, horizontally distributed on a computer screen. Four buttons are located below the four possible stimulus locations, and participants are asked to press the button below the visual stimulus that appears as quickly as possible. In their original study, Nissen and Bullemer [107] compared a con-

[1] Although it should be noted that explicit sequence knowledge is not at all necessary for learning (see e.g. [89, 107])

dition using random stimulus locations with a condition using a repeating, deterministic sequence, and found evidence for implicit sequence learning: participants in the deterministic sequence showed larger reduction in response times than participants in the random condition.

Most of the sequence learning literature has focused on cued paradigms such as the SRT task, in which participants have to respond to sequences of stimuli that appear. However, it seems clear that sequence learning in daily life is often not learned by simply chaining stimulus–response associations [87]. Instead, acquiring new action sequences is better characterized as *exploratory*, in which people try several alternatives before discovering the correct one.

In one recent study, Kachergis et al. [77] adapted the SRT task to a reinforcement learning paradigm. In this task, participants were not *cued* by the stimuli, but had to *explore* the four alternatives to find out which one was correct. Participants could collect points by predicting the next stimulus correctly. A strong correlation was observed between behavior on the SRT task and its reinforcement learning adaptation in terms of response time and accuracy per sequence position. Interestingly, the final scores were bimodally distributed, suggesting that participants used different strategies. Although purely stimulus-based control is impossible in this paradigm, it is clear that the accuracy of participants' action plans showed a large range of variance. Although their study investigated both the SRT task and its reinforcement learning adaptation, the study had a between-subject design, making it impossible to examine characteristics of participants that produce effects that are common to both tasks.

5.1.3 The current study

In scenarios where both stimulus-based control and plan-based control are possible, participants may strategically (or perhaps even randomly) choose an executive control mode. In the current study, we investigated predictors of executive control mode in an SRT task and action plan formation in a reinforcement learning task in which plan-based control is the only control mode available.

Earlier research has shown that visuospatial working memory capacity predicts both implicit and explicit sequence learning performance [16, 17]. In this study, we will look at visuospatial working memory capacity and IQ measurements as predictors of executive control mode that reflect cognitive limitations. One possibility would be that some participants simply do not have the cognitive capacity to form (long enough) action plans. Another possibility would be that control modes are chosen strategically or preferentially. The formation of an action plan might reflect individual differences in the need for structure. That is, some people may prefer to actively predict the future according to a plan or schema instead of waiting for stimuli to arrive, while others might want to delegate control to the external environment [105].

5.2 Method

5.2.1 Participants

Forty undergraduate and graduate students (13 males, 27 females) were recruited from Leiden University. Participants either received course credit or were paid 6.50 euro for participation. All participants had normal or corrected-to-normal vision. The total duration of the experiment was approximately 90 minutes.

5.2.2 Materials

In order to assess possible predictors of participant behavior, several tasks and questionnaires were administered.

Fluid intelligence

Fluid intelligence was estimated using a shortened, 10-minute version of the Raven's Standard Progressive Matrices (SPM) test [124]. It measures the individual's ability to form perceptual relations and for analogical reasoning. It is a widely used test to measure fluid intelligence, independent of language and schooling, and is considered to have excellent reliability [24]. The number of correct responses in 10 minutes over all participants are normalized to a distribution with mean 100 and SD 15, resulting in an estimated IQ score.

Locus of control

To investigate the influence of an individual's locus of control on control mode, we administered the Levenson Multidimensional Locus of Control Scales [88], a 24-item questionnaire consisting of three subscales: (1) internality, (2) powerful others, and (3) chance. People who have an internal locus of control tend to perceive reinforcement as a result of one's behavior, while people with an external locus of control tend to perceive it as a result of factors beyond one's control. It could be hypothesized that people with an internal locus of control are more likely to engage in plan-based control, while people with an external locus of control are more environment-driven.

Personal need for structure

To assess participants' tendency to seek out structured ways of dealing with the world, we administered the Personal Need for Structure scale [158]. This questionnaire quantifies people's need for simple structure, and consists of 12 statements (e.g. "I enjoy having a clear and structured mode of life.") which the participant can either agree or disagree with, rated on a 6-point scale. It has been shown to have good reliability and validity [105]. It has been hypothesized that personal need for structure reflects a strategy for simplifying the world due to a general lack of intellectual abilities, but the correlation between the PNS scale and IQ seems to be minimal [105]. It is therefore more likely to reflect a strategy that participants can choose to employ, and participants who score high on this measure could be more likely to actively search for structure in action sequences.

Visuospatial working memory

We assessed visuospatial working memory using the computer task from Bo et al. [17]. In their study, which used an adaptation of the visual working memory task used by Luck and Vogel [92], a relationship was found between visuospatial working memory capacity and performance on a serial reaction time task. In this task, participants were presented with a sample array for 100 ms followed by a blank screen delay of 900 ms, after which a test array was presented for 2000 ms. Participants were asked to determine whether the test array was different or similar to the sample array by pressing either D or S. Arrays consisted of 2–8 colored circles, and for each trial the test array was either the same as the sample array or different with one of the colors changed. Visuospatial working memory capacity was calculated as K = array size × (hit rate – false alarm rate). The average K across all array sizes was computed to estimate visuospatial working memory capacity [17]. Participants completed 140 trials in total.

Trajectory SRT task

The trajectory SRT task is an adaptation of Nissen & Bullemer's serial response time task [107]. It maps the four buttons of the original SRT task to four squares located on the corners of a computer screen, requiring participants to move the mouse cursor to each square that lights up [74, 75]. Unbeknownst to participants, the sequence is a repeating sequence of 10 items. Speed-up over time compared to a condition with a random sequence is thought to reflect implicit learning of the sequence. In the current study, we used a different sequence (3–2–4–2–1–4–3–4–2–1) than in the original SRT task to prevent carryover effects between this task and the RL task. The complete task consisted of 800 movements (80 repetitions of the 10-item sequence).

In order to assess first-order frequency effects, the sequence was designed in such a way that it consisted of 8 straight movements, and 2 diagonal movements. After completing the 800 movements, participants were asked if they noticed any structure within the experiment, and if so, were asked to reproduce the sequence.

Reinforcement learning task

The RL task is an adaptation of the trajectory SRT task (see above), with the difference being that the next stimulus is not cued, but to be discovered by the participant through trial-and-error [77]. Participants moved to one of the four squares, and received feedback by the square turning green in the case of a correct movement, and being returned to the center of the screen in the case of an incorrect movement. Points were awarded for correct movements (+1 point), and deducted for incorrect movements (−1 point), and participants were instructed to maximize their amount of points. The amount of points collected was continuously visible to the participant, their progress in the task, however, was not. The task ended after 800 correct movements of the original SRT sequence (4–2–3–1–3–2–4–3–2–1).

A participant having knowledge of the sequence before starting and who never made a mistake would therefore make 800 movements directly to valid targets, receiving a theoretical maximum score of 800 points. A participant with no memory of even the previous target they had tried could make an infinite number of mistakes, never finishing the experiment. If participants would not repeat the same invalid target more than once when seeking each target (i.e. an elimination strategy), a participant would expect on average to score 0 points, as the expected value of completing one movement successfully is 0 using this strategy[2]. Participants were not told that there was a repeating deterministic sequence, let alone details such as how long the sequence was.

[2] 33% chance of 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).

5.2.3 Design and procedure

All participants performed both the trajectory SRT task, as well as the reinforcement learning task. The order of the two tasks was counterbalanced over participants, and the two tasks used different sequences to prevent carryover effects.

Participants were seated at a computer after having given their informed consent. All subsequent tasks were performed on the computer. First, the Personal Need for Structure questionnaire was completed, followed by the Levenson Multidimensional Locus of Control questionnaire, the visuospatial working memory task, and Raven's SPM. After this, participants were given a 5-minute break. Participants then completed, in counterbalanced order, the trajectory SRT task and the reinforcement learning task.

5.3 Results

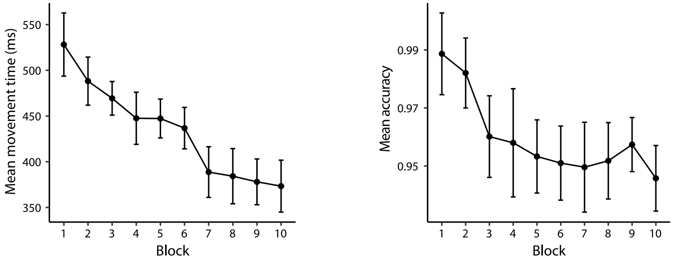
5.3.1 Trajectory SRT task

Data preparation

Prior to analysis, movement times >1500 ms were removed, and the experiment was divided into 10 blocks of 8 sequence repetitions. As an analysis of data collected earlier (described in Chapter 4) using a random sequence showed no significant difference in movement times between straight and diagonal movements, there was no correction applied for the somewhat larger distance required to make diagonal movements.

Response times

Comparative analyses were performed using the means of participants' median movement time, with the movement time defined as the time between cue onset (stimulus changing color) to touching any part of the stimulus with the mouse cursor. Median movement time to a target was 464 ms (*SD* = 223 ms). Participants' movement time decreased from 546



(a) Participants' movement time decreased over time, indicating learning of the sequence.

(b) Error rates increased during the first three blocks, but remained relatively stable during the rest of the task.

Figure 5.1 | Movement times and accuracy for the trajectory SRT task. Error bars indicate within-subject 95% CI.

ms in the first block to 413 ms in the tenth block, indicating learning of the sequence, $F(9, 360) = 15.80$, $p < .001$, $\eta^2_G = .126$.

Accuracy was high across all blocks of the experiment, but especially so during the first two blocks. There was an effect of time on accuracy, $F(9, 360) = 4.50$, $p < .001$, $\eta^2_G = .042$, indicating some degree of speed-accuracy tradeoff. However, after the third block movement times are still decreasing, while accuracy remains stable. Both movement times and accuracy are shown in Figure 5.1.

Explicit sequence knowledge

Participants were grouped into an implicit knowledge group and an explicit knowledge group. Only those 13 participants who could correctly recall the complete repeating sequence after having completed the task were considered to have explicit knowledge. Participants with explicit sequence knowledge had a significantly larger working memory capacity (2.87 vs. 2.25, $t(28.08) = 2.95$, $p = .006$, $d = 1.11$), but did not differ on estimated IQ, the Levenson Multidimensional Locus of Control scales,

Factor	*df*	*F*	η^2_G	*p*
Block	9, 342	23.94	.17	< .001
Block × Knowledge	9, 342	8.37	.07	< .001
Frequency	1, 38	106.00	.09	< .001
Frequency × Knowledge	1, 38	4.43	.004	.042
Frequency × Block	9, 342	2.75	.005	.004
Knowledge	1, 38	6.44	.089	.015

Table 5.1 | Results of analysis of variance on movement times.

or Personal Need for Structure scales (ts < .81, ps > .42).

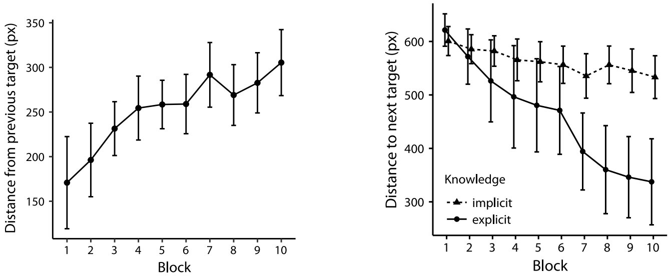
Modes of executive control

Similar to Tubau et al. [160], we used frequency effects (i.e. the facilitation of responses to frequent (straight) compared to infrequent (diagonal) transitions) to determine whether participants engaged in either stimulus-based or plan-based control. An analysis of variance revealed main effects of block, frequency, and knowledge on movement time (see Table 5.1). Overall, participants with explicit sequence knowledge had faster movement times (M = 398 ms) than participants without (M = 485 ms), and frequent (straight) movements were performed faster (M = 417 ms) than infrequent (diagonal) movements (M = 496 ms).

Predictive movements

As the task progressed, participants made an increasing amount of movement during the ITI—in the absence of a stimulus, $F(9, 342) = 6.53$, $p < .001$, $\eta^2_G = .053$. Total ITI (predictive) movement, defined as the distance from the previous target at the onset of the next target, increased from 171 pixels in block 1 to 305 pixels in block 10. There was no main effect of knowledge. Results are shown in Figure 5.2a.

Similar to Dale et al. [34], we can then define *correct* predictive movement as the distance to the next target at target onset. An analysis of



(a) Participants made increasingly larger movements during the ITI.

(b) Larger movements during the ITI reflect correct prediction of the next stimulus, as initial distance to the stimulus decreased over time.

Figure 5.2 | Predictive movements in the trajectory SRT task. Participants made increasingly larger predictive movements, which reflects correct prediction of the next stimulus. This effect was stronger for explicit than for implicit learners. Error bars indicate 95% CI.

variance using block and knowledge as factors shows a main effect of block, meaning that distance to next target decreased from 609 pixels to 474 pixels, or that correct predictive movement increased over time, $F(9, 342) = 32.36$, $p < .001$, $\eta_G^2 = .22$.

In the final block of the task, predictive movements (defined as movements larger than 300 pixels during the ITI, but not necessarily toward the correct target) appeared to show a mixed distribution over participants. Where some participants hardly showed any movement during the ITI, others had almost half of all their movements classified as predictive. Hartigan's dip test of unimodality [58] confirms this observation, $D = .079$, $p = .038$.

While implicit learners hardly increased their correct predictive movements, explicit learners showed a strong increase over time, as evidenced by a block × knowledge interaction, $F(9, 342) = 14.00$, $p < .001$, $\eta_G^2 = .11$. Results are shown in Figure 5.2b.

Centering behavior

In Chapter 4, participants in the random condition showed more centering behavior than those in the deterministic condition. This finding suggested that centering is a strategy that can be employed in the absence of reliable sequence knowledge, minimizing the distance to possible targets. Indeed, centering behavior, defined as the proportion of the ISI spent in the center 100 × 100 pixels of the screen, was highest for participants without explicit sequence knowledge, $t(30.46) = 2.34$, $p = .026$, $d = .85$.

5.3.2 Reinforcement learning task

As explained in Section 5.2.2, maximum score on the reinforcement learning task was 800, with the most basic elimination strategy leading to 0 points. Mean score was 525, ranging from 140 to 774 points. Distributions of scores was non-normal, with a large group of participants scoring 700 points, and a group scoring quite low. For subsequent analyses, a midpoint split on 457 points was performed, dividing the participants into low and high performers.

Predicting task performance

Low performers on the reinforcement learning task had a significantly lower estimated IQ of 91.4, compared to high performers with an estimated IQ of 104.9, $t(39) = 3.06$, $p = .004$, $d = .98$. Also, low performers had a significantly lower visuospatial working memory capacity of 2.13 vs. the high performers' 2.65 capacity, $t(39) = 2.40$, $p = .021$, $d = .77$. Results are shown in Figure 5.3. IQ and visuospatial working memory capacity were uncorrelated, $r(39) = .213$, $p = .181$.

There was no difference between the two groups on the Levenson Multidimensional Locus of Control scales, $t(39) = .27$, $p = .790$, and no difference on the Personal Need for Structure scale, $t(39) = .28$, $p = .780$.

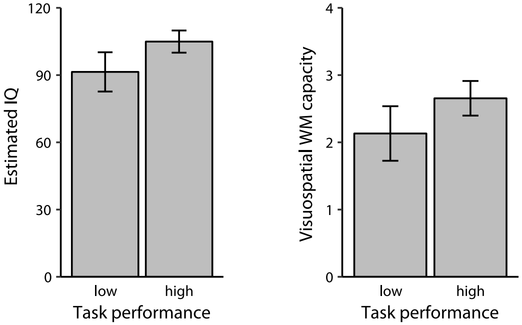


Figure 5.3 | Differences in estimated IQ and visuospatial working memory capacity between low and high performers on the reinforcement learning task. Error bars indicate 95% CI.

Explicit sequence knowledge was strongly related to task performance, as the 23 participants with explicit sequence knowledge had a far higher final score (M = 634) than participants without explicit knowledge (M = 375), $t(24.67) = 4.61$, $p < .001$, $d = 1.86$.

Stimulus- vs. plan-based control In the SRT task, two measures of executive control mode are used. First, explicit knowledge of the sequence is considered to be an indicator of a plan-based control mode. Second, the amount of correct predictive movements is evidence of the existence of an action plan, implying a plan-based control mode.

Participants with explicit sequence knowledge in the SRT task were more likely to have acquired explicit sequence knowledge in the reinforcement learning task, McNemar's $\chi^2(1, N = 40) = 4.5$, $p = .034$. This suggests that the acquisition of explicit knowledge in both tasks relies on a common mechanism or dependency. However, the amount of correct predictive movements in the SRT task was not related to the final score in the reinforcement learning task, $r(38) = -.025$, $p = .880$, nor did explicit knowledge in the reinforcement learning task relate to correct predictive movements in the SRT task, $t(38) = 1.32$, $p = .195$.

In summary, participants using plan-based control in the SRT task did not score higher on the reinforcement learning task, but participants with explicit knowledge formation in the SRT task *were* more likely to acquire explicit knowledge on the RL task. This suggests that predictive movements and explicit knowledge do not similarly reflect successful plan formation, and may not be equally good indicators for a plan-based control mode.

5.4 Discussion

5.4.1 Movement trajectories

Learning was evident in both the trajectory SRT task and the reinforcement learning task. In the trajectory SRT task, the findings of Tubau et al. [160] were replicated. The trajectory paradigm allowed us to find further evidence for a plan-based mode of control: participants made increasingly large movements toward the next stimulus, but participants with explicit knowledge of the sequence did more so than those with implicit knowledge. Instead, participants without explicit knowledge showed centering behavior during the ITI, moving the mouse to a position equidistant to all possible targets.

This centering strategy has been described in the literature (e.g. [34]), but has not before been associated with quality of prediction or sequence knowledge. Our results show that this behavior is a function of explicit sequence knowledge. It has been suggested that this centering behavior is an artifact of the spatial layout of the task, but we hypothesize that the centroid of any polygon defined by response locations should be a preferred (optimal) resting place when waiting for an uncertain stimulus. Future studies should be able to shed light on this theory.

5.4.2 Limitations preventing plan formation

In the reinforcement learning task, final scores showed a bimodal distribution, similar to what has been reported in [77]. The low-performing and high-performing groups differed in IQ and working memory capacity, but did not differ in personal need for structure or locus of control. This suggests that sequence learning performance in an exploratory paradigm is not determined by personal characteristics or preferences, but by cognitive limitations.

In both the SRT task and the reinforcement learning task, explicit sequence knowledge was predicted by visuospatial working memory capacity. Earlier research by Bo et al. [17] showed a relationship between visuospatial working memory capacity and performance on a non-trajectory SRT task, but the current study shows that this holds in an exploratory paradigm as well and predicts explicit sequence knowledge. The observation that participants who were more likely to acquire explicit sequence knowledge in the SRT task were also more likely to acquire it in the reinforcement learning task further corroborates this finding.

5.4.3 Suggestions for future research

A promising approach to investigating this relationship is by modeling the learning process in the reinforcement learning task (see Chapter 4 for an example). IQ and visuospatial working memory could be compared to the learning rate and state space in reinforcement learning models that are fit to the performance of individual participants. This may shed further light on the exact mechanisms that explain the wide range of performance on exploratory sequence learning.

Another possible explanation of the diverse learning outcomes could be rooted in different beliefs about the task. Participants were not told that the response locations would be a repeating, deterministic sequence. They may have instead believed it was to some extent probabilistic—as many psychological tasks are. Different assumptions about the task may lead participants to arrive at different strategies, with variable success in the task. Participants expecting a random sequence may be less inclined to predict the next stimulus and are—in the current paradigm—indistinguishable from participants expecting a deterministic sequence but unable to learn it due to intellectual limitations. However, manipulating these variables is straightforward and could be an interesting avenue for future research.