# The Predictive Brain Must Have a Limitation in Short-Term Memory Capacity

Sabrina Trapp[1,2] , Thomas Parr[3], Karl Friston[3], and Erich Schröger[1]

[1]Institute of Psychology, University of Leipzig; [2]Institute of Sport Science, Bielefeld University; and [3]Wellcome Centre for Human Neuroimaging, Institute of Neurology, University College London

## Abstract

Traditionally, short-term memory (STM) has been assessed by asking participants to remember words, visual objects, or numbers for a short amount of time before their recall or recognition of those items is tested. However, this focus on memory for past sensory input might have obscured potential theoretical insights into the function of this cognitive faculty. Here, we suggest that STM may have an important role in predicting future sensory input. This reconceptualization of STM may provide a functional explanation for its capacity limitation.

The ability to maintain information over the short term is probably a hallmark of human cognition because it is required for various domains, such as perception, action planning, and language. Humans cannot think, reason, or calculate, and possibly cannot even perceive, without this core functioning of the brain. Whereas the concept of short-term memory (STM) usually refers solely to the storage of information, working memory is proposed to involve additional executive processes, such as integration and manipulation of information, and is considered to support complex cognitive activities, such as language processing, reasoning, and problem solving (Baddeley, 2003).

Empirically, STM is examined using paradigms such as the delayed-match-to-sample and change-detection tasks, in which participants are asked to encode a given sensory input in order to eventually recognize or recall it at a subsequent time point. These and other related tasks all have revealed a robust, yet puzzling, finding, namely, that humans are severely constrained in the amount of information they can keep in STM (or mentally manipulate in working memory). On average, it is estimated that humans can keep just four items in STM (Cowan, 2001).[1]

Traditionally, the focus has been on addressing the mechanisms underlying this limitation. Cognitive processes such as decay or interference of information (Barrouillet & Camos, 2009; Oberauer & Lin, 2017) and subcycles in brain oscillation (Lisman & Idiart, 1995) have been identified as putative mechanisms that limit STM. To date, only a few authors have proposed a functional account for the STM limitation, suggesting, for example, that it improves the efficiency of memory search (e.g., Dirlam, 1972; MacGregor, 1987), benefits language acquisition (e.g., Elman, 1993), facilitates detection of covariation in the environment (Kareev, 2000), or serves as a device for action control (Heuer et al., 2020) or as a core component of the eye movement system (Van der Stigchel & Hollingworth, 2018).

According to Marr (1982), complex information-processing systems like the human brain can be analyzed at three different levels: (a) the *computational* level, at which the problem the system must solve is specified; (b) the *algorithmic* level, at which the means of solving the identified problem (and the representational vehicles involved) are described; and (c) the *implementation* level, at which the operation of the system's physical substrate (e.g., neurons and synapses) is described. Here, we consider STM's capacity limitation from a functional, or computational, point of view.

**Corresponding Author:**
Sabrina Trapp, Institute of Sport Science, Bielefeld University
Email: Sabrina.trapp@uni-bielefeld.de

Therefore, we are agnostic with regard to the actual cognitive and neural mechanisms implementing the limit. Rather, we explore the problem the brain is confronted with and how a limitation in STM may serve to solve this problem.

## The Predictive Brain

In the past decade, the interest in *predictive processes* has grown dramatically (for a recent review, see Clark, 2013). From this perspective, the brain is proactive: It does not merely rely on bottom-up information, but also adds expectations during the perceptual inference process. Several empirical studies have provided evidence for a brain that predicts its sensory input (e.g., Alink et al., 2010). Large-scale frameworks, such as the free-energy principle, suggest that minimizing surprise can give rise to brain functions such as perception, memory, and action (Friston, 2010). The concept of a predictive brain invited novel functional accounts for core faculties of the human mind: For example, from this perspective, attention is not primarily conceived as a selection process that is required because the system would otherwise drown in sensory overflow, but rather is considered as a mechanism that assigns weights to predictions and prediction errors (Friston, 2010). In a similar vein, long-term memory (LTM), traditionally considered as a device that stores past information, has been reconsidered as a system that provides the prerequisites for the simulation of future events (Schacter & Addis, 2007). The future-first hypothesis even posits that "our ability to revisit the past may be only a design feature of our ability to conceive of the future" (Suddendorf & Corballis, 2007, p. 303). But can the concept of the predictive brain offer something for understanding STM?

## Prediction and Short-Term Memory

How could STM be linked to the process of prediction? The key idea of predictive frameworks is that incoming sensory input is compared with prior information or expectations, and only deviations, or *prediction errors*, are fed forward in the sensory hierarchy (see Clark, 2013). Although many studies have investigated the neural correlates of prediction and processing of prediction error, and fitted behavioral data with predictive models, there has been little discussion of related psychological or cognitive processes, such as how sensory information is compared with prior expectations, how priors are retrieved from LTM and made accessible during a comparison with sensory input, and what their representational format is. One option is that the sensory input is compared with *all* prior information stored in LTM. Given the high speed with which perception is accomplished, this seems rather unlikely.

Alternatively, the brain could use a mechanism that highlights and temporarily maintains the part of prior knowledge from LTM that is relevant for the perceptual decision or task at hand. In most cognitive models, STM is considered elevated or highlighted information from LTM, that is, activity that tags a selected part as relevant for current processing (e.g., Ruchkin et al., 2003). Thus, STM could in principle provide optimal machinery for predicting sensory input, as it could maintain relevant priors in a highly accessible state for subsequent comparison processes. There is some empirical evidence that STM is required for predictions. In one relevant study, Travis and colleagues (2013) used a visual search paradigm called *contextual cuing*, in which participants search through seemingly random layouts for a specific target whose identity remains the same throughout the experiment. Unbeknownst to them, some layouts are repeated several times during the experiment, and participants learn to find the position of the target faster in these layouts than in others. Travis et al. found that when participants had to maintain additional information in STM while performing this visual search task, the facilitation was impeded, which indicates that STM is required for predicting an object's location. Furthermore, Cashdollar et al. (2017) demonstrated that STM capacity was correlated with the neural correlates of prediction in a task that required participants to attend to a series of images. These images had a probabilistic distribution that allowed for the prediction of some of them. Participants' STM capacity correlated with neural correlates of the anticipation process, which indicates that STM has a role in predicting object identities as well.

The idea that STM is relevant for predictive processes is also supported by research from the domain of action planning and execution. The term *working memory* was originally introduced in the context of planning behavior, to refer to a system in which "plans can be retained temporarily when they are being formed, or transformed, or executed" (Miller et al., 1960, p. 207). It has been suggested that "working memory is 'nothing more' than the preparation to perform an action, whether it be oculomotor, manual, verbal, or otherwise" (Theeuwes et al., 2009, p. 198). Specifically, working memory is considered important for storing both the movement-related visuospatial information and action plans (Postle, 2006). A recent study combining electroencephalography and functional MRI demonstrated preactivation of motor areas when the visual content retained in working memory was linked to specific actions to be performed in the future, thus supporting the idea that working memory plays a role in action planning (van Ede, 2020). The link between working memory and

action planning is highly consistent with the principle of minimizing surprise and prediction error, because an important cause of sensory data is the course of action pursued by the organism that experiences these data. If humans predict their sensations on the basis of their planned actions, they must represent the data generated by their courses of action both in the past and in the future. In the following, we argue that the limitation on the capacity of STM can be explained if one views STM as a derivate of a prediction device.

## Predictions Must Be Capacity Limited

To make accurate predictions, the brain must represent all uncertainties of events. However, a key feature of the dynamic natural environment is that predictions become generally more uncertain the further one predicts into the future. Therefore, the costs associated with the representation of uncertainties of predictions rise dramatically with temporal distance: As in a chess game, the number of possible configurations (i.e., states of the environment) rises exponentially with the number of moves to be predicted. For a person to make an informed decision about the next move, all these possibilities and their uncertainties have to be represented.

Events in the natural environment are inherently stochastic in that the probability of one event depends on that of the previous event. For instance, if a man one has just encountered reaches out his hands, the probability that his gesture is friendly depends on whether he previously displayed a friendly facial expression; if instead his facial expression suggested aggression, the probability that his reaching hands signal an act of violence is much higher. To bypass this inherent unpredictability of future events and their stochastic dependence, it would be optimal to use a limited window of prediction. Ideally, this window should not end too close to the present, because then it could not efficiently guide perception and action, but neither should it extend too far into the future, because then the representations would become too costly (computationally complex). Figure 1 shows the number of possible events as a function of temporal distance from the present into the future in a stochastic environment. After only a few time steps, predictions into the future (a) become too costly because of the need to represent all possible events and (b) become mostly useless because even the most probable event occurs with low probability. Consequently, an optimal strategy for the brain is to predict only short sequences that are long enough to allow for adjusting behavior but short enough to avoid both the exponential explosion of possibilities and the exponential decay in the probability of the most likely sequence.

The basic idea can be expressed more technically in terms of the balance between accuracy and complexity in representing trajectories of states in the world. In brief, current formulations of active inference and predictive processing that rest on sequences or successions of discrete states (e.g., using Markov decision processes) consider a small number of discrete steps from the beginning of a time epoch to the end. As time progresses, these sequential representations play the dual roles of STM and prospective memory (i.e., representations of the short-term past and future). These representations endow a generative model with temporal depth and are necessary to evaluate the quality of an action in terms of anticipated outcomes. Crucially, these sorts of generative models can be optimized in relation to their evidence (or variational free-energy bound). Because the evidence is the difference between accuracy and complexity, there must be an optimal number of time steps into the past (or future) to maintain a given degree of prediction accuracy. If there are too many time steps, then the degrees of freedom in the generative model will increase its complexity in a way that is not balanced by an accompanying increase in accuracy, because future states of the world become increasingly uncertain.

It is interesting to note that nearly all the current simulations using these deep temporal models (with STM) use a handful of steps into the future—usually between four and six (Friston et al., 2017). Furthermore, the continuous time homologue corresponds to the order of generalized coordinates of motion (effectively, a Taylor series expansion of a trajectory) in continuous-state generative models of the sort used by predictive coding: Generalized motion is usually truncated at about fourth-order motion (Friston, 2008). In summary, there must be an optimum number of discrete events (or orders of motion) for representing a dynamic world, depending on the volatility of this world and the degrees of freedom used by generative models. Anecdotally, it seems that four to six appears to be a universal range for this order.

By saying that the capacity limitation of STM is adaptive, we mean that it results from an optimal trade-off between the constraints on the complexity of computation in biological systems and the accuracy benefits. As the accuracy benefits saturate (i.e., reach a ceiling) with increasing STM capacity, further increases in capacity carry additional computational cost (complexity) without benefit. This means that a capacity limit is to be expected according to the proposed functional role of STM.

This logic applies even for future events that are not sequential, such as a visual event at one point in time. In the dynamic environment, fine details in the visual scenery, such as the visual angle of an object in space,
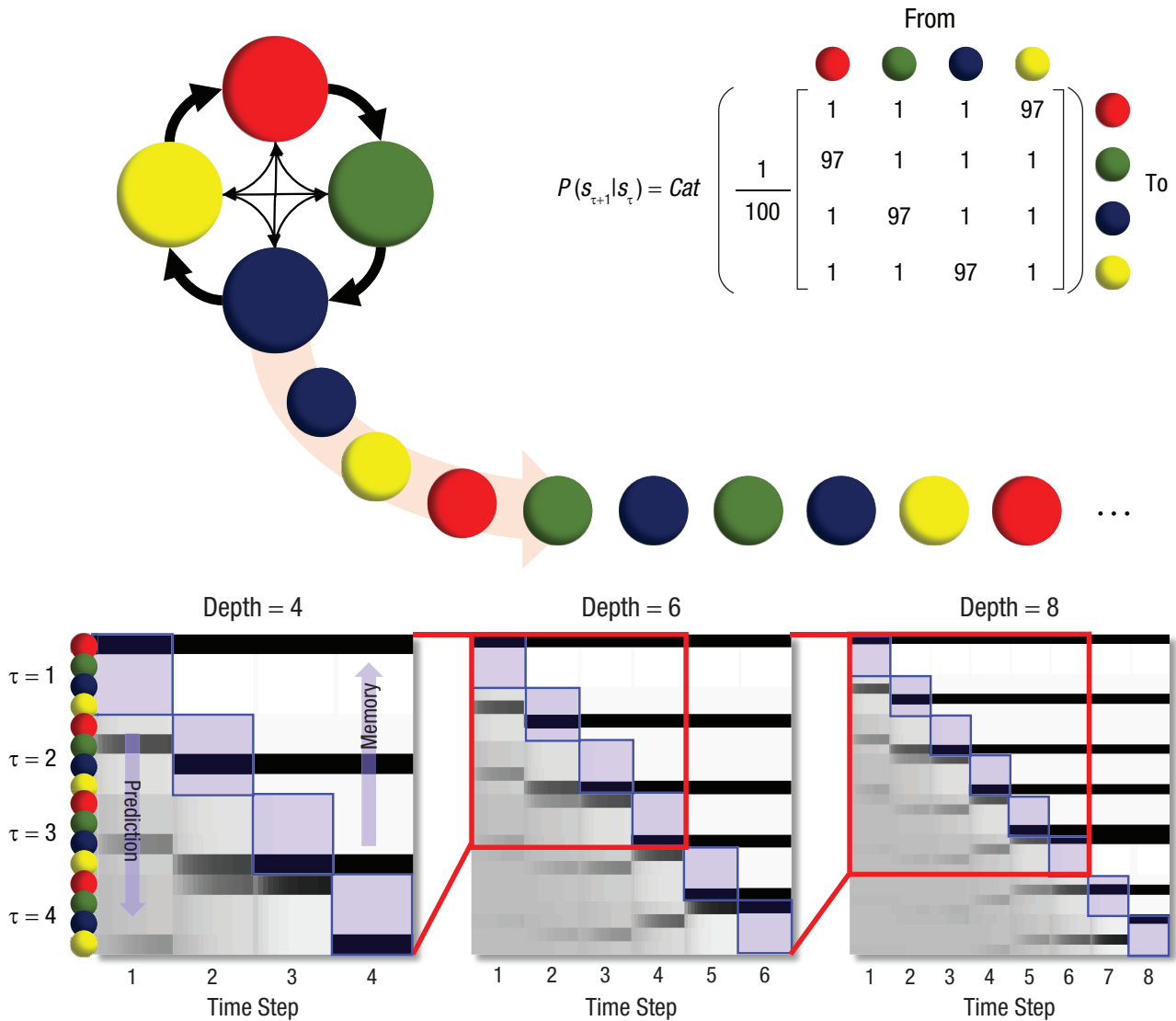
From

$$P(s_{\tau+1}|s_\tau) = Cat\left(\frac{1}{100}\begin{bmatrix} 1 & 1 & 1 & 97 \\ 97 & 1 & 1 & 1 \\ 1 & 97 & 1 & 1 \\ 1 & 1 & 97 & 1 \end{bmatrix}\right)$$

To

Depth = 4    Depth = 6    Depth = 8

$\tau = 1$

$\tau = 2$

$\tau = 3$

$\tau = 4$

Memory

Prediction

Time Step

Time Step

Time Step

**Fig. 1.** Temporal depth, prediction, and short-term memory span. This figure summarizes the key idea in this article using a simple example in which an experimental trial consists of a sequence of colored balls. The sequence is generated stochastically, according to the categorical (*Cat*) probability transition matrix at the upper right, which shows how the probability of one color ($s_{\tau+1}$) depends on that of the previous color in the sequence ($s_\tau$). This dependence is illustrated schematically at the upper left; most (97%) of the time (thick arrows), the red ball is followed by the green, then blue, then yellow balls. However, 3% of the time (thin arrows), these rules are violated. This construction is a simple example of a situation in which the future has some degree of predictability but can deviate from what has been predicted. The three plots in the lower part of the figure illustrate the inferences a subject could draw if the subject's internal model incorporates the statistics of this sequence. (See Friston et al., 2017, for details of the construction of these plots.) Each column in these plots represents a time point during the trial (i.e., the most recent position in the sequence). The rows indicate the alternative colors that might be present at each of those times. The shading of each cell represents the neural encoding of a posterior probability of the indicated color at the indicated time point. White represents a probability of 0, and black a probability of 1; darker shades of gray indicate increasing probabilities. The cells along the diagonals (highlighted in blue) represent beliefs about the present. Cells below the diagonals represent predictions about the future, and those above the diagonals represent short-term memory. Our key message is manifest in two ways in these plots: First, as time progresses (*x*-axis), those units that were predicting the future become representations of past states. This implies an equivalence between the predictive depth at the start of the trial and short-term memory capacity at its end. Second, each of the three plots represents a different temporal depth, or horizon, that could be used to represent the sequence. The first columns show that at Time Step 1, it is possible to make meaningful predictions (even if they turn out to be wrong) about Time Step 4 (lower left plot) and less confident predictions about Time Step 6 (lower middle plot), but no confident predictions about Time Step 8 (lower right plot). One can remove from the model the time steps for which no confident predictions can be made and select a model with a more efficient depth of four to six time steps. Given that the predictive depth at the start of the trial and short-term memory capacity at the end of the trial are equivalent, truncating the model's temporal depth also truncates the short-term memory span.

tend to change at a fast timescale, whereas more abstract information, such as the identity of an object, does not change, or changes more slowly. The temporal architecture of visual memory may map onto the scale of predictability: It is well established that iconic memory can store fine detail but has a memory span of only a few hundreds of milliseconds, whereas visual STM can store only around four items, but at a much longer timescale of several seconds. There are also reports of a memory system with an intermediate timescale and a larger memory capacity than classical visual STM (Sligte et al., 2008). The temporal hierarchy of visual memory might mirror the predictability of natural visual input at different timescales.

This separation of timescales is ubiquitous in hierarchical generative models, in which higher levels predict short trajectories at lower levels. Reading provides an intuitive example of this: Sentences predict short sequences of words, which themselves predict short sequences of letters (Friston et al., 2018). This setup has also been used in simulations of working memory tasks involving delay periods to account for sustained representation of an item that transcends the time period for which it is visible (Parr & Friston, 2017). The same kind of structure may support mnemonic strategies such as chunking, in which a sequence of short sequences, often around four to six items long, is used to remember a long sequence (e.g., Mathy & Feldman, 2012).

One may argue that STM is not necessarily used for prospective processes only. However, all experimental tasks calling on STM ultimately require participants to maintain information for comparison with future sensory information (the probe) and/or for prospective actions (the response). From this perspective, an STM task constitutes a predictive-processing problem. Information that is relevant for perceptual inference in the present or near future is extracted from the environment and then compared with incoming sensory information, and an internal inference and/or an external action is executed to resolve any discrepancy.

In addition, the structure of the brain has been conserved over the course of evolution, and new structures are assumed to be the result of shifts in axonal projection patterns, that is, differentiation of existing structures (Krubitzer & Kaas, 2005). It is therefore possible that structures or mechanisms previously developed for one purpose, such as predictive processing, may be used for others, such as memorizing over the short term, over a phylogenetic timescale. However, a system that serves any novel purpose will necessarily be endowed with features that served the original purpose, and those features may seem irrelevant when co-opted by other processes. Consider the reconceptualization of LTM mentioned earlier. If one considers LTM as a structure for the conservation of the past, humans' liability to false memories seems suboptimal. There is a large literature on how humans construct new, fictive episodes by adding new information after the encoding phase of the original exposure (for a review, see Loftus, 2005). However, if LTM developed for using past information to simulate *future* episodes, such errors make more sense, as they can be considered a side effect of the flexibility of LTM in composing novel episodes (Schacter & Addis, 2007). The same idea holds true for STM. STM can be placed in the service of inferring both the past and the future, but its capacity limitation makes most sense if one looks at STM from the perspective of a predictive brain.

## Implications for Further Theorizing and Research

We have presented a computational perspective on STM's capacity limitation. That is, we have focused on the problem the system has to solve, namely, enabling the most accurate and confident predictions possible. This perspective reveals that a capacity limitation of STM is to be expected. Our perspective may provide new research avenues that deviate from the current focus on STM as a function that stores past sensory input. Previous work on predictive processes has primarily focused on neural consequences of deviations from expectations (i.e., prediction errors). The question of how the brain prepares for upcoming information has been addressed less extensively, and research in this area may be stimulated by considering STM as activity that represents the top-down predictions that are maintained for subsequent comparison with sensory input. Another avenue for future work would be the assessment of individual differences, that is, whether and how individuals with low versus high STM capacity differ in their predictive abilities on various levels.

### Recommended Reading

Baddeley, A. (2003). (See References). An excellent introduction to working memory research, providing an overview of the concept, its history, and tasks that probe working memory processes, and outlining challenges for the future.
Clark, A. (2013). (See References). A broad overview of the key ideas of predictive processing and the evidence supporting the assumptions of predictive frameworks, along with brief discussions of important frameworks, such as the Bayesian brain and free-energy principle.
Friston, K. (2010). (See References). One of the most accessible introductions to the free-energy principle, outlining the explanatory power of this theory and encompassing both perception and action.

Marr, D. (1982). (See References). An excellent, well-written, and seminal book in which the visual system is conceptualized as an information-processing system.

Schacter, D. L., & Addis, D. R. (2007). (See References). A review that highlights the function of long-term memory for simulating the future and contrasts this functional approach with traditional accounts.

## ORCID iD

Sabrina Trapp [iD] https://orcid.org/0000-0002-5889-0419

## Note

1. It is important to mention, however, that this number is by no means fixed. Recent STM models rely more on variation-in-noise paradigms, in which all items are encoded but noise makes some items encoded more poorly than others (e.g., Schneegans & Bays, 2016; van den Berg et al., 2012).

## References

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., & Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. *Journal of Neuroscience*, *30*(8), 2960–2966.

Baddeley, A. (2003). Working memory: Looking back and looking forward. *Nature Reviews Neuroscience*, *4*(10), 829–839. https://doi.org/10.1038/nrn1201

Barrouillet, P., & Camos, V. (2009). Interference: Unique source of forgetting in working memory? *Trends in Cognitive Sciences*, *13*(4), 145–146. https://doi.org/10.1016/j.tics.2009.01.002

Cashdollar, N., Ruhnau, P., Weisz, N., & Hasson, U. (2017). The role of working memory in the probabilistic inference of future sensory events. *Cerebral Cortex*, *27*(5), 2955–2969.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral & Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral & Brain Sciences*, *24*(1), 87–114. https://doi.org/10.1017/S0140525X01003922.

Dirlam, D. K. (1972). Most efficient chunk sizes. *Cognitive Psychology*, *3*(2), 355–359. https://doi.org/10.1016/0010-0285(72)90012-6

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99. https://doi.org/10.1016/0010-0277(93)90058-4

Friston, K. (2008). Hierarchical models in the brain. *PLOS Computational Biology*, *4*(11), Article e1000211. https://doi.org/10.1371/journal.pcbi.1000211

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Friston, K., FitzGerald, T., Rigoli, F., Schwartenbeck, P., & Pezzulo, G. (2017). Active inference: A process theory. *Neural Computation*, *29*(1), 1–49. https://doi.org/10.1162/NECO_a_00912

Friston, K. J., Rosch, R., Parr, T., Price, C., & Bowman, H. (2018). Deep temporal models and active inference. *Neuroscience and Biobehavioral Reviews*, *90*, 486–501. https://doi.org/10.1016/j.neubiorev.2018.04.004

Heuer, A., Ohl, S., & Rolfs, M. (2020). Memory for action: A functional view of selection in visual working memory. *Visual Cognition*, *28*(5–8), 388–400.

Kareev, Y. (2000). Seven (indeed, plus or minus two) and the detection of correlations. *Psychological Review*, *107*(2), 397–402. https://doi.org/10.1037/0033-295X.107.2.397

Krubitzer, L., & Kaas, J. (2005). The evolution of the neocortex in mammals: How is phenotypic diversity generated? *Current Opinion in Neurobiology*, *15*(4), 444–453. https://doi.org/10.1016/j.conb.2005.07.003

Lisman, J. E., & Idiart, M. A. (1995). Storage of 7 +/− 2 short-term memories in oscillatory subcycles. *Science*, *267*(5203), 1512–1515. https://doi.org/10.1126/science.7878473

Loftus, E. F. (2005). Planting misinformation in the human mind: A 30-year investigation of the malleability of memory. *Learning & Memory*, *12*(4), 361–366. https://doi.org/10.1101/lm.94705

MacGregor, J. N. (1987). Short-term-memory capacity: Limitation or optimization? *Psychological Review*, *94*(1), 107–108. https://doi.org/10.1037/0033-295X.94.1.107

Marr, D. (1982). *Vision: A computational approach*. Freeman & Co.

Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition*, *122*(3), 346–362. https://doi.org/10.1016/j.cognition.2011.11.003

Miller, G. A., Galanter, E., & Pribram, K. H. (1960). *Plans and the structure of behavior*. Holt.

Oberauer, K., & Lin, H. Y. (2017). An interference model of visual working memory. *Psychological Review*, *124*(1), 21–59.

Parr, T., & Friston, K. J. (2017). Working memory, attention, and salience in active inference. *Scientific Reports*, *7*(1), Article 14678. https://doi.org/10.1038/s41598-017-15249-0

Postle, B. R. (2006). Working memory as an emergent property of the mind and brain. *Neuroscience*, *139*(1), 23–38.

Ruchkin, D. S., Grafman, J., Cameron, K., & Berndt, R. S. (2003). Working memory retention systems: A state of activated long-term memory. *Behavioral & Brain Sciences*, *26*(6), 709–728.

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: Remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *362*(1481), 773–786. https://doi.org/10.1098/rstb.2007.2087

Schneegans, S., & Bays, P. M. (2016). No fixed item limit in visuospatial working memory. *Cortex*, *83*, 181–193. https://doi.org/10.1016/j.cortex.2016.07.021

Sligte, I. G., Scholte, H. S., & Lamme, V. A. F. (2008). Are there multiple visual short-term memory stores? *PLOS ONE*, *3*(2), Article e1699. https://doi.org/10.1371/journal.pone.0001699

Suddendorf, T., & Corballis, M. C. (2007). The evolution of foresight: What is mental time travel, and is it unique to humans? *Behavioral & Brain Sciences*, *30*(3), 299–313. https://doi.org/10.1017/S0140525X07001975

Theeuwes, J., Olivers, C. N., & Chizk, C. L. (2005). Remembering a location makes the eyes curve away. *Psychological Science*, *16*(3), 196–199. https://doi.org/10.1111/j.0956-7976.2005.00803.x

Travis, S. L., Mattingley, J. B., & Dux, P. E. (2013). On the role of working memory in spatial contextual cueing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*(1), 208–219. https://doi.org/10.1037/a0028644

van den Berg, R., Shin, H., Chou, W. C., George, R., & Ma, W. J. (2012). Variability in encoding precision accounts for visual short-term memory limitations. *Proceedings of the National Academy of Sciences, USA*, *109*(22), 8780–8785.

Van der Stigchel, S., & Hollingworth, A. (2018). Visuospatial working memory as a fundamental component of the eye movement system. *Current Directions in Psychological Science*, *27*(2), 136–143.

van Ede, F. (2020). Visual working memory and action: Functional links and bi-directional influences. *Visual Cognition*, *28*(5–8), 401–413. https://doi.org/10.1080/13506285.2020.1759744