Deciding to Remember: Memory Maintenance as a Markov Decision Process

Jordan W. Suchow
Institute of Cognitive and Brain Sciences
University of California, Berkeley
suchow@berkeley.edu

Thomas L. Griffiths
Department of Psychology
University of California, Berkeley
tom_griffiths@berkeley.edu

Abstract

Working memory is a limited-capacity form of human memory that actively holds information in mind. Which memories ought to be maintained? We approach this question by showing an equivalence between active maintenance in working memory and a Markov decision process in which, at each moment, a cognitive control mechanism selects a memory as the target of maintenance. The challenge of remembering is then finding a maintenance policy well-suited to the task at hand. We compute the optimal policy under various conditions and define plausible cognitive mechanisms that can approximate these optimal policies. Framing the problem of maintenance in this way makes it possible to capture in a single model many of the essential behavioral phenomena of memory maintenance, including directed-forgetting and self-directed remembering. Finally, we consider the case of imperfect metamemory — where the current state of memory is only partially observable — and show that the fidelity of metamemory determines the effectiveness of maintenance.

1 Introduction

Working memory is a storage system that actively holds information in mind and allows for its manipulation, providing a workspace for thought [1, 2]. Its capacity is strikingly limited, perhaps to only a few sights or sounds [3]. Using working memory is effortful: pupils dilate, skin conductance rises, and secondary tasks become impossible to perform well [4]. Much of the research on working memory has focused on characterizing its limits and determining what gives rise to them. For example, working memory capacity is known to be lower in young children and the elderly [5], correlates strongly with a person's fluid intelligence [6, 2], is affected by sleep schedule [7], and can be impaired in people with mental disorders such as schizophrenia [8]. From this work, we have learned a considerable amount about how much can be remembered and who is best at remembering it.

Information held in working memory is malleable [9]. It can, for example, be remembered and forgotten intentionally through the processes of directed forgetting and directed remembering, which prioritize some experiences over others for later access [10, 11]. These directed maintenance mechanisms are closely related to cognitive control and to the top-down processes that determine our conscious thoughts from moment to moment [12]. At times, these control processes can backfire, causing unwanted thoughts and memories to linger despite our best intentions [13].

Given the flexibility available to the working memory system, a question naturally arises: What is the optimal way to maintain memories? What is the space of possible maintenance strategies, and how successful is each of them in retaining information over short durations?

We approach this question by likening working memory maintenance to a sequential decision process in which, at each moment, a cognitive control process decides which memories to prioritize.

We focus on a particular kind of sequential decision process known as the *Markov decision process* (MDP) [14], which provides an abstract mathematical framework for describing decision-making in a setting that is partly under control of the decision-maker (here, the maintenance process) and partly under control of the environment (here, the degradation process). Besides being well-suited to describing the problem of memory maintenance, the MDP has the added benefit of being one of the most well-understood problems in the mathematics and psychology of reinforcement learning. Thus, having established the connection, existing concepts and tools from reinforcement learning can be brought to bear on the dynamics of memory maintenance.

The plan of the paper is as follows. Section 2 describes the essential behavioral phenomena of memory maintenance and control. Section 3 formulates the problem of memory maintenance as an MDP. Section 4 begins by describing the form of a solution to the maintenance problem — a maintenance policy – and proceeds by computing the optimal policy under various reward functions. Section 5 shows how the optimal policy, and cognitively-plausible approximations thereof, can reproduce the behavioral phenomena from Section 2. In Section 6, we extend our framework to the case of imperfect metamemory, describing memory maintenance in a partially observable mind – i.e., in situations where the maintenance system has incomplete or uncertain information about the current status of actively-held memories. Section 7 discusses the results and suggests new avenues of research made possible by formulating the problem of memory maintenance in this way.

2 Phenomena of memory maintenance and control

The essential behavioral phenomena of active memory maintenance and control involve monitoring, prioritizing, and controlling memories:

Monitoring comes in the form of *metamemory*, an awareness of one's memories and the systems that store them. Metamemory is often studied in the context of long-term memory, where it is invoked to explain phenomena such as tip-of-the-tongue states and the feeling of knowing [15, 16, 17]. Healthy individuals have a rich set of metamemory skills that guide learning, decision making, and action [18]. Neurological diseases, such as Alzheimer's and Korsakoff's syndrome, adversely affect metamemory judgments, causing a mismatch between what is remembered and what is believed to be remembered [19].

Memories can be forgotten intentionally. In experiments on this process of so-called "directed forgetting", participants study some information and are then directed to remember or forget specific elements of what was studied [10, 11]. Memory tends to be better for the to-be-remembered information than for the to-be-forgotten information. For example, in Woodward & Bjork [20], participants studied a list of words and were later asked to recall as many of them as possible. This is the popular *free recall* paradigm used extensively in studies of long-term memory. Following each word's presentation, a cue appeared instructing the participants to remember or to forget the word. Later, participants were asked to recall all the words from the studied list, regardless of how those words initially had been marked. The recall task was challenging. Critically, its difficulty depended on how the word had been marked: words marked as to-be-remembered were recalled 23.3% of the time, whereas those marked as to-be-forgotten were recalled only 4.7% of the time. This is the hallmark of directed forgetting, which has been demonstrated in both long- and short-term memory [20, 11].

Directed forgetting is intimately related to cognitive control and to the processes that determine our conscious thoughts from moment to moment [12]. For example, increasing cognitive load decreases people's ability to suppress unwanted thoughts [21], and young children and the elderly have deficits in attentional processing, which makes it more difficult for them to abandon memories and thoughts that are no longer relevant [22].

3 Computational framework: the Markov decision process

A Markov decision process is defined by a state space, a set of possible actions, a transition model, and a reward function. Each is defined in turn below:

State space. We suppose that there is a memory-supporting commodity, akin to attention, that can be divided into *quanta*, each of which is assigned to a particular memory. The quanta might, for example, represent cycles of a time-based refreshing process [23] or neural populations in prefrontal

cortex that represent "token" encodings of visual events [24]. The more of the commodity assigned to a memory, the stronger and more robust it is. The state of working memory is then an allotment of the quanta to each memory, which may receive the entire commodity, only a portion of it, or perhaps none at all. The state space thus forms a $(K-1)$ regular discrete simplex, where K is the number of memories held in working memory and where the discretization is determined by the number of quanta N.

Actions. At each time step, the maintenance process selects a quantum as the recipient of maintenance. Thus the set of possible actions A is of size N, one action per quantum, and does not depend on the state.

Transition model. The transition model specifies the probability of moving from one state of memory to another and is thus a formal model of memory degradation. We will make use of the transition model proposed in [25] — i.e., a Moran process, a model of evolution in finite populations that originated in population genetics [26] and which has been used to describe dynamic processes in diverse settings. Under the Moran process, at each time step a quantum degrades because another quantum interferes with it or replaces it. The degraded quantum is chosen randomly, uniformly across all the quanta. The interfering (or replacing) quantum is determined by the action chosen by the maintenance process. We can write the state as an allotment of quanta to memories, $s = [n_1, n_2, ..., n_K]$, summing to N, the number of quanta. At each time step, one of the n's is incremented and one is decremented. The incremented n is determined by the chosen action — if the chosen action maintains a quantum belonging to that memory, it is deterministically incremented. The decremented n is chosen with probability proportional to n because the quanta are all equally likely to degrade. This defines a transition model $\Pr(s' \mid s, a)$, which gives the probability of landing in state s' given that the agent took action a while in state s [27, 28].

Reward function. Finally, there is the reward function. By definition, the agent's goal is to maximize the total reward that is received. The reward function is a mapping from states to an amount of reward that is received for landing in that state. In the case of most working memory tasks, which are episodic (in the sense that information arrives all at once and is then discarded at the end of the trial), and which have a retention interval that is known to the participant, the reward function is time-varying, taking on a value of zero everywhere until the moment of the test, at which point it becomes positive for some states and (possibly) zero for others. For simplicity, we assume that the retention interval is chosen in such a way (e.g., from an exponential distribution) that the reward function is stationary. The specifics of the reward function inevitably depend on the demands of the task and are usually implicit in the experiment's design and feedback mechanism. For example, tasks using the "continuous partial report paradigm" require participants to hold information in mind for a fixed duration, e.g., 2000 ms, with reward provided in proportion to the similarity between the participant's response and the true value. Other tasks provide all-or-none feedback.

We will consider three reward functions relevant to the goals of a memory maintenance system. The first applies to tasks with an all-or-none design in which the memorizer receives full credit for having remembered enough about the cued memory to access it (i.e., having at least k quanta assigned to it at the time of the test, where k is the strength of the weakest accessible memory) and otherwise receives no reward. This reward function is appropriate when scoring performance using a high-threshold model [29, 30], considering only the probability of remembering while ignoring accuracy. In the second, the memorizer is rewarded for having at least one sufficiently strong memory (i.e., one with greater than some threshold number of quanta), but where remembering something about everything is unnecessary. In the third, there is an imbalance across memories in the reward given for remembering them: some are more valuable than others.

4 Maintenance policies, optimality, and approximations

The Markov decision process is a general framework for describing the problem of sequential decision making, but it does not specify the particular strategy used by the agent to make a decision. That strategy is defined by a policy, a function that specifies an action (or probability distribution over actions) for each possible state. Much of modern research on MDPs focuses on finding the optimal policy, one that maximizes the (possibly time-discounted) reward.

The simplest maintenance policies do not depend on the current state of memory. Rather, they produce the same behavior in every state. Borrowing terminology from game theory, in which a player can adopt a strategy that does not depend on the behavior of the opponent (e.g., a player in the Prisoner's Dilemma who always defects), we call these maintenance policies *unconditional* [31]. An example of an unconditional maintenance policy is `all-i`, which always selects the ith quantum as the target of maintenance. A second unconditional strategy is `random`, which selects a target at random, uniformly over all quanta — this maintenance policy is equivalent to a neutral Moran process.

Conditional policies, in contrast, depend on the state (e.g., a player in the Prisoner's Dilemma who plays tit-for-tat, responding to cooperation with cooperation and to defection with defection). In the context of memory maintenance, consider for example the strategy `all-j`, which selects a quantum uniformly from among those assigned to memory *j* if one exists, otherwise choosing randomly among all the quanta.

The optimal policy is conditional. Using linear programming, we computed the optimal policy for a time-discounted variant of the above MDP under each of the reward functions described above, setting $N = 10$, $K = 3$, and the discount factor to 0.99. The optimal policy is different under each reward function, reflecting the differing demands of the task. When the reward function encourages having at least one highly-stable memory, the optimal policy tends to maintain memories that are already stable, preferring to select a quantum assigned to a memory with an above-median allocation of quanta 64% of the time. In contrast, when the reward function encourages good performance on the task, which requires storing more than just one memory, the optimal policy tends to maintain memories that are least stable, preferring to select a quantum assigned to a memory with an above-median allocation of quanta only 29% of the time. When the reward function encourages prioritization of a particular memory, the optimal policy deterministically maintains that memory so long as it has not fully degraded, in which case it chooses randomly among the others — this is the `all-j` maintenance policy described above. At a minimum, then, any cognitive implementation of memory maintenance must be able to selectively maintain memories according to their strength and according to their identity.

The optimal policy can be approximated by a simple strategy that rests on plausible cognitive mechanisms, inspired by a psychological principle known as Luce's choice axiom [32, 33]. According to the axiom, when faced with a choice among alternatives, a decision-maker will exhibit 'matching behavior', selecting options with probability proportional to their value. Matching behavior was originally studied in the context of learning theory, where value is defined as the expected reward [34, 27]. Thus if two levers offer rewards in a ratio of 2:1, an individual that displays matching behavior will press the more rewarding lever twice as often. Here, value is akin to memory strength and is defined by the number of quanta dedicated to a memory.

In practice, it is common to consider a generalization of matching behavior in which a real-valued parameter L determines the decision-maker's sensitivity to the signal. In this so-called "softmax" generalization of matching behavior, the probability of selecting option a from the set of alternatives A is given by

$$p(a) = \frac{v(a)^L}{\sum_{b \in A} v(b)^L},$$

where $v(x)$ is the strength of the signal generated by x and where L determines the decision maker's sensitivity to the signal [27, 35].

Five values of *L* are particularly significant for the process of memory maintenance. When $L = 0$, the process is unconditional (i.e., insensitive to the signal). This corresponds to a neutral process. When $L = 1$, the process gives preference to objects in proportion to how strongly they are currently represented. When $L \to \infty$, the winner takes all. In contrast, when $L = -1$, the process gives preference to objects in proportion to how *weakly* they are currently represented, and in the limit $L \to -\infty$, the loser takes all.

The Luce family of policies can be extended to give graded preference to certain memories over others. To do this, we first define a priority function *f* that assigns a score to each memory. For example, memories A, B, and C may receive scores of 4, 3, and 1, meaning that A has 4× the priority of C and B has 3× the priority of C. Quanta are selected with probability proportional to the

priority score of the memory to which it is assigned. For a system with N quanta, of which n_{A} are assigned to memory A, n_{B} to B, and n_{C} to C, the probability of selecting a quantum q that is of type j is given by

$$\Pr(q) = \frac{f(j)}{\sum_{j \in \{\mathrm{A,B,C}\}} f(j) n_j}.$$

This is equivalent to adding selective pressures to the neutral process and allows for prioritization and graded directed-remembering.

5 Reproducing the behavioral phenomena

The `Luce` family of maintenance policies, which approximate the optimal policy for the MDP defined in Section 4, can reproduce the effects of prioritization, directed forgetting, and self-directed remembering in a single model. We simulated performance of a memorizer who uses the `Luce` family of maintenance policies in a directed-remembering task (Fig. 1a), a priority-based graded directed-remembering task (Fig. 2a), and a self-directed remembering task (Fig. 1c). The simulated observer shows all three forms of maintenance behavior described in Section 2. In the directed-remembering task, the agent selectively maintains the target memory at the expense of the others (Fig. 1a). In the priority-based graded directed-remembering task, the agent maintains the target memories better than the non-target memory, while devoting more resources to the target memory with the higher priority (Fig. 1b). In the self-directed remembering task, the agent selectively maintains the least-stable memory and remembers more because of it (Fig. 1c).

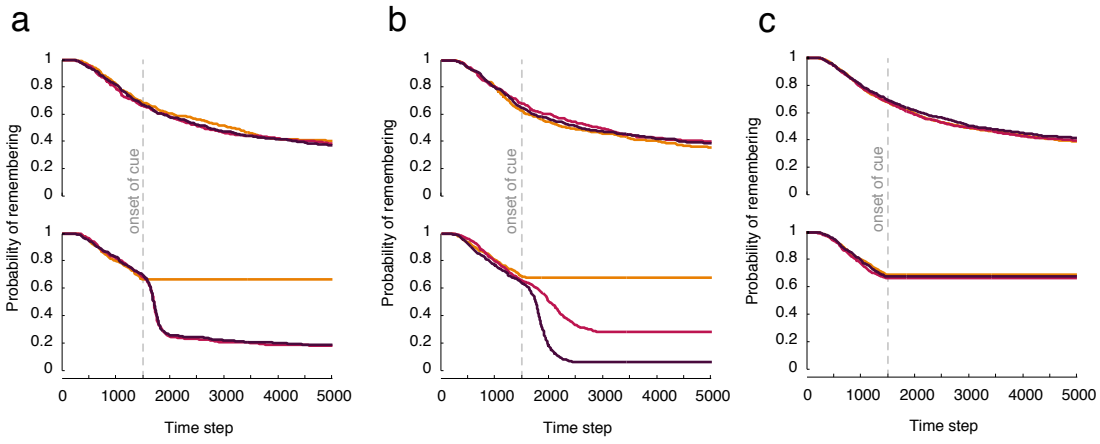

Figure 1: Reproducing directed remembering, priority-based directed remembering, and self-directed remembering by simulating the `Luce` conditional maintenance policy. Panels show forgetting functions for each of three objects that were presented (purple, yellow, and red lines), averaged over 10,000 trials. A forgetting function tracks the amount of information stored in memory as it falls over time. The dashed vertical line marks the onset of the cue. In the upper panels, the participant uses the `random` policy, which gives equal priority to all three objects. Before the cue, the participants behave identically. After the cue, the behavior diverges. In Panel A (directed remembering), performance is better for the target object (yellow) than for the other two. In Panel B (priority-based directed remembering), performance is better for the high-priority memory (yellow) than for the low-priority memory (red), and worst for the non-target memory (purple). In Panel C (self-direct remembering), performance is improved because maintenance is directed to whichever memory is currently least stable. Simulations were run with parameters $N = 64$ and $K = 3$ for 5000 steps. The cue appeared at time step 1500.

6 Partially observable minds

The framework of a Markov decision process makes a strong commitment to the accessibility of the memory state to the memory maintenance system: it assumes perfect, real-time, no-cost metamemory. However, metamemory is imperfect [15, 18].

By generalizing the MDP to a partially-observable world, we can accommodate situations of imperfect or costly metamemory. A partially observable world is one in which the agent does not know exactly what state it is in, making it impossible to directly carry out conditional policies that depend on the state. Often the agent has available some instrument (a "sensor") for measuring or sensing the state. In the case of memory maintenance, the sensor is metamemory. The agent uses the sensor to update its beliefs about the state. Thus the *partially observable Markov decision process* (POMDP) extends the MDP through the introduction of a sensor model, which describes the information about the state that is provided by each observation, and a belief state, which is a probability distribution over the state space that embodies the agent's beliefs about the current state [36, 37]. The Dirichlet distribution is a convenient representation of uncertainty about the state of memory resource allocation because it is the conjugate prior for multinomial data.

In a partially observable mind, inefficiencies of metamemory limit the efficacy of flexible maintenance behaviors. This is because in a world where the future depends on the past, one who does not even know the present cannot suitably plan for what is to come. We demonstrate this dependence by defining a simple metamemory agent and then simulating its behavior with different levels of efficiency. Metamemory observations made by the agent come in the form of object labels sampled with probability proportional to their strength (that is, the number of quanta assigned to them). This defines the sensor model. The agent is initially unaware of the allocation of the commodity, represented by a belief state initially set to a Dirichlet distribution with concentration parameters 1, 1, and 1, which is equivalent to a uniform distribution over all possible allocations. At each time step, the agent makes *m* observations. We assume that the metamemory system has no memory of its own and thus considers only the observations made at the current time step (see below for a brief discussion of optimal filtering, in which the metamemory system also considers past observations). To avoid the problems caused by sampling zero quanta of a certain type, we use additive smoothing by adding one to all the counts. These counts are used by the Luce policy, with exponent 1. The efficiency of metamemory can be varied by altering the number of observations made at each time step. This formulation makes it possible to vary efficiency between two extremes. At one extreme, $m = 0$ and the agent gains no information about the state. At the other extreme, in the limit $m \to \infty$, the agent has perfect information about the state. Intermediate efficiencies lead to intermediate performance (Figure 2).

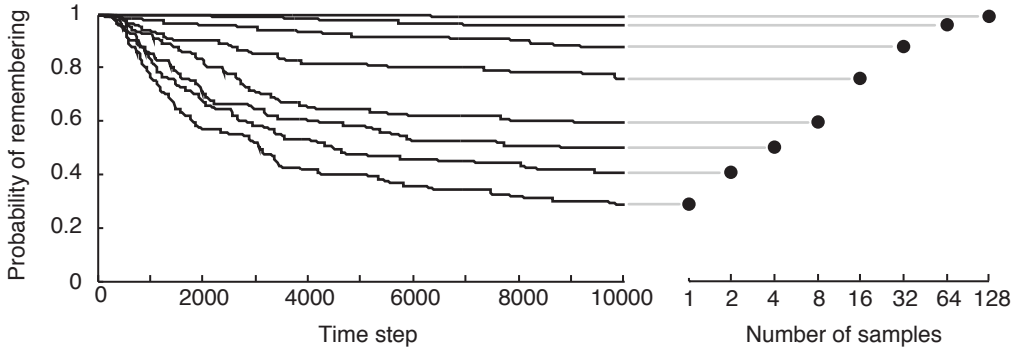

Figure 2: Inefficiencies of metamemory limit the efficiency of memory maintenance. On the left are forgetting functions for a simulated agent whose memory is only partially observable. At each time step the agent draws *m* quanta (with replacement) and observes their assignment. Selection happens according to the procedure in the main text. On the right, performance increases with the number of samples taken. Simulations were run with settings $N = 128$, $K = 12$, and $L = -1$.

7 Discussion

In this paper, we approached problem of memory maintenance by demonstrating an equivalence to a Markov decision process in which, at each moment, a cognitive control mechanism selects a memory as the target of maintenance. The challenge of remembering is then finding a maintenance policy well-suited to the task at hand. We computed the optimal policy under various conditions and defined plausible cognitive mechanisms, embodied by the `Luce` policy, that can approximate these optimal policies. Framing the problem of maintenance in this way makes it possible to capture in a single model many of the essential behavioral phenomena of memory maintenance, including directed remembering, priority-based directed remembering, and self-directed remembering. Finally, we considered the case of imperfect metamemory — where the current state of memory is only partially observable — and show that the fidelity of metamemory determines the effectiveness of maintenance.

Perhaps the biggest payoff that comes from framing the problem of memory maintenance in this way is the set of new questions that it makes possible to ask.

For example, one might ask where maintenance policies come from. Specifically, how are they learned? Methods such as temporal difference learning have emerged as candidate learning mechanisms used in the brain to learn policies that guide behavior, and it has become popular to relate this particular class of learning algorithms to known reward circuitry in the brain [38, 39, 40]. Particularly relevant is the work of [41], who discuss methods for learning to use working memory by temporal difference methods. Specifically, they showed how temporal difference learning can be used to shape representations in the prefrontal cortex so that they are useful for working memory [41]. Also relevant is the work of [42], who developed an "actor/critic" model of the neural substrates of working memory and cognitive control. They showed that an active gating mechanism that controls the contents of working memory can be learned through learning mechanisms from reinforcement learning [42].

Finally, it may be useful to consider other resource allocation tasks that are similar in structure to that of memory maintenance — e.g., scheduling and queuing. Much of the original work on these problems came from the field of operations research, which originated from military planners in WWII and which today considers the optimal solutions to decision making and resource allocation tasks in a variety of settings, often in the context of organizational behavior [43] or electronic systems [44, 45]. Having made the link to these related problems, it may be fruitful to consider known solutions as candidate psychological mechanisms. For example, queuing theory is a set of tools for considering resource allocation tasks that feature the continuous arrival of entities that require the resource (e.g., callers to a company's customer support center) [46]. Most of the popular working memory tasks are episodic, with information arriving all at once and then being discarded at the end of the trial. Our visual experience is not always so episodic; rather, it is sometimes necessary to update the contents of working memory with new information or redirecting maintenance in light of new goals [47, 48]. Looking towards queuing theory, for example, may provide insight into this problem of maintenance in the face of continuously-arriving information.

References

[1] A. Baddeley, "Working memory," *Science*, vol. 255, pp. 556–559, 1992.

[2] N. Cowan, *Working memory capacity*. Psychology Press, 2005.

[3] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychological Review*, vol. 63, pp. 81–97, March 1956.

[4] D. Kahneman, *Attention and effort*. Prentice Hall, 1973.

[5] A. R. Dobbs and B. G. Rule, "Adult age differences in working memory.," *Psychology and Aging*, vol. 4, no. 4, pp. 500–503, 1989.

[6] A. R. A. Conway, M. J. Kane, and R. W. Engle, "Working memory capacity and its relation to general intelligence," *Trends in Cognitive Sciences*, vol. 7, no. 12, pp. 547–552, 2003.

[7] M.-R. Steenari, V. Vuontela, E. J. Paavonen, S. Carlson, M. Fjällberg, and E. T. Aronen, "Working memory and sleep in 6-to 13-year-old schoolchildren," *Journal of the American Academy of Child & Adolescent Psychiatry*, vol. 42, no. 1, pp. 85–92, 2003.

[8] P. S. Goldman-Rakic, "Working memory dysfunction in schizophrenia.," *The Journal of Neuropsychiatry and Clinical Neurosciences*, vol. 6, pp. 348–357, 1994.

[9] J. Jonides, R. L. Lewis, D. E. Nee, C. A. Lustig, M. G. Berman, and K. S. Moore, "The mind and brain of short-term memory," *Annual Review of Psychology*, vol. 59, pp. 193–224, 2008.

[10] W. S. Muther, "Erasure or partitioning in short-term memory.," *Psychonomic Science*, vol. 3, pp. 429–430, 1965.

[11] R. A. Bjork, D. Laberge, and R. Legrand, "The modification of short-term memory through instructions to forget," *Psychonomic Science*, vol. 10, pp. 55–56, 1968.

[12] C. N. Macrae, G. V. Bodenhausen, A. B. Milne, and R. L. Ford, "On regulation of recollection: The intentional forgetting of stereotypical memories," *Journal of Personality and Social Psychology*, vol. 72, no. 4, pp. 709–719, 1997.

[13] D. M. Wegner, "How to think, say, or do precisely the worst thing for any occasion," *Science*, vol. 325, pp. 48–50, 2009.

[14] M. L. Puterman, *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, Inc., 1994.

[15] J. H. Flavell and H. M. Wellman, "Metamemory," in *Perspectives on the development of memory and cognition* (R. V. Kail and J. W. Hagen, eds.), pp. 3–33, Erlbaum, 1977.

[16] H. M. Wellman, "Tip of the tongue and feeling of knowing experiences: A developmental study of memory monitoring," *Child Development*, vol. 48, pp. 13–21, 1977.

[17] A. S. Brown, "A review of the tip-of-the-tongue experience.," *Psychological Bulletin*, vol. 109, no. 2, pp. 204–223, 1991.

[18] J. E. Metcalfe and A. P. Shimamura, *Metacognition: Knowing about knowing.* The MIT Press, 1994.

[19] J. K. Pannu and A. W. Kaszniak, "Metamemory experiments in neurological populations: A review," *Neuropsychology Review*, vol. 15, no. 3, pp. 105–130, 2005.

[20] A. E. Woodward and R. A. Bjork, "Forgetting and remembering in free recall: Intentional and unintentional.," *Journal of Experimental Psychology*, vol. 89, no. 1, pp. 109–116, 1971.

[21] D. M. Wegner and R. Erber, "The hyperaccessibility of suppressed thoughts.," *Journal of Personality and Social Psychology*, vol. 63, no. 6, pp. 903–912, 1992.

[22] M. Hartman and L. Hasher, "Aging and suppression: Memory for previously relevant information," *Psychology and Aging*, vol. 6, no. 4, pp. 587–594, 1991.

[23] E. Vergauwe, P. Barrouillet, and V. Camos, "Visual and spatial working memory are not that dissociated after all: A time-based resource-sharing account.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 35, no. 4, pp. 1012–1028, 2009.

[24] H. Bowman and B. Wyble, "The simultaneous type, serial token model of temporal attention and working memory.," *Psychological Review*, vol. 114, no. 1, pp. 38–70, 2007.

[25] J. W. Suchow, B. Allen, M. A. Nowak, and G. A. Alvarez, "Evolutionary dynamics of visual memory," *Journal of Vision*, vol. 13, p. 20, 07 2013.

[26] P. A. P. Moran, "Random processes in genetics," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 54, no. 1, pp. 60–71, 1958.

[27] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction).* The MIT Press, 1998.

[28] W. J. Ewens, *Mathematical population genetics: I. Theoretical introduction*, vol. 27. Springer, 2004.

[29] S. J. Luck and E. K. Vogel, "The capacity of visual working memory for features and conjunctions," *Nature*, vol. 390, pp. 279–281, 1997.

[30] N. Cowan, "The magical number 4 in short-term memory: A reconsideration of mental storage capacity," *Behavioral and Brain Sciences*, vol. 24, no. 01, pp. 87–114, 2001.

[31] R. Axelrod and W. D. Hamilton, "The evolution of cooperation," *Science*, vol. 211, pp. 1390–1396, 1981.

[32] R. D. Luce, *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.

[33] R. J. Herrnstein, "Relative and absolute strength of response as a function of frequency of reinforcement," *Journal of the Experimental Analysis of Behavior*, vol. 4, no. 3, p. 267272, 1961.

[34] W. K. Estes, "Of models and men.," *American Psychologist*, vol. 12, no. 10, pp. 609–617, 1957.

[35] E. Vul, *Sampling in human cognition*. PhD thesis, Massachusetts Institute of Technology, 2010.

[36] G. E. Monahan, "State of the art—a survey of partially observable Markov decision processes: Theory, models, and algorithms," *Management Science*, vol. 28, no. 1, pp. 1–16, 1982.

[37] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996.

[38] J. R. Hollerman and W. Schultz, "Dopamine neurons report an error in the temporal prediction of reward during learning," *Nature Neuroscience*, vol. 1, no. 4, pp. 304–309, 1998.

[39] J. P. O'Doherty, P. Dayan, K. Friston, H. Critchley, and R. J. Dolan, "Temporal difference models and reward-related learning in the human brain," *Neuron*, vol. 38, no. 2, pp. 329–337, 2003.

[40] P. Dayan and Y. Niv, "Reinforcement learning: the good, the bad and the ugly," *Current Opinion in Neurobiology*, vol. 18, no. 2, pp. 185–196, 2008.

[41] M. T. Todd, Y. Niv, and J. D. Cohen, "Learning to use working memory in partially observable environments through dopaminergic reinforcement," in *Advances in Neural Information Processing Systems*, pp. 1689–1696, 2008.

[42] R. C. O'Reilly and M. J. Frank, "Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia," *Neural Computation*, vol. 18, no. 2, pp. 283–328, 2006.

[43] H. A. Taha, *Operations research: An introduction*. Pearson/Prentice Hall, 2007.

[44] K. J. Åström and B. Wittenmark, *Computer-controlled systems: Theory and design*. Courier Dover Publications, 2011.

[45] A. Silberschatz, P. B. Galvin, and G. Gagne, *Operating system concepts*, vol. 8. Wiley, 2013.

[46] L. Kleinrock, *Queueing systems. Volume 1: Theory*. Wiley, 1975.

[47] A. Sandberg, A. Lansner, K. M. Petersson, and Ö. Ekeberg, "A palimpsest memory based on an incremental Bayesian learning rule," *Neurocomputing*, vol. 32, pp. 987–994, 2000.

[48] L. Matthey, P. Bays, and P. Dayan, "Probabilistic palimpsest memory: Multiplicity, binding and coverage in visual short-term memory," in COSYNE, pp. III–48, 2012.