Control of complex actions in humans and robots

Kleijn, R.E. de

CHAPTER 7

Summary and general discussion

7.1 Summary of this dissertation

The research described in this dissertation tried to shed light on the relation between complex action control in humans and robots. Taking the acquisition of action sequences as an example, a paradigm for the study of sequential action was introduced, and several models were discussed that can account for sequence learning and execution.

7.1.1 How human and robotic complex action control are related

First, the main obstacles in the way of autonomous, everyday action execution by robots were discussed from a cognitive psychological viewpoint in Chapter 2. Four main categories of problems are identified that need to be dealt with in order to make truly flexible, autonomous robots: (1) the integration of symbolic and subsymbolic planning; (2) the integration of feedforward and feedback planning and execution mechanisms; (3) the structure of action representation; and (4) the contextualization of action control.

Early AI planners, such as STRIPS [47], were designed to reach an intended goal state from an initial state through symbolically represented subac-

tions. This symbolic nature of action representation has many advantages: it allows, for example, for easy manipulation of action components leading to efficient planning. Early approaches in the study of human sequential action also assumed a symbolic representation of action sequences, with subsymbolic (sensorimotor) triggers responsible for timing. Both James [69] and Washburn [167] suggested a *chaining theory* of sequential action, in which the sensory feedback produced by executing the subaction at t_0 would trigger the execution of the subaction at time t_1. However, several empirical findings seem to be incompatible with a chaining account of sequential action. For example, such models cannot account for context effects as found in studies into anticipatory lip rounding, in which facial muscles adapt to sounds that are to be produced later in time [14] or Gentner's typewriting studies that showed a large amount of movement in anticipation of subactions several units ahead [51]. Instead, models that integrate symbolic and subsymbolic representations such as the typewriting model suggested by Rumelhart and Norman [131], seem to be more promising. In this model, the correct temporal order of subaction execution is ensured by feedforward inhibition.

As the field of robotics advances from repetitive, predictable actions such as factory work to highly dynamic and complex actions in everyday life, feedforward control systems alone are no longer sufficient. On the other hand, feedback systems are often slow as they require information from the environment to be produced and detected. Successful integration of feedforward and feedback control systems is needed to create agents that are both fast and adaptive. The existence of feedforward planning mechanisms in humans is demonstrated by the relation between onset delay and sequence complexity in finger and arm movements [59], as well as Eriksen et al.'s [44] linguistic studies on number pronunciation. However, feedback control mechanisms are essential for filling in parameters unavailable or unreliable at planning time, such as object weight and required grip strength. Hybrid architectures, in which skeleton action plans are generated by feedforward mechanisms, and where parameters are filled in by feedback processes seem to combine the best of both

worlds [53, 64].

Another difficulty in complex action planning is that the meaning and purpose of subactions vary with the goal that they serve to accomplish. In AI planners, the function of goals is to guide the selection of task components, and in cognitive processing models such as ACT-R goals reduce the search space, making task preparation more efficient [33]. Some authors have argued against the representation of goals for two reasons [18]. First, goals themselves may be context-dependent, and as such require different subactions to accomplish them. Second, many everyday activities such as taking a walk do not always have clearly defined goals. Others, however, emphasize that it is the representation of goals that makes useful action plan manipulation such as subaction substitution or skipping possible [33]. Alternatively, *implicit* goal representation from a TEC viewpoint can be viewed as a kind of "intentional weighting" mechanism in which relevant features are activated more than others, priming the agent to execute different subactions [65, 95]. Whatever the exact nature of goal representation, it is clear that some form of end-state representation is necessary to generate flexible behavior.

Chapter 3 discussed how the relationship between cognitive psychology and cognitive robotics developed over time. After breaking away from philosophy, psychology found itself depending on unreliable, subjective information. In a push toward reliable, empirical observation as the basis of a scientific psychology, behaviorism emerged as the method that could put psychology on par with the natural sciences. However, behaviorism proved untenable as a general theory of human behavior as it could not account for fundamental cognitive processes such as language and memory, leading to what is now known as the neocognitive revolution. Meanwhile, in the 1950s the field of artificial intelligence arose from cybernetics, mathematics, and computer science, and over the following decades expert systems such as MYCIN and symbolic AI (now known as GOFAI, *good old-fashioned AI*) were able to show impressive results. Also, computers were slowly beginning to gain public interest. Cognitive psychologists started to wonder if humans are like computers: input–output devices

with sensory information as input and behavior as output, known as the so-called computer analogy. Meanwhile, roboticists were considering animal behavior as a foundation for robot control. Some early cognitive robots were roughly inspired by biology [21], but even more specific parallels could be drawn between humans and robots.

The problem of integrating feedforward and feedback control in robotics had gained interest as task demands for robots became less predictable. Where the absence of a feedback loop in a factory environment may not be a big problem as long as all manipulanda are in the correct location and orientation, feedback is required in almost all situations in the outside world. Brooks' *subsumption architecture* [21] was a response to traditional GOFAI and showed that complex behavior could emerge without the traditional separation of feedforward and feedback systems. However, this architecture worked for rather low-level behavior such as wandering, avoiding, and homing, and it is unclear how well it would scale up to more complex situations. More complex, goal-directed behavior in robots is usually the product of a *planner*[1]. This component takes an intended state, compares it with the initial state, and determines the actions to take in order to successfully reach the intended state. Traditional planners such as STRIPS fail when one of the subactions cannot be successfully completed, and backtrack to try alternative subactions.

7.1.2 Empirical studies on sequence learning

One of the most widely used paradigms in sequence learning is the serial reaction time (SRT) task [107]. In this task, participants are asked to press the button associated with one of four horizontally distributed stimuli. Unbeknownst to the participants the four stimuli appear in a repeating, deterministic sequence. Over time, participants show a larger decrease in response times compared to a random sequence, indicating learning of the sequence. However, due to the discrete nature of this task it is

[1]Although both Brooks [22] and Braitenberg [20] are excellent examples of apparent complex behavior *without* a planner.

impossible to investigate interstimulus processes such as prediction or context effects [146].

In Chapter 4, we described an adaptation of the SRT task into the continuous domain. Instead of four discrete buttons associated with stimuli, we presented four squares in the corners of a computer screen, with the instruction of moving the mouse cursor as fast as possible to the square that changes color. This type of data collection allows researchers to capture the temporal dynamics of cognitive processes and the interaction between them [48, 146, 148]. First, we were able to replicate Nissen & Bullemer's [107] original findings: more speedup in the deterministic, repeating sequence than in a random sequence. Second, we showed that this speedup was due to predictive responses made during the ITI, and that participants employed different strategies. While some participants actively moved the cursor to the next target during the ITI, others used a centering strategy in which they moved cursor to a central location equidistant from all possible alternatives, a phenomenon reported earlier in the literature [34, 38].

Due to the questionable ecological validity of the SRT task—after all, everyday sequence learning is not often characterized by merely responding to attention-grabbing stimuli—we adapted the SRT task to a reinforcement learning paradigm. In this task, participants no longer could respond to squares changing color but had to actively explore the alternatives, receiving a 1-point reward when choosing the correct alternative, and a reward of –1 for choosing an incorrect alternative. Participants varied widely in the amount of points collected. To investigate possible causes, we fit three model-free reinforcement learning models: (1) Q-learning, (2) SARSA, and (3) Q-learning with eligibility traces.

Reinforcement learning models are a class of machine learning models that learn what to do in order to maximize reward, roughly inspired by operant conditioning in cognitive psychology. As such, the learner is not told explicitly what to do—as is the case in supervised learning—but has to discover which actions produce the highest reward through trial-and-error. In traditional reinforcement learning models, each possible action

that can be taken in a given state has a certain *value*: the immediate reward the action will yield plus the total amount of reward that can be expected in the future. In order to keep track of these values, they are often stored in a table[2], mapping discrete actions in discrete states to Q-values.

The models as used in their current form were not able to approach the final scores of the best human participants. However, Q-learning performed better than SARSA, and Q(λ) produced even better results. The relatively bad performance of Q-learning—which was quite surprising given the relative simplicity of the task—could be due to the specific action selection policy used. This is further explained in Section 7.2.2.

In the study described in Chapter 4, we found centering behavior to be a function of uncertainty, and a large variance in scores attained on the reinforcement learning task. To further examine these phenomena, the study described in Chapter 5 used a larger sample and a within-subject design. We wanted to investigate the factors that predict successful plan formation, and compare performance between the responsive SRT task and the exploratory reinforcement learning task. Participants in an SRT task can rely on two modes of executive control: stimulus-based control and plan-based control [160]. Under stimulus-based control, participants are prepared to respond to stimuli in an automatized fashion, delegating control to the external stimulus. Under plan-based control, an internal representation of the motor plan is made. These two modes of executive control can be strategically chosen under some circumstances. In a reinforcement learning paradigm, stimulus-based control is not a viable strategy, as there are no external stimuli to respond to. Participants were asked to perform both tasks described in Chapter 4 in randomized order, as well as complete measures of IQ, visuospatial working memory, need for structure, and locus of control.

For the SRT task, we used three measures of plan-based control: (1) the

[2]The action–value function need not necessarily be represented as a table. In fact, much progress has been made in the last years using (deep) neural networks as action-value function approximators, see e.g. [98].

acquisition of explicit knowledge about the sequence, (2) predictive movement toward the correct target in the inter-stimulus interval, and (3) the magnitude of frequency effects. Participants who acquired explicit sequence knowledge made increasingly larger predictive movements over the course of the task, whereas participants without explicit sequence knowledge hardly did so. Of all predictors, only visuospatial working memory predicted the acquisition of explicit sequence knowledge.

For the reinforcement learning task, both visuospatial working memory and IQ predicted final score. This suggests that the formation of an action plan in the current paradigm is limited by cognitive capacity, although another explanation could be that people with high IQ or WM are more likely to *actively look* for structure in sequential tasks.

In Chapter 6, we investigated the centering behavior described in Chapters 4 and 5 in more detail. We used a simulated robotic arm controlled by an artificial neural network to perform the same task as the one described in the earlier chapters: moving the mouse toward a stimuli that appear in a deterministic, repeating order. In one condition, the networks were provided with accurate information about the next stimulus, similar to human participants that have learned the sequence and are able to predict the next one. In another condition, the networks were given a random stimulus location as a prediction, making the prediction uninformative in that it contains no useful information about the next stimulus. In a third condition, we did not provide any stimulus location as a prediction, i.e. the input to the prediction units were fixed at zero.

We found that the networks that were given accurate predictions evolved predictive behavior. They moved toward the next stimulus after touching the current one, but before the next one appeared. The networks with either random or no prediction developed a centering strategy similar to the one described in Chapters 4 and 5: they moved the cursor to the center of the screen, an optimal location to wait for the next stimulus to appear.

7.2 Discussion and future directions

7.2.1 Sequential action under stimulus-based and plan-based control

Two modes of executive control were discussed and studied in this dissertation: stimulus-based and plan-based control. Our paradigm was a hybrid between our earlier trajectory SRT work and Tubau et al.'s [160] study into stimulus-based vs. plan-based control. In our design, we used a sequence with straight (left–right or up–down) movements being more frequent than diagonal movements in order to examine frequency effects, which were found by Tubau et al. [160] to decrease under plan-based control. However, due to the increased dimensionality of our paradigm there are many more possible frequency effects in play: horizontal repeat or switch, vertical repeat or switch, diagonal or straight, stimulus location, etc. This reduced the usability of frequency effects as a measure of plan-based control.

Other shortcomings with the used paradigm can be identified. Although the use of the original SRT sequence allows for a straightforward comparison with earlier work (e.g. [107]), this sequence is not specifically designed for the analyses conducted in our work. For example, with four alternatives the distances between alternatives are not identical, as diagonal movements require longer distances than straight movements. Although an analysis of response times between diagonal and straight movements in the random condition did not show an effect of movement type on response times, other properties of these movements could affect our results. For example, Burk et al. [23] found that movement distance affects decision making, and this could have made diagonal movements a less attractive choice for participants because they require more effort to perform. Additionally, location and transition probabilities are not balanced in the standard SRT sequence. A similar, balanced three-alternative paradigm could be used in future research to remove these confounds.

Another interesting avenue of research would be the role of stimulus probability on centering behavior. If the centering behavior described in

this dissertation is indeed due to minimization of mean travel distance to stimuli, altering stimulus probabilities would cause the centering location to shift toward more probable stimuli. This can be investigated both by using human participants as subjects, or in a simulated robotics paradigm such as the one described in Chapter 6.

7.2.2 Reinforcement learning: action selection and parameter fitting

The studies described in this dissertation compared human performance on a sequential reinforcement learning task with the performance of three reinforcement learning models: Q-learning, SARSA, and Q(λ). For a reinforcement learning model to perform well, the method of action selection it uses needs to balance between *exploitation*, using the information it has gathered from experience and that is stored in its Q-table, and *exploration*, allowing the model to try other and possibly better actions. At the start of any task or learning process, the Q-table may have been initialized to zero, or filled with small, random values. Either way, the information it contains is uninformative, and therefore should not be used for action selection. Different action selection policies deal differently with this problem. Several different action selection policies are used in the literature:

- **greedy**: the agent always selects the action that maximizes the value estimate;
- **random**: the agent always selects an action at random;
- ϵ-**greedy**: the agent selects the action that maximizes the value estimate Q with probability $1 - \epsilon$, otherwise it selects an action at random;
- **softmax**: the agent selects an action based on weighted probabilities by applying a softmax function over the value estimates. A temperature parameter τ can be used to control the spread of the softmax distribution.

The greedy policy could be considered purely exploitative, while the random policy is purely explorative. It should be clear that neither policy will provide good results in the paradigms described in this dissertation, as the greedy policy will always choose the action that happens to have the associated highest random value at Q-table initialization, while the random policy will never use the information stored in the Q-table. In the study described in Chapter 4, an ϵ-greedy policy was used. However, preliminary analyses of the data (not described in this dissertation) show that both softmax and another policy have the potential of outperforming even humans. The policy involves temporal decay of random action rate ϵ in the ϵ-greedy policy. ϵ is initialized to a relatively high value at the start of the sequence, exploring all possible actions and updating the Q-table with associated rewards. As the Q-values stabilize over the course of the experiment, ϵ decreases, making use of the informative Q-values that now populate the Q-table.

Also, the learning rule and action selection policy interact, as is clear from their definitions. 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 $\gamma \in [0, 1]$:

$$Q'(s,a) = (1-\alpha)Q(s,a) + \alpha(r + \gamma Q[s', \underset{a'}{\operatorname{argmax}}(Q[s', a'])]) \tag{7.1}$$

SARSA, on the other hand, 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 uses the experience tuple <s, a, r, s', a'>:

$$Q'(s,a) = (1-\alpha)Q(s,a) + \alpha(r + \gamma Q[s', a']) \tag{7.2}$$

Under a greedy action selection policy, Q-learning and SARSA are equivalent[3], and will update Q with the maximum attainable reward in state s': Q-learning by definition, and SARSA by virtue of always selecting

[3]Although note that Q-learning first updates Q, and selects the next action based on the updated Q, while SARSA chooses the action first and then updates Q.

the action that will yield the maximum attainable reward. Future studies should investigate the influence of action selection policies and their parameters on model performance in the paradigms discussed in this dissertation.

Also, if these reinforcement learning models are shown to be able to outperform humans in the task described in Chapters 4 and 5, parameter fitting could shed light on the nature of individual differences between human participants if the models turn out to be identifiable using specific cost metrics. For example, a final score of only 200 points could be due to either a low value of learning rate α, placing too little weight on the latest reward, or a too high value of random action rate ϵ, taking too many exploratory actions instead of exploiting the information in the Q-table. Instead, by looking at the learning trajectory, and using it as an error function, it could be possible to make these models identifiable. As another interesting manipulation, the reward schedule of the reinforcement learning task could be manipulated. By making certain rewards contingent on (a series of) earlier actions, differences in discount rate γ could be investigated, making this paradigm quite versatile for explaining individual differences.

7.3 Conclusion

This dissertation concerned itself with everyday action, and the mechanisms by which humans and robots are able to perform it. First, we described the fundamentals of everyday action, and explained that it is not as simple as the word implies. Also, we described the capacities a robot should have in order to perform everyday action. Next, we investigated the similarities and differences between human and robotic action control. Several mechanisms by which motor control is learned (e.g. motor babbling and reinforcement learning) are already common to both human and robotic action control.

The adaptation of the SRT task into a trajectory paradigm allows for the observation of predictive processes in sequential action control, and

shows that participants tend to adopt either a predictive or reactive strategy. Our results suggest that the quality of the action plan that is formed is a function of individual limitations in IQ and visuospatial working memory. The reinforcement learning models investigated did not perform as well as humans, but we suspect that the specific action selection policy used was partly to blame.

Participants who did not generate a reliable action plan tended to engage in centering behavior: moving the cursor to the center of the screen in anticipation of the next stimulus. We showed, using an evolutionary robotics approach, that this behavior evolves in an artificial neural network that controls a robot arm as a function of prediction quality. This suggests that this behavior is an emerging strategy caused by task constraints. The optimality of this behavior should be investigated further by manipulating target frequency and location.

Overall, the paradigms presented in this dissertation are well-suited to investigate both symbolic sequential action in the form of reinforcement learning, as well as sensorimotor action control in the form of evolved motor behavior in a robot arm controlled by an artificial neural network. Both paradigms provide ample opportunity for manipulation to further investigate the commonalities between complex human and robot action control.

Acknowledgments

The authors would like to thank Denis O'Hora and an anonymous reviewer for helpful comments and feedback on the study described in Chapter 4.