Processes Behind Reinforcement Learning: An examination of the temporal progression of processes behind successful and unsuccessful reinforcement learning strategies

Berg, Myrna van den

Processes Behind Reinforcement Learning

An examination of the temporal progression of processes behind successful and unsuccessful reinforcement learning strategies

Myrna S. van den Berg

Cognitive Neuroscience

Leiden University

28-01-2022

Supervisor:

Dr.ir. R.E. de Kleijn

Department of Psychology

Leiden University

Abstract

The recent trend towards gamification could lead to an increase of the number of learning problems that need to be solved through reinforcement learning. It is therefore important that people learn how to solve reinforcement learning problems. Understanding which learning strategies people use, the processes behind them, and identifying sub-optimal learning strategies could prove very beneficial for teaching people the best reinforcement learning strategies. This study examines the processes behind reinforcement learning strategies through cognitive modeling. A reinforcement learning model was fitted on human behavior on a reinforcement learning problem. Some people were able to fully solve the problem and others were not. The temporal parameter trajectories of these two groups were compared to each other. The group that solved the problem showed expected results with a lot of learning and exploration at the start and less learning and more exploitation towards the end of the task. The other group started similarly but then started to learn less while exploration remained high. This could indicate that these people would benefit from short learning sessions after which they are able to focus on something else.

Processes Behind Reinforcement Learning

An examination of the temporal progression of processes behind successful and unsuccessful reinforcement learning strategies

1 Introduction

The ability to learn is an integral part of human and animal life. As soon as we are born, we learn (Gray, 2011; Holt, 1967; Iverson, 2010). It is, therefore, no surprise that several research fields related to psychology and biology (e.g. developmental psychology, neurology, and social psychology) have been interested in *when* and *how* humans and animals learn (Bransford et al., 2000; Gray, 2011; Kolb & Whishaw, 2009).

In recent decades, learning has expanded from an ability exclusive to humans and animals to an ability that can be incorporated into artificial agents and machines (e.g. computers). (Sutton & Barto, 2018). Often these machine learning algorithms can solve specific learning problems better than humans but lack the ability to transfer their newly obtained knowledge to (slightly) different learning problems. Even algorithms specifically designed to transfer knowledge from one problem to another, remain limited in their ability to transfer knowledge successfully (Lazaric, 2012). Nevertheless, the proper algorithm should be able to always solve a specific learning problem at least as well as well-performing humans.

Conversely, humans are not always able to flawlessly solve a learning problem. Different people tend to approach problems differently and some approaches might lead to a less-than-optimal or unsuccessful result (Nisbet & Shucksmith, 2017). To be able to understand why some people can learn successfully and others cannot, it is recommended to investigate the learning strategies they used when solving the learning problem (Williams & Burden, 1997, as cited in Lee, 2010).

There are several ways humans, animals, and machines learn (Gray, 2011; Sutton & Barto, 2018). One way is through trial-and-error-based learning (Thorndike, 1898). Little to no information is given to the learner, forcing the learner to try out actions and learn

the consequences of these actions afterward. Once at least some consequences are known, the learner can select actions that lead to the preferred outcome. A rather well-known trial-and-error-based learning method is reinforcement learning (Kaelbling et al., 1996; Sutton & Barto, 2018); originally based on the principles of operant conditioning (Skinner, 1981)

This study aims to examine the process behind reinforcement learning strategies. A recent trend towards gamification (Zainuddin et al., 2020) will likely result in an increasing amount of learning taking place through reinforcement learning in the future. Therefore, understanding the processes behind successful and unsuccessful reinforcement learning strategies could prove necessary in teaching people how to avoid unsuccessful reinforcement learning strategies in environments where reinforcement learning becomes ever more present. We take a first look into the processes behind reinforcement learning in a very controlled environment.

Before diving into the fine details of the study, the concept of reinforcement learning, how the technique of cognitive modeling allows to research the processes of learning, and the definition and examination of previous work on learning strategies will be examined.

1.1 Reinforcement Learning

Reinforcement learning has been a subject of interest for the last few decades (Kaelbling et al., 1996) and can refer to a learning problem or a self-learning model (Szepesvári, 2010). *Reinforcement learning as a learning problem* has been part of the psychology field since Pavlov and Skinner started their research into conditioning (Clark, 2004; Skinner, 1981). Especially Skinner (1981) used reinforcement learning to train pigeons to perform specific behaviors, such as turning a full circle. This method would become known as operant conditioning. Skinner discovered that positive reinforcements (e.g. food) would encourage the learner to perform the same actions again, while negative reinforcements (e.g. a shock) would discourage the learner to perform the same actions

again.

In its simplest form, reinforcement learning is a reward-based learning method to learn a sequence of steps (Janssen & Gray, 2012; Kaelbling et al., 1996). Rewards can be anything that a learner wants to earn at that moment (e.g. points or money) and can be earned cumulatively for each N-number of steps taken or earned all at once at the end of the problem (Sutton & Barto, 2018). Using these rewards it is possible to solve a reinforcement learning problem without needing to exactly specify how this feat can be accomplished. One only needs to know how to maximize the total reward obtained from the problem (Sutton & Barto, 2018; Szepesvári, 2010), or to minimize the costs of negative rewards (Bertsekas & Tsitsiklis, 1995).

Since little to no instructions are given, learning takes place through trial-and-error (Kaelbling et al., 1996; Sutton & Barto, 2018). The learner has to figure out the rules, action consequences, and rewards by themselves to achieve the highest possible reward. Along the way, the learner is bound to make errors. However, these should decline over time, signaling that the learner is learning to solve the reinforcement learning problem (Kaelbling et al., 1996).

Reinforcement learning is part of our everyday life. The most straightforward presence of reinforcement learning is in video games, where the correct actions progress the game or lead to large rewards. However, reinforcement learning also happens more covertly when, for example, an enthusiastic reaction from a friend for a nice birthday gift encourages you to make another effort the next year. However, in the near future, reinforcement learning might start to play an even larger role in our daily lives. Recently there has been a rise in gamification in which game elements (e.g. rewards) are used to encourage learning (Zainuddin et al., 2020). Hence, increasing knowledge on the topic could prove beneficial and perhaps even vital for how well people are able to learn in the future.

In recent decades, interests in *reinforcement learning as a self-learning model* increased with the large increase in computing power (Sutton, 1992). Nowadays,

reinforcement learning models have been updated; expanded on; and combined with other (self-learning) models, such as, deep neural networks to create deep reinforcement learning. Several action value update algorithms have been introduced, such as Q-learning (C. J. Watkins & Dayan, 1992; C. J. C. H. Watkins, 1989), Temporal difference learning (TD-learning) (Sutton, 1988), the Monte Carlo method and algorithm (Metropolis et al., 1953; Metropolis & Ulam, 1949), as well as actions selection algorithms such as adaptive ϵ-greedy (Tokic, 2010) and softmax (Sutton & Barto, 2018). By combining reinforcement learning with a neural network and some human knowledge, Silver et al. (2016) were able to build AlphaGo and defeat the world GO champion. A year later, Silver et al. (2017) created AlphaGoZero without human knowledge and defeated AlphaGo.

These days, companies have also shown a great interest in implementing reinforcement learning models for commercial purposes. Especially tech companies, such as Tesla, Amazon, and Netflix, have been using reinforcement learning in their products (e.g. self-driving cars, speech assist, and video recommendation respectively). Furthermore, reinforcement learning models are now being used for advancement in healthcare. For example, models have been used to mine biological data in an effort to better understand pathologies (Mahmud et al., 2018) and to design new drugs (Popova et al., 2018). Lastly, reinforcement learning models have been implemented in household appliances, such as automated vacuum cleaners, to perform or optimize performances on tasks, such as learning where the most dust tends to gather (Mahmood et al., 2018; Sutton & Barto, 2018). This is by no means an extensive list of applications. Reinforcement learning models are used in many different ways, for many different purposes.

Reinforcement learning models have also found their way back to the research field of psychology. For example, models have been used as statistical tools (e.g. for bootstrapping) or as a cognitive model able to give insight into underlying processes in the human brain (Sutton & Barto, 2018). A variation of the latter will be conducted in this research.

1.2 Cognitive Modeling

Cognitive models have become a crucial tool for research in the (neuro)cognitive field (Busemeyer & Diederich, 2010). They allow researchers to have a look into the 'black box' that is our brain (Anderson, 2014) and help unravel the mysteries behind, for example, memory, attention, decision making, and learning. Actions, processes, and thoughts behind these systems can be modeled by stripping the theories and ideas behind them down and translating them into algorithms and the parameters within (Anderson, 2014). When the correct model is built, the model should be able to perform the required behavior, such as solving a learning problem. This way theories about processes in the human brain can be explored, developed, and tested (McClelland, 2009).

Although cognitive models might be associated with computers, computers are not necessary to engage in cognitive modeling (Anderson, 2014). Cognitive modeling has been part of cognitive psychology long before Turing (1936) would lay the groundwork for the first electronic computer. Nevertheless, computers made modeling a lot more efficient and convenient (Anderson, 2014). The recent rise in computing power has enabled the creation of increasingly complex and computationally demanding cognitive models (McClelland, 2009).

Researchers have started to use reinforcement learning models as cognitive models (Daw & Frank, 2009; Janssen & Gray, 2012). For example, they have adjusted basic reinforcement learning models to allow them to solve a specific learning problem (Balkenius & Winberg, 2004), indicating that something similar should be happening in the brain. Others have used reinforcement learning models to enhance the performance of their previously existing cognitive models (Khalid et al., 2020).

Several research groups have used variations of reinforcement learning models to solve, variations of, multi-armed bandit learning problems (**kretchmar2002parallel**; Besbes et al., 2014; Bouneffouf & Féraud, 2016; Even-Dar et al., 2006; Sutton & Barto, 2018). In these types of problems, learners need to learn which of the one-armed bandits

presented to them results in the highest turnover reward. Typically, the rewards are given based on a chance system. Variations of this problem include non-stationary versions where the underlying model of the problem changes (Besbes et al., 2014) and variations in which the participant is informed beforehand of the reward trends they will encounter (Bouneffouf & Féraud, 2016). Reinforcement learning models have been able to solve learning problems like these successfully (Sutton & Barto, 2018), suggesting that, according to the ideas behind cognitive modeling, similar reinforcement learning mechanisms are likely to be present in humans when they are able to solve the problems.

In this research, we will also use a reinforcement learning algorithm as a cognitive model to solve a learning problem that could be considered a variation of a multi-armed bandit task. As mentioned before, cognitive models are created with a theory or notion of some processes that might underlie specific behaviors. Then a model that can only do those processes is built in hopes of achieving similar behaviors as observed in humans. The question usually can be stripped down to: is this specific algorithm able to solve the given learning problem? In the current research, the focus is slightly different. Instead of seeing which parameters are needed or which variation of the algorithm is necessary to solve a learning problem, the focus will be on if and how parameters in a typical reinforcement learning model change over time and differ between participants when the model is fitted to behave like a human. According to Sutton and Barto (2018), a reinforcement learning model typically includes a learning rate parameter, a discounted reward rate parameter, and, depending on the exact algorithm, one or two exploration parameter(s).

To elaborate, the learning problem used in this study could be considered a combination of a multi-armed bandit task (Sutton & Barto, 2018) and a serial reaction time (SRT) task (Nissen & Bullemer, 1987) in which participants need to respond to a series of stimuli. The former is present in the task because participants needed to select from options given and were then rewarded or punished for their choice. However, instead of learning which fixed option would give the best rewards, the participants had to learn a

sequence determining which option contained the best reward. Crucially, to ensure variation in how successful people are in solving the learning problem, the problem needed to be difficult enough that some people would fail the task and others would successfully solve it.

1.3 Learning Strategies

Learning strategies can be defined as "a sequence of procedures for accomplishing learning" (Schmeck, 2013). They are the specific behaviors or steps taken in hopes of solving a learning problem (Scarcella & Oxford, 1992). Which learning strategy people use depends on the learning problem and their own personal preference (Schmeck, 2013). The latter depends in turn on the skills they possess and how they use these skills (Schumaker & Deshler, 1992), their expertise on (the field of) the problem (Bransford et al., 2000), and their learning style (i.e. the general approach to a learning problem, R. L. Oxford, 2003). The latter tends to be fairly consistent (Entwistle & Peterson, 2004; Schmeck, 2013). Two people solving the same learning problem might vastly differ in their learning strategy and one person solving two different learning problems might select different learning strategies for each problem. Someone might even decide halfway through solving the problem to change their learning strategy to increase performance. Nevertheless, while people could be using many learning strategies each time they learn, they tend to use a small subset of strategies they prefer.

Various research fields (i.e., animal studies, cognitive psychology, developmental psychology, social psychology, education, and artificial intelligence) have been interested in studying learning strategies (Leelawong & Biswas, 2008; Weinstein & Underwood, 1985). Unsurprisingly, a considerable number of studies on learning strategies focus on how people, in particular children, adolescents, and students, learn in and outside the classroom (Mayer, 1988; Weinstein & Underwood, 1985). These studies often focus on topics these people have to learn, such as learning a language (R. Oxford & Crookall, 1989) and

mathematics (Berger & Karabenick, 2011). The focus on this group is unsurprising because people have to learn a lot in a relatively short time during this early period in their lives (Gray, 2011). This is likely the period where people develop and stabilize most of their learning strategies. Thus, people could benefit all their life from an early correction of learning strategies that lead to an unfavorable outcome. To do this a lot of information about learning strategies is needed including information about the processes behind them.

Many measurements have been created to investigate what learning strategies people use (Nisbet & Shucksmith, 2017; Schmeck, 2013). Again, a large part of these studies focus on how people study (Weinstein & Underwood, 1985), but not on the cognitive processes happening during learning. These measurements often involve observing behavior or asking for the subjective experience of people (R. Oxford & Crookall, 1989). While these measurements can give a lot of insight into many aspects of learning, they lack the ability to examine the processes behind learning. However, the latter is possible with cognitive modeling (Anderson, 2014; McClelland, 2009) allowing for an expansion of the knowledge on reinforcement learning strategies that could not be accessed through the classic measurements.

1.4 Current Study

In conclusion, the current research aims to study the processes behind reinforcement learning strategies of people who are successful in solving the learning problem and compare them to the processes of people who are unsuccessful in solving the learning problem. For this purpose, a relatively standard reinforcement learning model will be used as a cognitive model. Learning strategies can be extracted from the model by interpreting the progression of parameter values during problem-solving when the model is fitted on human behavior. The information gained from this research not only increases knowledge but could also help to combat unsuccessful learning strategies with a more targeted approach. Presently, reinforcement learning might not be very overtly present in daily life

and there might not seem to be many reinforcement learning problems to solve. However, this could soon change due to the rise of gamification.

Expectations regarding reinforcement learning strategies are difficult to make, especially for the people that will be unable to successfully solve the learning problem. It all depends on the progress of the parameters. For example, one of the parameters is *learning rate* (α). One could expect the learning rate to drop for people that quickly solve a reinforcement learning problem, since no additional learning is needed, but remain relatively high for people that do not solve the problem, indicating that they try to keep learning and solving the problem, but fail on some other aspect. On the other hand, the learning rate of people that don't solve the problem might quickly drop, preventing them from learning and solving the problem properly. However, what is certain is that temporal analyses will be conducted for each parameter in which the two groups (i.e. successful and unsuccessful reinforcement learning problem solvers) will be compared.

2 Methods

2.1 Participants

The participants were 40 students from Leiden University that participated for either a fixed monetary reward or fixed credits score. Of these participants 13 were male and 27 were female. All of the participants completed the experiment without complications.

2.2 Materials

To examine the processes behind participants' strategies during a reinforcement learning task with a cognitive model, both a reinforcement learning problem and a reinforcement learning model were needed. These were, however, not the only two measurements in the study. This study was part of a larger umbrella project that also measured IQ scores with Ravens Progressive Matrices, Locus of Control, participants'

personal need for structure, and working memory. In addition, participants had to complete a serial reaction time (SRT) task during the studies. However, this study focuses exclusively on the reinforcement learning problem and the reinforcement learning model. Therefore these other measurements will not be mentioned again.

2.2.1 The Reinforcement Learning Problem

The reinforcement learning problem used for this study was the same as De Kleijn et al. (2018) used in their studies. The participants were shown four squares on the screen, one in each quadrant (top-left = 1, top-right = 2, bottom-left = 3, bottom-right = 4). The squares could color red or green and give -1 or +1 points depending on whether they were the incorrect or correct square respectively for that trial. In each trial, three squares were incorrect and one was correct. The next trial would start when the correct square was found. A different square would then become the correct one and the correct square became an incorrect square. The goal was to score as many points as possible. Both the participants and the model had to finish 800 trials. To reach the highest possible score both the participants and the model had to learn a sequence of ten consecutive correct positions in a row. This sequence was derived from Nissen and Bullemer (1987): 4-2-3-1-3-2-4-3-2-1 and was proven to be difficult enough for some, but not all, people to fail to learn the full sequence.

The task was performed on a 21-inch monitor computer at Leiden University with a 1024x768 resolution and 60Hz refresh rate. The acceleration of the cursor was disabled which was suggested by Fischer and Hartmann (2014).

2.2.2 The Reinforcement Learning Model

A rather default reinforcement learning model, described by Sutton and Barto (2018), was used. The model contained a Q-learning (C. J. Watkins & Dayan, 1992) algorithm (i.e. a model-free reinforcement learning algorithm), the softmax action selection algorithm, and the maximum likelihood estimation for optimization.

2.2.3 The Q-Learning Algorithm

The Q-learning algorithm assigns action values to each action in a trial. These action values depend on the reward the model expects to receive directly from that action immediately and in the future after a series of actions. The higher the reward, the higher the action value will be. Equation 1 shows the algorithm used in this study to calculate action values. The values are updated after every action.

As seen in equation 1, the algorithm is parameterized by α (i.e. learning rate parameter) and γ (i.e. discounted reward rate parameter). The higher α is the more the model focuses on new information and forgets old information. Higher γ values mean a higher emphasis on future rewards, while lower γ values mean a higher emphasis on immediate rewards. The values of both parameters ranged between 0 and 1.

$$Q(s_t, a_t) \leftarrow (1 - \alpha) \cdot Q(s_{old,t}, a_{old,t}) + \alpha \cdot (r_t + \gamma \cdot max(Q(s_{t+1}, a)) \quad (1)$$

2.2.4 The Softmax Action Selection Algorithm

The softmax action selection algorithm calculates the likelihood that each action will be selected for each trial (see equation 2). These probabilities depend on the action values calculated by the aforementioned Q-learning algorithm.

The probabilities calculated range from 0 to 1 excluding 0 and 1. This allows for some occasional exploration when a task is fully learned. This, in turn, could help change strategies after a possible change in the rewards values. The exploration of other options is not completely random, as is the case with the -greedy action selection algorithm. Instead, each option has an individual probability of being selected, reflecting outcome differences between these other options. While a score of -1 might be bad, it is not nearly as bad as a score of -100. The softmax algorithm skews the probabilities towards the lesser of the two evils, while the -greedy algorithm assigns equal selection probability to both options. Hence, the softmax algorithm is generally the preferred action selection algorithm.

The softmax algorithm contains an exploration parameter τ (or sometimes called *temperature*). The lower this parameter value is, the larger the difference between the action selection probabilities is. This leads to an increased chance that the model will exploit actions with known favorable outcomes.

$$P(s,a) = \frac{e^{Q(s,a)/\tau}}{\sum e^{Q(s)/\tau}} \tag{2}$$

2.2.5 Optimize the Model's Fit

The model was fitted onto the behavior of each participant. To do this, the three parameters discussed needed to be configured so that the behavior trajectory of the model would come as close as possible to the behavioral trajectory of each participant. To be able to test how far the model deviated from the human trajectory, we needed an error metric. The maximum likelihood, or the minimum log-likelihood, was selected for this purpose. The model already demanded a lot of computational power and the softmax algorithm would already calculate the likelihood of each action. Therefore, by using the minimum log-likelihood, little to no extra computational power would be needed.

For each action a participant would take, the model would pick the same action and return the probability of that selection. The higher the returned probability, the more likely the model would have done the same as the participant. Then the log values were calculated over these probabilities (see equation 3) so that the values could be accumulated and optimized towards zero.

$$LL = -2Log(P(\textit{model picks the same action as the participant})) \tag{3}$$

2.3 Procedure

Participants were asked to come to Leiden University to participate in the study. They would first sign an informed consent letter before they were allocated to a computer in a quiet room. Participants were given no instructions on how to do the task.

The task had already been started by the experimenter. The participants saw a white screen with four black squares in each corner of the screen, a mouse cursor in the middle, and a score of zero points. There was no indication or cue for the participants to start. They could move the mouse whenever they wanted. Once they moved their cursor over one of the squares, that square would become either red or green and the participant would either lose or gain 1 point, respectively. The square remained colored until the participant moved his cursor to a different square. The next trial would start when the participant found the green square.

After participants had completed the task, they were asked whether they thought there was any sequence present in the task. When the participants agreed that there indeed was a sequence, then they were asked to give the sequence. Participants that correctly could recall the full sequence, irrespective of the starting point in the sequence, were marked as having explicit knowledge of the sequence.

2.4 Analysis

The main analysis conducted in this study evaluated the temporal progression of the parameter values while solving the learning problem. To extract this temporal progression, parameters needed to be optimized several times during the task. Every optimization would take a lot of computational power and therefore a lot of time. Therefore, we decided to limit the number of times the model had to optimize the parameter to every 100 trials, resulting in 8 optimizations. Furthermore, for each optimization, all previous trials would be used and not only the 100 new trials. The information learned in previous trials is not suddenly forgotten by humans and so should also be used for accurate optimization of the parameters.

To find the optimal parameter configurations grid-searches were performed. In these searches, α and γ values could range between 0 and 1. After exploring what the minimum and maximum possible values of τ would be in our sample, the search space was set to

comfortably fit these values. The search space of the parameter value was set between 0.5 and 4. The parameter configurations with the lowest log-likelihood scores would then be returned.

After extracting the data, a within-subject mixed model ANOVA was performed to test the main and interaction effects of the temporal trajectory of each parameter, and task performance of people that performed well or poorly.

3 Results

All participants solved the problem to some degree. All participants managed to have a final positive score even though there were three incorrect options and one correct option in each trial. The scores ranged from 140 points to 774 points ($M = 523.8$, $SD = 207.53$). On average, participants needed 1074 moves to finish the reinforcement learning task. Almost two-thirds of the participants ($N = 26$, 65%) managed to explicitly recite the whole sequence directly after they finished the task.

3.1 Learning Rate (α)

The within-subject mixed model ANOVA for the learning rate parameter, α, showed a main effect of temporal change ($F(1, 278) = 61.27$, $p < .001$), no main effect of group difference between people with and without explicit knowledge of the task ($F(1, 153) = 0.06$, $p = .803$), and a significant interaction effect between temporal change and explicit knowledge ($F(1, 278) = 18.06$, $p < .001$).

For both groups (i.e. with and without explicit knowledge of the sequence) the learning parameter value decreased over time (see figure 1). The parameter decreased more (i.e. to a lower point) for participants without explicit knowledge than for participants with explicit knowledge of the sequence

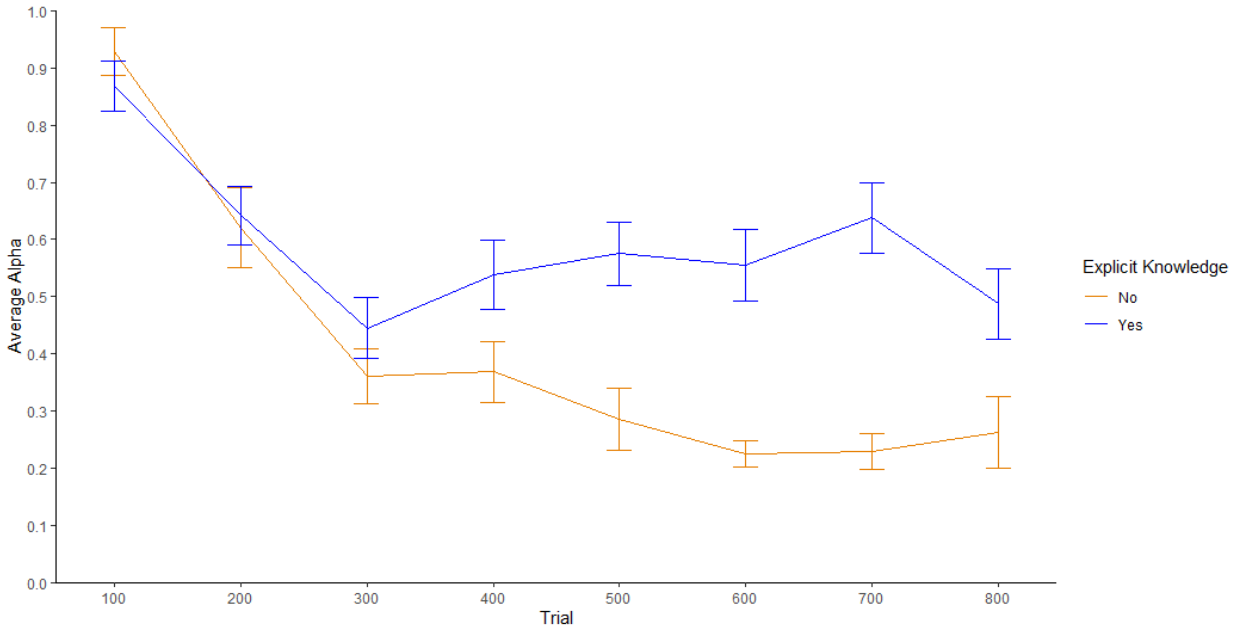

Figure 1. *Temporal change of the learning rate parameter (α) split by explicit knowledge.*

3.2 Discounted Reward Rate (γ)

The within-subject mixed model ANOVA of the discounted reward rate parameter showed a main effect of temporal change of the parameter value during the task ($F(1, 316) = 82.88$, $p < .001$), a main effect of explicit knowledge ($F(1, 316) = 29.92$, $p < .001$), and an interaction effect of temporal parameter value and explicit knowledge ($F(1, 316) = 33.23$, $p < .001$).

However, these results cannot and should not be interpreted. The values of the parameter were on the boundary of the feasible parameter spaces. Elaborations and discussions of this issue and the parameter itself will be covered down in the discussion.

3.3 Exploration (τ)

Lastly, the within-subject mixed model ANOVA for the exploration parameter showed a main effect of temporal change of the parameter value during the task ($F(1, 278) = 29.94$, $p < .001$), no significant effect of the explicit knowledge of the two groups ($F(1, 66) = 2.79$, $p = .0996$), and a significant interaction effect of the temporal parameter value

change and explicit knowledge ($F(1, 278) = 86.03$, $p < .001$).

Similar to the learning parameter, the exploration parameter shows a temporal effect and an interaction effect. Thus, the parameter changes over time regardless of performance. However, when the two groups are split, as shown in figure 2, we see that participants with explicit knowledge of the sequence decreased in their exploration value while participants without explicit knowledge seem to increase in their exploration value. Suggesting that people with explicit knowledge decreased in the amount they explored during the task, while participants without explicit knowledge increased in the amount they explored during the task.

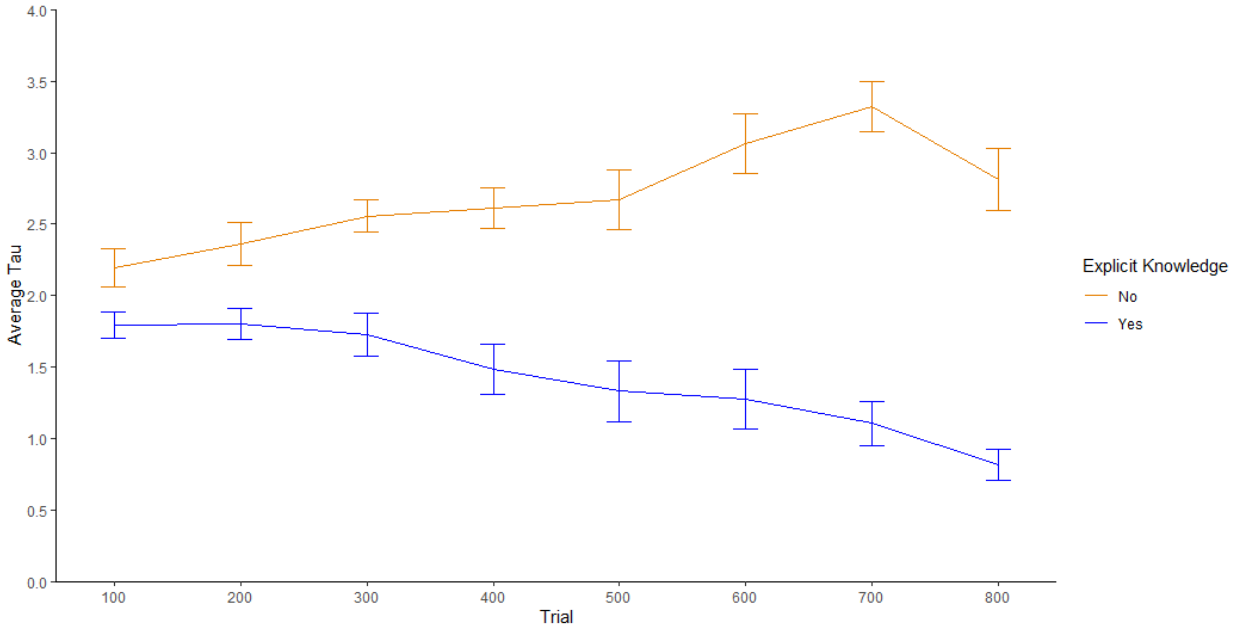

Figure 2. *Temporal change of the exploration parameter (τ) split by explicit knowledge.*

3.4 Additional Exploration and Discoveries

Testing if participants have explicit knowledge of the sequence is only one way to split the group. Since the final score could also represent participants' performances, the group was also split by the mean final score. An advantage of this method is that both groups now contain the same number of participants ($N = 20$). The most notable difference is that, in addition to the aforementioned significant effects, the main effect of

group difference for the exploration parameter τ became significant in a similar analysis to the one done before ($F(1, 100) = 18.42$, $p < .001$).

Furthermore, a strong correlation was found between the model fit score (i.e. log-likelihood) and the performance of the participants measured by their final score ($r(38) = .9977$, $p < .001$, see figure 3). Indicating that the better a participant performed on the reinforcement learning task, the better the model could fit its behavior to the actions of the participant.

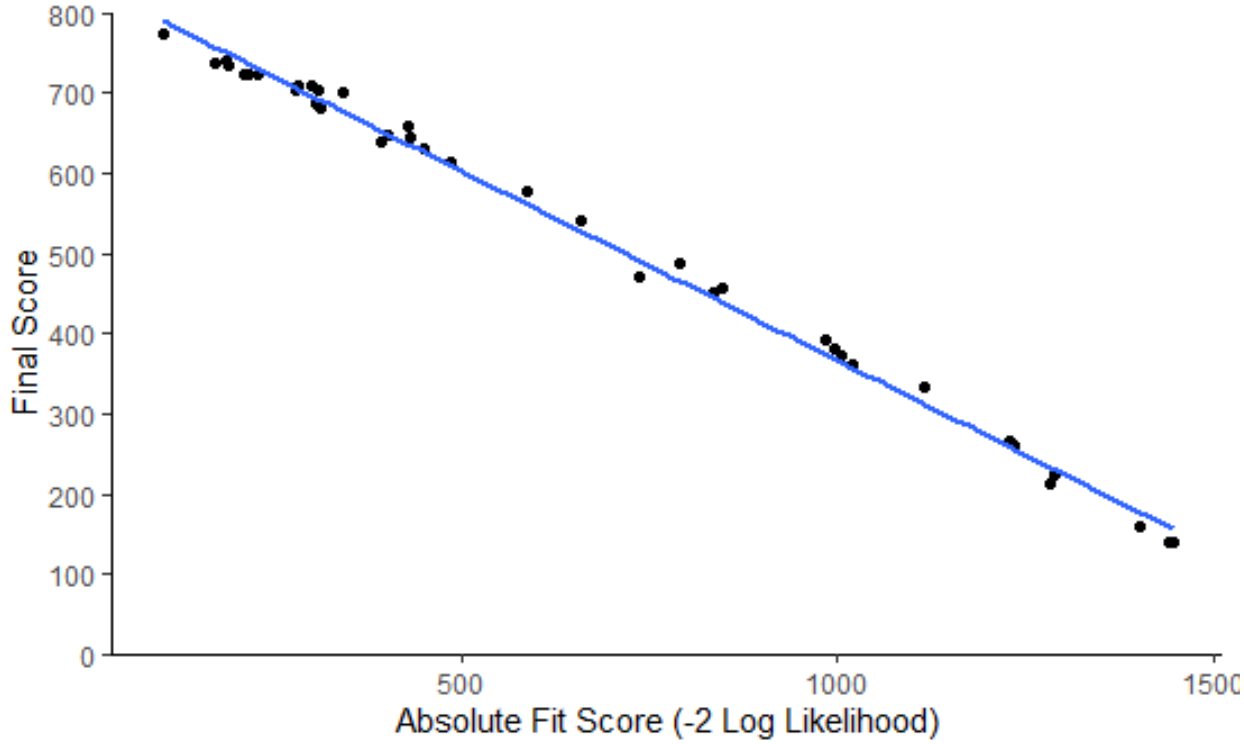

Figure 3. *Correlation between the performance of the participants (reflected by their final score of the task) and the fit score of the model (i.e. -2 log likelihood score of the model) for each participant.*

4 Discussion

4.1 Summary of the Findings

In summary, this study aimed to examine the processes behind reinforcement learning with a cognitive model to see if there is a difference in the learning strategies of participants that were able to fully solve the learning problem (i.e. the explicit knowledge group) and those that could not fully solve it (i.e. the non-explicit knowledge group). For this purpose, a relatively standard reinforcement learning model with three parameters (i.e. learning rate, discount reward rate, and level of exploration) was fitted on human behavior on the reinforcement learning problem. The temporal trajectories of the parameters for participants with and without explicit knowledge were analyzed and compared. Learning rate (α) decreased relatively early for both groups but decreased most for the non-explicit knowledge group. Furthermore, discount reward (γ) also showed main and interaction effects, suggesting a strong increase for both groups to focus on future rewards instead of immediate rewards. However, the data had some singularity issues. Third, the exploration parameter (τ) seemed to stay the same or even increase for the non-explicit knowledge group and steadily decrease for the explicit knowledge group. Lastly, a strong correlation between model fit scores and participants' performances was found.

4.2 Discussions and Implications of the Findings

4.2.1 Learning rate

The temporal decrease of the learning rate value (α) of the explicit knowledge group is in line with our expectations. The problem never changes and therefore participants in this group do not need to learn new information. Perhaps more surprisingly, the learning rate parameter value of the non-explicit knowledge group also decreased. The value decreased even more than the value of the explicit knowledge group. This decrease indicates that they likely stopped learning a lot of new information before they were able

to fully solve the problem, preventing this group from fully solving the learning problem. A possible explanation could be that the participants in this group had little to no attention and/or motivation to fully solve the problem. Motivation and learning strategies have been shown to influence each other before (Berger & Karabenick, 2011).

Hence, both groups stopped learning a lot of new information relatively early in the task. This could indicate that the attention span of both groups had reached its limit, reducing the amount of attention they can pay to new information. The explicit knowledge group had the benefit of already having solved the problem that the non-explicit knowledge group did not have. On the other hand, it could be that the explicit knowledge group could learn more information, but there simply was no new information to be learned. The lesser decrease could indicate a readiness to learn new information when presented allowing the learning rate value to quickly increase to a high value. In any case, it seems that for the non-explicit knowledge group, spending more time on the same issue in hopes of learning at some point might not work. Intuitively, we tend to think that if people spend more time on something, they will eventually learn to do it. However, here it seems that people might benefit more from short periods of learning than long learning sessions.

4.2.2 Discount reward rate

The discount reward rate values (γ) increased for both groups until values very close to one were reached. This resulted in a singularity issue because the values were on the edge of the limited valid value space. A proper analysis could therefore not reliably be conducted.

When interpreted, a close-to-one value would indicate that the participants heavily favored looking at the reward of the next trial over the reward in the current trial when deciding in the current trial. However, in both trials, the rewards were the same and no larger future reward could be obtained. Thus, the rewards in each trial were independent of the rewards in all other trials. Consequently, a look-ahead parameter would likely not

have been necessary for this model. In addition, since the rewards are independent for each trial, one would expect the value to decrease to zero. It is not necessary to look ahead when considering the action of the current trial. However, the opposite happens. This could have been a side effect of the parameter not fitting its application properly. In hindsight, the model would have likely fitted the learning problem better without this parameter. Additional research is necessary to study the process of this parameter and its effects on reinforcement learning strategies.

4.2.3 Exploration parameter

The exploration parameter value (τ) was the only value that showed a different temporal trajectory between the two groups. The explicit knowledge group showed a steady decrease in the amount they explored. In other words, they seem to increasingly exploit answers they know are correct. Since the sequence of correct answers never changed, exploiting the knowledge they had would lead to high rewards.

On the other hand, the non-explicit knowledge group showed a steady exploration value and perhaps even increased the value. This group at no point seemed to start exploiting the knowledge they did obtain, even if it was only a small amount. The increase in exploration value during the task might even indicate that the participants in this group might have discarded knowledge obtained in favor of extra exploration.

This increase in exploration might have been on purpose. Perhaps participants felt that their learning strategy would not help them solve the learning problem and so they changed strategies. The increase in exploration could signal people mentally starting over with learning halfway through the task. Unfortunately, this would likely be futile since learning new information was already at a very low point. Another reason for the increased exploration could be that some of these participants started the task poorly, having little luck in selecting the correct tiles in the first few trials resulting in a large negative score they had to compensate. Going through this might have demotivated some participants to

put in the effort to learn properly. They did not know how many trials they would get to turn the score and so they might have thought that their scores were doomed to be bad scores. Additional research is needed to test these theories and if the exploration truly stays the same or increases.

Furthermore, τ was also the only parameter for which the main effect of group differences became significant after the groups were split differently. Splitting the groups at the mean final score, instead of splitting by explicit knowledge, resulting in the main effect of group difference for the exploration parameter. When split by explicit knowledge, the main effect was already close to significance with a value just under 0.1. The new split resulted in equal group sizes, causing an increase in statistical power (Lachin, 1981) and thus probably also a significant result. Another contribution to the shift toward significance could be because the participants that were moved from the 'well-performing' group to the 'poorly-performing' group had learning strategies and parameters progressions more similar to the latter than the former. They, for example, solved the problem relatively late in the task. This would increase the difference between the two groups and result in an increased chance for a significant effect.

4.2.4 Learning strategies

Combining the learning rate and exploration parameters to examine participants' learning strategies, we see that the explicit knowledge group had a learning strategy one might expect when solving a problem like this. From start to end they: explored their options, learned which ones were best, and stuck to those options. Once they didn't need to learn as much new information, they decreased their learning rate and favored old information over new information. Since this group managed to solve the problem, one could state that this was the correct learning strategy for this learning problem.

On the other hand, the non-explicit knowledge group showed a different learning strategy. They also declined the amount of new information that they would learn in favor

of the old information around the same time as the other group. However, they seem to have done this without having properly solved the full problem yet. This group then also seemed to explore more over time, but due to the low learning rate were unable to learn the new knowledge discovered through exploration. It even seems that information already learned is not exploited in favor of more exploration.

From the decrease in the learning rate, we can derive that simply putting more time in one go into the learning problem will accomplish little toward solving the problem. When also taking exploration into account we see that even if one goes looking for new information after, for example, changing their learning strategy, the low learning rate prevents them from storing this new information. However, these strategies could have also been caused by a lack of motivation (Berger & Karabenick, 2011) causing people to pay less attention to the problem they need to solve. A better learning strategy would, therefore, probably be to take a break or focus shortly on something else than this learning problem. In practice, this could mean that students in a class might benefit from short lessons and then do some exercises, or that learning through gamification should come in short bursts.

4.2.5 Log-likelihood optimization

The reinforcement learning model used in this study increased in participants' fit scores the better these participants performed (i.e. the higher they scored). There were likely two causes resulting in this high correlation: the fit scores for each trial and the number of moves made. The model is created in such a way that it will always try to reach the goal of gaining as many points as possible. Hence, the model will almost always, except at the start of the task, be more likely to select the correct option and thus fit well for participants that also performed that action. For the other actions, the likelihood the model would select those options would be higher and thus the log-likelihood would be lower. Each time an incorrect action is performed by the participant, the log-likelihood value would be higher than for the correct action. Furthermore, the poorer the participants

performed the more moves they had to make. The model needed to also perform every move resulting in a higher number of fit scores that were summed. Hence, the sum of the fit scores was higher due to higher individual fit values for bad performances, but also because the more values were summed the worse the performance was.

4.3 Limitations

As mentioned before, the present research was part of a larger study in which many other variables were also measured in one sitting. These other measurements could have influenced the results. Participants had to spend a long time on all of the tests, which might have made them rather tired and/or unmotivated to continue performing well. Even though all these tests were conducted, the only demographic that was asked for was gender. Other relatively standard demographics (e.g. age or education level) were not asked for. This makes it impossible to know if the results are in any way generalizable and whether the sample looked balanced in terms of age distribution. This study likely would have benefited from a more direct approach, where only the participants only needed to solve this learning problem and fill in a few demographics.

Furthermore, the task itself was relatively long. On average participants needed 1074 moves, meaning that, on average, participants made 275 incorrect moves. Due to the nature of the task, where it is a lot more likely for participants to make mistakes at the beginning of the task and the task would only continue to the next trial with a correct action, we can assume that most mistakes were made in the first 300 trials by most of the participants. After these trials, the participants had to do another 500 identical trials, which seems like a bit too many trials for the learning problem presented. It could have disengaged and/or demotivate some of the participants, especially if they were not able to learn the task.

Another limitation was the algorithm itself. As mentioned before, while the algorithm was perfectly capable of learning the task correctly, the discount reward (i.e. γ)

was somewhat unnecessary to solve the current problem. While the discount reward is a standard part of many, if not most, reinforcement learning models, for this study it would have been cleaner to leave it out. An additional side effect would be that only two parameters need to be optimized, reducing the computational power the fitting process needs.

4.4 Future Works

Thus, one of the findings in this study was that the discount reward rate (γ) was redundant in the current model. Nevertheless, it would still be interesting to see what the effect of the discount reward rate would be for people that can solve a reinforcement learning problem and people who cannot. A different learning problem would be needed to research this. An interesting variation on this problem would be when not all actions need perfect to reach the optimal reward. This could manifest as something like a set of lives in games or allowing students a few mistakes, but still being able to pass the exam. It would allow people that don't fully solve the problem to still achieve this reward, which is quite realistic in our world. People are generally allowed to make some mistakes and still succeed. For the current non-explicit group, interesting questions are: Would they focus only on the large rewards or only the smaller intermediate rewards? Would they be able to obtain a large reward at all? When there are multiple large rewards, would they focus on a subset of them or possibly try to collect all of them? How successful would they be in collecting multiple large rewards? Could these larger rewards aid the participants in some way to fully solve the problem, by, for example, forcing participants to 'chunk' the information (Groome, 2014; Miller, 1956)? Finally, could these larger rewards work as an incentive to fully solve the learning problem, especially for participants that had a bad start to the task?

Not only could the task be adjusted to investigate the discount reward rate, but it could also be adjusted to add more variation, or an earlier cutoff moment, to make the task

more enjoyable for the participants and perhaps research the influence of motivation on the performance. One could for example change the rules during the task and examine the effects on the parameter trajectories. With the right setup, one might be able to determine if the decrease in learning rate in one or both groups was due to a lack of attention and/or motivation.

Furthermore, an experiment in which a reinforcement learning task and the reinforcement learning model are used without having all the other measurements and where the exploratory findings found in this study are tested, is needed to validate the results that were found. Also, this study would likely benefit from a larger (and perhaps more diverse) sample of participants and a more extensive measure of the demographics to, for example, investigate whether educational level influences how long the learning rate stays high.

Lastly, reinforcement learning is only one way of learning. Cognitive models can be created for all sorts of tasks. These types of models, when used and implemented properly with the correct underlying theories and ideas, could give additional insight into the black box that is the human brain for other types of learning problems.

4.5 Conclusion

In conclusion, this study examined the process behind solving a reinforcement learning problem to better understand reinforcement learning strategies. At the start, people learn a lot of new information, but they quickly change towards a strategy that consolidates old information instead of learning new information. Notably, this change was larger for people that were unable to fully solve the learning problem, suggesting that they kept holding on to incorrect old information. Furthermore, everyone started with a relatively high amount of exploration which is expected when solving reinforcement learning problems. People that solved the problem increasingly exploited the correct actions, while people that did not fully solve the problem kept exploring and seemed to

even increase how much they explored. This could indicate that these people tried to change strategies, but were unable to store the new information or that they might have lost attention and/or motivation to keep striving toward the best possible reward. These people might benefit from short learning sessions and then focus on something else, such as practical application of the material or a break.

References

Anderson, B. (2014). *Computational neuroscience and cognitive modelling*. Sage.

Balkenius, C., & Winberg, S. (2004). Cognitive modeling with context sensitive reinforcement learning. *Proceedings of AILS 04 (Report/Lund Institute of Technology, Lund University; 151)* (pp. 10–19). Department of Computer Science, Lund University.

Berger, J.-L., & Karabenick, S. A. (2011). Motivation and students' use of learning strategies: Evidence of unidirectional effects in mathematics classrooms. *Learning and Instruction*, *21*(3), 416–428.

Bertsekas, D. P., & Tsitsiklis, J. N. (1995). Neuro-dynamic programming: An overview. *Proceedings of 1995 34th IEEE Conference on Decision and Control*, *1*, 560–564.

Besbes, O., Gur, Y., & Zeevi, A. (2014). Stochastic multi-armed-bandit problem with non-stationary rewards. *Advances in Neural Information Processing Systems*, *27*, 199–207.

Bouneffouf, D., & Féraud, R. (2016). Multi-armed bandit problem with known trend. *Neurocomputing*, *205*, 16–21.

Bransford, J. D., Brown, A. L., Cocking, R. R., et al. (2000). *How people learn* (Vol. 11). Washington, DC: National academy press.

Busemeyer, J. R., & Diederich, A. (2010). *Cognitive modeling*. Sage.

Clark, R. E. (2004). The classical origins of Pavlov's conditioning. *Integrative Physiological Behavioral Science*, *39*(4), 279–294.

Daw, N. D., & Frank, M. J. (2009). Reinforcement learning and higher level cognition: Introduction to special issue. *Cognition*, *113*(3), 259–261. https://doi.org/https://doi.org/10.1016/j.cognition.2009.09.005

De Kleijn, R., Kachergis, G., & Hommel, B. (2018). Predictive movements and human reinforcement learning of sequential action. *Cognitive Science*, *42*, 783–808.

Entwistle, N., & Peterson, E. (2004). Learning styles and approaches to studying. *Encyclopedia of Applied Psychology*, *2*, 537–542.

Even-Dar, E., Mannor, S., Mansour, Y., & Mahadevan, S. (2006). Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of Machine Learning Research*, *7*(6), 1079–1105.

Fischer, M. H., & Hartmann, M. (2014). Pushing forward in embodied cognition: May we mouse the mathematical mind? *Frontiers in Psychology*, *5*(1315), 1–4. https://doi.org/https://doi.org/10.3389/fpsyg.2014.01315

Gray, P. (2011). *Psychology* (6th ed.). New York, NY: Worth Publishers.

Groome, D. (2014). *An introduction to cognitive psychology: Processes and disorders* (3rd ed.). Psychology Press.

Holt, J. (1967). *How children learn*. Penguin Books Ltd.

Iverson, J. M. (2010). Developing language in a developing body: The relationship between motor development and language development. *Journal of Child Language*, *37*, 229–261. https://doi.org/10.1017/S0305000909990432

Janssen, C. P., & Gray, W. D. (2012). When, what, and how much to reward in reinforcement learning-based models of cognition. *Cognitive Science*, *36*(2), 333–358.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Khalid, B., Alikhani, M., & Stone, M. (2020). Combining cognitive modeling and reinforcement learning for clarification in dialogue. *Proceedings of the 28th International Conference on Computational Linguistics*, 4417–4428.

Kolb, B., & Whishaw, I. Q. (2009). *Human psychology* (6th ed.). New York, NY: Worth Publishers.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, *2*(2), 93–113.

Lazaric, A. (2012). Transfer in reinforcement learning: A framework and a survey. *Reinforcement Learning* (pp. 143–173). Springer.

Lee, C. K. (2010). An overview of language learning strategies. *Annual Review of Education, Communication & Language Sciences*, *7*, 132–152.

Leelawong, K., & Biswas, G. (2008). Designing learning by teaching agents: The Betty's Brain system. *International Journal of Artificial Intelligence in Education*, *18*(3), 181–208.

Mahmood, A. R., Korenkevych, D., Vasan, G., Ma, W., & Bergstra, J. (2018). Benchmarking reinforcement learning algorithms on real-world robots. *Conference on robot learning*, 561–591.

Mahmud, M., Kaiser, M. S., Hussain, A., & Vassanelli, S. (2018). Applications of deep learning and reinforcement learning to biological data. *IEEE Transactions on Neural Networks and Learning Systems*, *29*(6), 2063–2079.

Mayer, R. E. (1988). Learning strategies: An overview. *Learning and Study Strategies*, 11–22. https://doi.org/https://doi.org/10.1016/B978-0-12-742460-6.50008-6

McClelland, J. L. (2009). The place of modeling in cognitive science. *Topics in Cognitive Science*, *1*(1), 11–38.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Metropolis, N., & Ulam, S. (1949). The Monte Carlo method. *Journal of the American Statistical Association*, *44*(247), 335–341.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97.

Nisbet, J., & Shucksmith, J. (2017). *Learning strategies*. Routledge Kegan Paul plc.

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

Oxford, R., & Crookall, D. (1989). Research on language learning strategies: Methods, findings, and instructional issues. *The Modern Language Journal*, *73*(4), 404–419.

Oxford, R. L. (2003). *Language learning styles and strategies*. Mouton de Gruyter.

Popova, M., Isayev, O., & Tropsha, A. (2018). Deep reinforcement learning for de novo drug design. *Science Advances*, *4*(7), 1–14. https://doi.org/DOI:10.1126/sciadv.aap7885

Scarcella, R. C., & Oxford, R. L. (1992). *The tapestry of language learning: The individual in the communicative classroom*. Cambridge University Press.

Schmeck, R. R. (2013). *Learning strategies and learning styles*. Springer Science & Business Media.

Schumaker, J. B., & Deshler, D. D. (1992). Validation of learning strategy interventions for students with learning disabilities: Results of a programmatic research effort. *Contemporary Intervention Research in Learning Disabilities* (pp. 22–46). Springer.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *Nature*, *529*(7587), 484–489.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of go without human knowledge. *Nature*, *550*(7676), 354–359.

Skinner, B. F. (1981). Selection by consequences. *Science*, *213*(4507), 501–504.

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine Learning*, *3*(1), 9–44.

Sutton, R. S. (1992). Introduction: The challenge of reinforcement learning. *Reinforcement Learning* (pp. 1–3). Springer.

Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

Szepesvári, C. (2010). Algorithms for reinforcement learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, *4*(1), 1–103.

Thorndike, E. L. (1898). Animal intelligence: An experimental study of the associative processes in animals. *The Psychological Review: Monograph Supplements*, *2*(4), 551–553.

Tokic, M. (2010). Adaptive ε-greedy exploration in reinforcement learning based on value differences. *Annual Conference on Artificial Intelligence*, 203–210.

Turing, A. M. (1936). On computable numbers, with an application to the entscheidungsproblem. *Proceedings of the London Mathematical Society*, *2*(1), 230–265.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine Learning*, *8*(3-4), 279–292.

Watkins, C. J. C. H. (1989). *Learning from delayed rewards*. King's College, Cambridge United Kingdom.

Weinstein, C. E., & Underwood, V. L. (1985). Learning strategies: The how of learning. *Thinking and Learning Skills*, *1*, 241–258.

Zainuddin, Z., Chu, S. K. W., Shujahat, M., & Perera, C. J. (2020). The impact of gamification on learning and instruction: A systematic review of empirical evidence. *Educational Research Review*, *30*(100326), 1–23. https://doi.org/https://doi.org/10.1016/j.edurev.2020.100326