The rational analysis of memory

Samuel J. Gershman
Department of Psychology and Center for Brain Science
Harvard University

January 25, 2021

In Press, *Oxford Handbook of Human Memory*

Abstract

This chapter surveys rational models of memory, which posit that memory is optimized to store information that will be needed in the future, subject to the constraint that information can only be stored with a limited amount of precision. This optimization problem can be formalized using the framework of rate-distortion theory, which addresses the trade-off between memory precision and task performance. The design principles that emerge from this framework shed light on numerous regularities of memory, as well as how cognitive and environmental factors shape these regularities.

Introduction

One of the fundamental questions about memory is why we remember some things and forget other things. What is special about the things that we remember? Answers to this question have typically come in the form of process models that describe encoding and retrieval operations. Rational process models push the question back one notch further: why are our brains equipped with certain encoding and retrieval processes and not others? This *why* question can only be answered by appealing to *design principles*. Memory evolved because it fulfills an adaptive function. What is that function?

John Anderson was the first to systematically address this question (J. R. Anderson & Milson, 1989; J. R. Anderson & Schooler, 1991). Starting from the assumption that we remember information because we will need it in the future, Anderson derived a wide range of memory properties from ecological estimates of the need probability. The same logic has been extended to disparate areas of memory research, including semantic memory, working memory, serial order memory, reconstructive memory, classical conditioning, and motor adaptation. The goal of this chapter is to synthesize many of these phenomena within a unifying framework, focusing on recognition and reconstructive memory tasks as case studies.

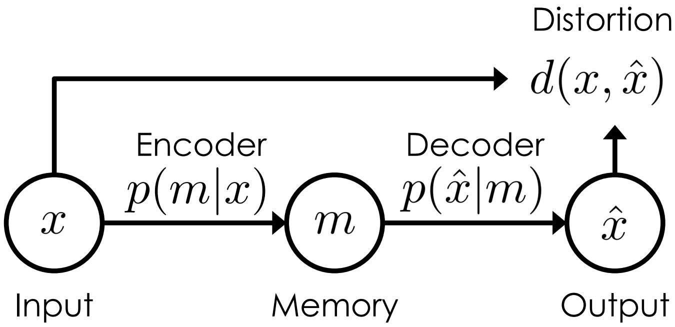

Figure 1: Memory formalized as a communication channel. The input signal x is sampled from an information source $p(x)$ and stochastically encoded into memory trace m with probability $p(m|x)$. The trace can later be decoded into output $\hat{x}$ according to $p(\hat{x}|m)$. This output is evaluated according to the distortion function $d(x, \hat{x})$.

Anderson's rational analysis was primarily concerned with *encoding*: how should information be prioritized for storage in memory such that useful information is more easily retrieved later? This chapter will also address the rational analysis of *decoding* (retrieval): how can we accurately reconstruct the information stored in memory, possibly corrupted by noise at encoding and/or retrieval? By placing the encoding and decoding problems within a single framework, we will arrive at a more complete picture of the design principles underlying human memory.

Memory as a communication channel

A general framework for the rational design of memory systems is schematized in Figure 1. According to this framework, memory is viewed as a communication channel that transmits information about the past into the future by first encoding the information into a memory trace, and then later decoding the information from that trace. Expressed in terms of traditional memory theory, decoding refers to the process of retrieving information from memory and possibly correcting errors based on background knowledge (i.e., the retrieval process is reconstructive rather than a veridical readout).

The communication channel is capacity limited, which means that not all information about the past can be transmitted accurately—encoding is *lossy.* Loss of information can occur at the encoding stage (noisy storage) or at the decoding stage (noisy retrieval), or both. The key point for now is that channel design must consider which information to prioritize, by selectively reducing storage and/or retrieval noise. If some piece of information is needed in the future but cannot be accurately reconstructed, a cost is incurred. The goal is to design a memory system that minimizes the expected cost. This is the problem addressed by the branch of information theory known as *rate-distortion theory* (see Sims, 2016, for a psychology-oriented overview).

More formally, an information source (the environment or experimental task) generates an input signal x with probability $p(x)$. For example, x might be a word, and $p(x)$ is the probability of encountering the word in the environment. To ground the framework more concretely, let us consider two examples. In an item recognition memory task, subjects judge whether items are old (O) or new (N). Thus, in our framework, $x \in \{O, N\}$. New items cannot have a trace stored in memory, so m in this case represents a signal sampled from some "noise" distribution. The information source $p(x)$ describes the proportion of old items in the test list. In a reconstructive memory task, subjects report a continuous variable x such as spatial location or line length, and the information source $p(x)$ is the distribution over locations or lengths.

The signal is encoded into memory trace m by sampling from the encoding distribution $p(m|x)$, where m may be a single number (as in signal detection models discussed below) or a vector representing a collection of perceptual features (e.g., color, shape, etc.). These abstract representations could arise from the collective activity of many neurons, but a detailed neurobiological account is beyond the scope of this chapter. I will look at some specific examples of abstract memory traces in subsequent sections (see also Cox and Shiffrin, Chapter 1.4, this volume).

In the case of episodic memory, the trace may encode information about the spatiotemporal context of the signal (e.g., the list position of an item in a serial memory task), such that each signal is effectively assigned a unique trace (this is the assumption of multiple trace theories; see Hintzman, 1988; Nadel, Samsonovich, Ryan, & Moscovitch, 2000). However, if episodic memory storage has limited capacity, then multiple signals may share elements of the same trace. The degree to which storage is capacity-limited has been the subject of long debate (Brady, Konkle, Alvarez, & Oliva, 2008; Hintzman, Grandy, & Gold, 1981; Shiffrin, Ratcliff, & Clark, 1990; Whitlow & Estes, 1979), and is relevant to the fundamental question of where memory distortions occur in the communication channel (either at storage, retrieval, or both).

At a later time point, the memory trace is retrieved, producing output $\hat{x}$ (an estimate of the signal, for example an old/new judgment in a recognition memory task), and evaluated according to a distortion function, $d(x, \hat{x})$, which quantifies the cost of reconstruction errors. In general, the distortion will be higher to the extent that x and $\hat{x}$ are different (e.g., incorrect old/new judgments in a recognition memory task). Three common distortion functions are given below:

1. Squared error distortion: $d(x, \hat{x}) = (x - \hat{x})^2$.
2. Absolute error distortion: $d(x, \hat{x}) = |x - \hat{x}|$.
3. Hamming (binary) distortion: $d(x, \hat{x}) = \mathbb{I}[x \neq \hat{x}]$, where $\mathbb{I}[\cdot] = 1$ when its argument is true, and 0 otherwise.

Typically, the squared and absolute error distortion functions are applied to continuous-valued signals (e.g., length or location in reconstructive memory tasks), whereas the Hamming distortion is applied to discrete signals (e.g., old/new judgments in a recognition memory task). The choice of distortion function is likely task-dependent, but may be constrained by some general principles. For example, lower reconstruction errors require higher precision, which may be metabolically costly.

In general, the decoder $p(\hat{x}|m)$ can be stochastic, which could arise in cases where the optimal decoder is approximated using a sampling algorithm due to computational constraints. For example, in some cases (discussed further below), the optimal decoder is the mean of the posterior distribution $p(x|m)$:

$$\hat{x} = \mathbb{E}_{p(x|m)}[x] = \sum_x p(x|m)x, \tag{1}$$

where the expectation operator $\mathbb{E}_p[\cdot]$ averages the quantity in the brackets under distribution p. For continuous variables, the summations are replaced by integrals. Intuitively, the optimal reconstruction is computed by adding up each possible signal weighted by its posterior probability, which is given by Bayes' rule:

$$p(x|m) = \frac{p(m|x)p(x)}{\sum_x p(m|x)p(x)}. \tag{2}$$

Bayes' rule says that a hypothetical signal will be a plausible reconstruction if (i) it is likely to have been encoded in trace m (the first term in the numerator), and (ii) it has high probability of being encountered in the environment (the second term in the numerator). If the space of signals is very large or continuous, the posterior mean may not be computationally tractable. Nonetheless, we may still be able to approximate the expectations with a set of samples, $\{x_1, \ldots, x_N\}$, drawn from the posterior:

$$\mathbb{E}_{p(x|m)}[x] \approx \frac{1}{N} \sum_{n=1}^{N} x_n \tag{3}$$

As N grows larger, the approximation becomes increasingly exact. This still leaves the problem of how to generate samples from the posterior, which may itself be computationally intractable. There are a number of ways to do this that are computationally tractable (Dasgupta, Schulz, & Gershman, 2017; Sanborn & Chater, 2016). A neural implementation of stochastic decoding via sampling has been studied by Savin, Dayan, and Lengyel (2014).

Optimal channel design

We can now formally define the channel design problem as finding the conditional distribution $p^*(\hat{x}|x)$, specified by an encoder-decoder pair, that minimizes the expected distortion:

$$p^*(\hat{x}|x) = \underset{p(\hat{x}|x)}{\operatorname{argmin}} \, \mathbb{E}_{p(x,\hat{x})}[d(x,\hat{x})]. \tag{4}$$

When x and m are discrete, this expectation corresponds to:

$$\mathbb{E}_{p(x,\hat{x})}[d(x,\hat{x})] = \sum_x p(x) \sum_m p(m|x) \sum_{\hat{x}} p(\hat{x}|m) d(x,\hat{x}). \tag{5}$$

In other words, the expected distortion is computed by averaging the distortion over randomness in the source $p(x)$, the encoding distribution $p(m|x)$, and the decoding distribution $p(\hat{x}|m)$.

If there are no constraints on $p(\hat{x}|x)$, then the optimal memory channel will encode and decode all input signals with perfect fidelity. We are concerned with the situation in which the channel

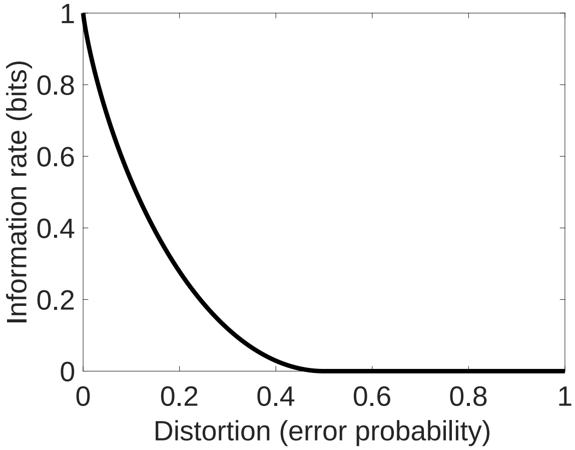

Figure 2: Rate distortion curve for a binary source $p(x = 1) = \frac{1}{2}$ and Hamming distortion.

has a fixed information rate, which is the number of bits available to encode an input signal into memory. More bits per signal means the signal can be encoded with higher fidelity. The channel design problem is then to find the conditional distribution that minimizes expected distortion for a fixed information rate. Intuitively, the optimal channel will encode high probability signals with higher fidelity; if a signal is rarely encountered, there is no point encoding it accurately. We return to this point later when we discuss recognition memory.

The trade-off between rate and expected distortion can be visualized by plotting the rate-distortion curve. An example is shown in Figure 2 for a binary source (random coin flip), $x \sim \text{Bernoulli}(\frac{1}{2})$, and the Hamming distortion function. The rate-distortion curve separates a suboptimal region (above the curve) from an infeasible region (below the curve) that cannot be achieved by any channel. Optimal channels always lie on the rate-distortion curve.

The rate-distortion analysis is extremely general—applying to any communication channel, be it artificial or biological—and encompasses a wide range of rational models as special cases. To simplify the exposition of these ideas, I will first consider the design of the optimal decoder, holding the encoder constant. I will then consider the design of the optimal encoder, holding the decoder constant.

Optimal decoding

The optimal decoder (holding the encoder constant) will in general depend on the distortion function. For a number of common cases, we can obtain the analytical solution (Berger, 1985).

1. For the squared error distortion, the optimal decoder is the posterior mean: $\hat{x} = \mathbb{E}_{p(x|m)}[x]$.
2. For the absolute error distortion, the optimal decoder is the posterior median.

3. For the Hamming (binary) distortion, the optimal decoder is the posterior mode (the signal with the highest posterior probability): $\hat{x} = \mathrm{argmax}_x \, p(x|m)$.

All of these special cases are deterministic functions of the posterior distribution over the input signal given the memory trace, computed according to Bayes' rule (Eq. 2). I now examine these models, and the evidence supporting them, more closely, using case studies from recognition and reconstructive memory tasks.

Recognition memory

Recall that in a recognition memory task, subjects judge whether items are old ($x = O$) or new ($x = N$). In building a model of this task, some theorists have found it useful to abstract away from the item encoder, which may result in a high-dimensional memory trace (see the section below on encoding), and instead focus on a one-dimensional "strength" variable s that is derived from the memory trace. The idea is that we can draw inferences about memory retrieval even in the absence of detailed knowledge about the memory traces themselves. As we will see, this tactic runs into problems. Nonetheless, it simplifies our examination of optimal decoding, and allows us to draw connections to signal detection theories of recognition memory.

Typically, the strength variable is modeled as a scalar drawn from a Gaussian distribution, $s|x \sim \mathcal{N}(\mu_x, \sigma^2)$, as illustrated in Figure 3. Strength variability can arise from multiple sources, including variability of pre-experimental experience as well as encoding or retrieval noise.[1] In signal detection theory, this is known as the equal-variance normal model, because old and new items differ in their mean trace strength but share the same variance (Green & Swets, 1974). One can generalize this distribution to the unequal-variance case (Egan, 1958; Mickes, Wixted, & Wais, 2007), where the old and new items differ in both their means and variances. Other strength distributions have also been studied (Glanzer, Adams, Iverson, & Kim, 1993); I will discuss more detailed encoding models (which eliminate the strength abstraction) below. Generally speaking, the choice of strength distribution is often a matter of convenience, permitting analytical tractability and transparency.

Under the Hamming distortion (the decoder is either correct or incorrect, as in the example of old/new recognition memory), we can express the optimal decoder as a decision rule that returns $\hat{x} = O$ whenever the likelihood odds ratio exceeds the prior odds ratio:

$$\frac{p(s|x = O)}{p(s|x = N)} > \frac{p(x = N)}{p(x = O)}. \tag{6}$$

The likelihood odds ratio expresses the evidence in favor of the hypothesis that an item is old rather than new—how likely is it that signal $x = O$ was encoded in memory m, relative to signal $x = N$? The prior odds ratio expresses the probability of new vs. old based only on knowledge of the information source $p(x)$. Intuitively, you need stronger evidence to report "old" if the frequency of new items exceeds the frequency of old items.

[1] Also note that the value of s should not be conceived literally as a "strength" variable, since it can take negative values. All values of s can be shifted by an arbitrary constant without affecting the results.

One convenience of the Gaussian encoder is that it facilitates analytical calculations of useful decision-theoretic quantities. The *hit rate* (probability of responding "old" when presented with an old item) is given by:

$$\mathrm{H} = p(\hat{x} = O|x = O) = \Phi\left(\frac{d'}{2} - \beta\right), \tag{7}$$

where $\Phi(\cdot)$ is the standard Gaussian cumulative distribution function (with mean 0 and variance 1), $d' = (\mu_O - \mu_N)/\sigma$ is the *sensitivity*, and $\beta = \ln \frac{p(x=N)}{p(x=O)}$ is the *criterion.* Similarly, the *false alarm rate* (probability of responding "old" when presented with a new item) is given by:

$$\mathrm{FA} = p(\hat{x} = O|x = N) = \Phi\left(-\frac{d'}{2} - \beta\right). \tag{8}$$

The sensitivity measures how discriminable old and new items are; sensitivity increases with the separation between the signal and noise distributions.[2] The criterion measures the subject's response bias for reporting old vs. new, depending on their subjective beliefs about the information source. When the criterion is larger, subjects will be more likely to report that an item is new. Some evidence suggests that people learn the optimal criterion based on trial-by-trial feedback (see below).

This signal detection model has been used to explain a number of phenomena in recognition memory (Glanzer, Hilford, & Maloney, 2009). The most extensively studied phenomenon is the *mirror effect*: manipulations that increase the hit rate generally also decrease the false alarm rate. If we assume that strengthening an item (e.g., through repetition or presentation time) increases d' (typically by increasing the signal strength, μ_O), then the mirror effect is immediately apparent from the expressions given above, where the hit and false alarm rates both depend on d' but with opposite signs.

The mirror effect shows that changing sensitivity pushes hit and false alarm rates in opposite directions. Another important implication of the signal detection model is that changing the criterion β should push hit and false alarm rates in the *same* direction. As the criterion increases, then subjects should tend to make both fewer hits and fewer false alarms. Assuming a Gaussian encoder, we can make an even stronger predictions. Let $\Phi^{-1}(\mathrm{H}) = \frac{d'}{2} - \beta$ denote the z-transformed hit rate (the inverse of the Gaussian cumulative distribution function evaluated at the hit rate), and similarly let $\Phi^{-1}(\mathrm{FA}) = -\frac{d'}{2} - \beta$ denote the z-transformed false alarm rate. We can then formulate the following relationship between the two measures:

$$\Phi^{-1}(\mathrm{H}) = d' + \Phi^{-1}(\mathrm{FA}). \tag{9}$$

In other words, the z-transformed hit rate should be a linear function (with slope 1) of the z-transformed false alarm rate. We can visualize this relationship by plotting the two quantities against each other in what is known as a z-transformed receiver operating characteristic (z-ROC) curve. Empirically, the z-ROC curve is linear, but its slope is typically around 0.8. This finding motivates the unequal-variance normal model (Egan, 1958; Mickes et al., 2007), in which the encoder

[2] As the variance goes to 0, sensitivity will get arbitrarily large—memories will not be confused when encoding noise is close to 0, no matter how similar their traces. Thus, similarity between memory traces is not sufficient to produce confusions.

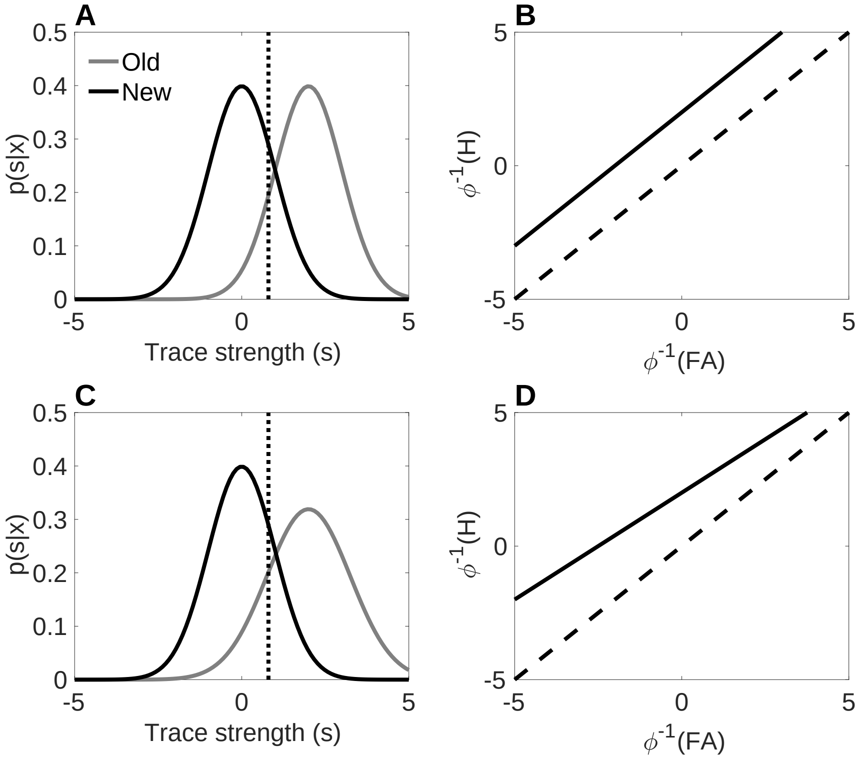

Figure 3: Illustration of signal detection theory applied to recognition memory. (A) Strength distributions for old and new items under the equal-variance normal model. The dotted line indicates the criterion. (B) The relationship between hit rate and false alarm rate, plotted on z-transformed axes (the z-ROC curve). Dashed line indicates the identify function. (C) Strength distributions for old and new items under the unequal-variance normal model, where trace strength has a higher variance for old items. (D) The z-ROC curve for the unequal-variance normal model.

for old items has higher variance (σ_O^2) than the encoder for new items (with variance σ_N^2). In this case, the slope of the z-ROC curve is σ_N/σ_O, as illustrated in Figure 3.

A third phenomenon predicted by the signal detection model is strategic adjustment of the criterion according to the relative base rates of new and old items. Rhodes and Jacoby (2007) reported an experiment in which items were presented on one of two sides of the screen, such that items were more likely to appear on one side of the screen. At test, subjects made more hits and more false alarms to items presented on the "mostly old" side, consistent with the adoption of a lower criterion value, as predicted by the distortion-minimizing criterion placement, $\beta = \ln \frac{p(x=N)}{p(x=O)}$. There is some evidence that the optimal criterion can be learned incrementally through corrective feedback (Han & Dobbins, 2009; Kantner & Lindsay, 2010; Scimeca, Katzman, & Badre, 2016).

Despite these successes, a number of challenges for the signal detection theory of memory have been identified. One challenge is that, even with extensive practice, recognition memory decisions appear to be stochastic (Thomas & Legge, 1970), such that the probability of responding "old" increases with the base rate, $P(x = O)$. This and related observations have motivated models that posit decision noise as an additional source of variance beyond the contribution of perceptual noise (Benjamin, Diaz, & Wee, 2009; Mueller & Weidemann, 2008). Within the rate-distortion framework, decision noise could arise mechanistically from stochastic decoding, as described above. If exact decoding is computationally costly, and this cost is reflected in the distortion function, then the optimal rate-distortion trade-off can be achieved by approximate decoding using a small number of samples from the posterior.[3] The optimal number of samples depends on the relative cost of sample generation and memory errors (see Vul, Goodman, Griffiths, & Tenenbaum, 2014).

A second challenge is that nonlinear z-ROC functions are sometimes reported (Heathcote, 2003; Ratcliff, McKoon, & Tindall, 1994), whereas the Gaussian signal detection analysis described above always requires the z-ROC functions to be linear. Nonlinearities could arise from a threshold-based decision process in which evidence is dichotomized into two discrete states, as has been proposed by some authors (e.g., Yonelinas, 1999). Alternatively, nonlinearities could arise from decision noise (Malmberg & Xu, 2006; Ratcliff et al., 1994). This possibility is appealing from a rate-distortion perspective in light of the relationship between stochastic decoding and decision noise.

There are bigger problems that cannot easily be addressed by small extensions of the abstract model presented here. By focusing on the one-dimensional strength variable, I have abstracted away the structure necessary to account for major laws of recognition memory. In particular, perceptual and temporal proximity between items exerts substantial influence on memory performance by causing retrieval interference. It has long been argued that such interference is the primary determinant of forgetting (McGeoch, 1932). This principle has been formalized in modern computational theories (e.g., Brown, Neath, & Chater, 2007; Howard & Kahana, 2002; Mensink & Raaijmakers, 1988) that capture a wide range of regularities, such as recency (later items in a list are remembered better), contiguity (memory for an item is facilitated by retrieval of another item at a nearby list position), similarity (performance is typically impaired by inter-item similarity), and association (performance is typically impaired by repeated co-occurrence of items). Critically, these models rely upon a multidimensional representation of memory traces over which proximity is defined. Below

[3] Note that experimentally it is more straightforward to manipulate the prior (base rate), but this typically has a monotonic effect on the posterior.

I will review more detailed encoding models that follow in the tradition of multidimensional trace representations, and hence are in principle equipped to account for the regularities cited above.

Reconstructive memory

In reconstructive memory tasks, subjects report an estimate of a studied item's features. Many studies have observed that estimates are systematically biased by distributional information. For example, Huttenlocher, Hedges, and Vevea (2000) asked subjects to reproduce stimuli that varied along a single dimension, such as fish cartoons varying in fatness, squares varying in shade of gray, and lines varying in length. Estimates were systematically biased towards the mean of the stimulus distribution (the central tendency bias), and the bias was strongest for stimuli far from the mean. Similar results have been obtained with temporal reproduction tasks, in which subjects report an estimate of the time interval between two stimulus presentations (Acerbi, Wolpert, & Vijayakumar, 2012; Jazayeri & Shadlen, 2010), and serial recall tasks, in which subjects recall the sequence of items presented during a study phase (Botvinick & Bylsma, 2005). Hemmer and Steyvers (2009) have shown that multiple sources of distributional information can bias estimates; for example, reconstructing the size of an apple depends both on the distribution of apple sizes and the distribution of fruit sizes. Distributional information can also be derived from simultaneously presented information. For example, Brady and Alvarez (2011) found that size estimates for a target stimulus were biased towards the mean size of simultaneously displayed stimuli.

The effect of distributional information can be understood naturally from the perspective of Bayesian decoding. Following previous studies (e.g., Acerbi et al., 2012; Huttenlocher et al., 2000; Jazayeri & Shadlen, 2010), let us assume a Gaussian information source, $x \sim \mathcal{N}(\mu_0, \sigma_0^2)$, and a Gaussian encoder, $m|x \sim \mathcal{N}(x, \sigma^2)$. Under these assumptions, the optimal Bayesian decoder is the same for squared error and absolute error distortion functions:

$$\hat{x} = wm + (1 - w)\mu_0, \tag{10}$$

where

$$w = \frac{\sigma_0^2}{\sigma^2 + \sigma_0^2}. \tag{11}$$

Intuitively, the optimal signal estimate is a linear combination of the memory trace and the prior mean, where the weighting depends on their relative reliability (a consequence of the balance between likelihood and prior in Bayes' rule). When the encoder has low noise variance relative to the prior variance (small σ^2 relative to σ_0^2), the memory trace will be weighted more. The reconstruction bias is the expected difference between the estimate and the true signal:

$$\mathbb{E}_{p(m|x)}[x - \hat{x}] = (x - \mu_0)(1 - w). \tag{12}$$

Thus, bias goes to 0 as encoder noise goes to 0 ($w \to 1$). Importantly, the bias always points in the direction of the prior mean (negative when $x < \mu_0$, positive when $x > \mu_0$), and the bias is largest when the signal is far from the mean (larger $|x - \mu_0|$), consistent with the empirical findings describe above.

Factors that affect encoding noise influence central tendency bias in accordance with the prediction of the model. Cognitive load, which plausibly increases encoding noise by consuming bits that could otherwise be used to store information about the signal, increases the central tendency bias (Allred, Crawford, Duffy, & Smith, 2016; Ashourian & Loewenstein, 2011; Huttenlocher, Hedges, & Duncan, 1991). A similar effect is achieved by increasing the delay between study and test (Crawford, Huttenlocher, & Engebretson, 2000; Olkkonen & Allred, 2014; Olkkonen, McCarthy, & Allred, 2014). I will later consider other properties of the information source and memory channel that affect encoding noise through the optimization of the encoder.

Optimal encoding

I now turn to the question of encoder design: how should memories be stored? Because the memory channel is constrained by the number of bits that it can communicate on average per signal, encoder design is essentially a problem of compression: how can the number of bits be reduced while maintaining sufficient information to achieve low distortion?

For illustration, imagine a reporter interviewing various witnesses and trying to remember two features for each interview: what was said and who said it. If the interviewer only ever interviews the same witness, then the number of bits needed to describe the witness identity is very small, while the number of bits needed to describe each witnesses testimony might be relatively large due to the diversity of content. In general, more diversity means more bits. Shannon's source coding theorem (Shannon, 1948) is a formal statement of this idea: the minimum number of bits needed on average to communicate a signal over a noiseless channel is given by the entropy of the source, $H(x) = -\sum_x p(x) \log p(x)$. A concrete example of a code that achieves optimal compression is *entropy coding*, where the number of bits allocated to signal x is $-\log p(x)$.

If the channel is noisy (e.g., due to storage or retrieval noise), then optimal compression requires allocating extra bits to signals in order to protect against transmission errors. The extra bits typically introduce redundancy into the code, so that if one bit is corrupted, the other bits can correct the error. In the simplest case, the code could just store multiple copies of the memory trace. More sophisticated techniques (e.g., parity check codes) can correct errors with fewer redundant bits, although these have yet to be studied systematically in the cognitive psychology literature.

An important insight into the encoding problem can be obtained by recognizing the connection between compression and statistical inference. Let us start by assuming a generic model of the information source, parametrized in terms of latent variable z:

$$p(x) = \sum_z p(x|z)p(z). \tag{13}$$

This corresponds to a generative process in which a latent variable z is first sampled from the prior $p(z)$, and then a signal x is sampled from the condition distribution $p(x|z)$. Continuing the reporter example, if some witnesses received privileged information ($z = 1$) and some did not ($z = 0$), then their conditional distributions over testimony (the signal, x) will cluster based on the value of z. Because these conditional distributions are lower entropy than the marginal distribution $p(x)$, knowing z allows the signal to be compressed further. In practice, z is typically unknown but can be estimated from x. The cost of transmitting both x and the estimate $\hat{z}$ is $-\log p(x, \hat{z})$, which is minimized by setting $\hat{z}$ to the value that maximizes the posterior. In this way, optimal compression and optimal inference are closely related.

So far, the discussion of optimal encoding has assumed that the goal is to store all information. However, rate-distortion theory allows for prioritization of certain information, depending on the choice of distortion function. If, for example, the reporter's employer cares only about what was said and not who said it, then there is no need to allocate any bits at all to the witness identity. In the interest of brevity, I will not pursue this angle, and focus instead on the role of the information source. The key idea is that if memory reflects probabilistic beliefs about latent variables, then we should be able to derive properties of memory from properties of those beliefs. To explore this idea, I will again use recognition and reconstructive memory tasks as case studies.

Recognition memory

The signal detection model described earlier cannot adequately account for a number of important recognition memory phenomena. One reason is that its representation of the individual items is too impoverished; all items are collapsed into "old" and "new" categories. This has prompted the development of more elaborate item-level models (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997), which also rely on a likelihood ratio computation at retrieval, but make different assumptions about how items are represented and encoded in memory.

McClelland and Chappell (1998) assumed that items are represented as vectors of binary features, stochastically generated from some distribution such that the same item can elicit different feature vectors across multiple presentations. Memory traces are vectors of feature probabilities (equivalent to feature expectations). Thus, the information constraint in this model is the lossy compression of noisy binary vectors (one for each item presentation) into probability vectors (one for each item). The feature probabilities are initialized randomly. When an item is encountered, the feature probabilities are updated to improve the match between the expected and observed item features. We can understand this updating process as a form of approximate encoder optimization. The latent variables in this case correspond to the unknown feature probabilities.

A key consequence of the updating process is *differentiation*: repeated experience with an item causes that item to be more distinguishable from other items, in the sense that the feature expectations become more distinct. McClelland and Chappell (1998) invoke differentiation to explain several phenomena that proved problematic for earlier models (Shiffrin & Steyvers, 1997, offer a similar account, though differing in a number of details). For example, strengthening some items on a list in a recognition memory task does not affect memory for other items on the list, a phenomenon known as the *null list strength effect* (Ratcliff, Clark, & Shiffrin, 1990). According to the differentiation mechanism, the strengthened items should become more dissimilar from the other items on the list (thus decreasing the hit rate for those items), while also becoming more dissimilar from the new test items (thus decreasing the false alarm rate). These factors roughly cancel each other out, leaving d' unaffected.

Anderson's approach to the rational analysis of memory (J. R. Anderson & Milson, 1989; J. R. Anderson & Schooler, 1991) was grounded more explicitly in claims about the information source (i.e., the environment). He posited that memories are consulted in order of their need probability. Because only a subset can be consulted, memories with higher need probability are more likely to be retrieved. We can situate his analysis in the rate-distortion framework by assuming that x corresponds to item labels and z corresponds to the parameters of a categorical distribution over items, $p(x = i) = z_i$, which defines need probability. When an item is encountered, z is adjusted to make that item more probable, similar to the updating process in McClelland and Chappell (1998). Anderson used the relationship between retrieval probability and need probability to derive the form of several canonical memory functions from naturalistic information sources. For example, J. R. Anderson and Schooler (1991) showed that need probability is a power function of item frequency (consistent with the power law of practice), and a power function of the interval between item repetitions (consistent with the power law of forgetting).

Need probability can also be manipulated experimentally. R. B. Anderson, Tweney, Rivardo, and Duncan (1997) had subjects study lists of numbers and then tested them probabilistically (i.e., subjects were asked to recall some proportion of the studied items) after a distractor-filled retention interval. Subjects had a lower rate of forgetting when the test probability was an increasing function of the retention interval, compared to when it was an increasing function of the retention interval, consistent with the predicted effect of need probability.

One apparent challenge for the relationship between need probability and memory is the finding that low-frequency words are often remembered better than high-frequency words (the *word frequency effect*; Glanzer & Bowles, 1976; Gorman, 1961; Lohnas & Kahana, 2013). As mentioned earlier, the rate-distortion analysis implies that high-frequency words should be remembered better, because these contribute more to the expected distortion. Some rational models have sought to accommodate the word frequency effect by assuming that low-frequency words are composed of more distinctive features, and hence their memory traces will be more discriminable from other traces (McClelland & Chappell, 1998; Shiffrin & Steyvers, 1997). In support of this hypothesis, Malmberg, Steyvers, Stephens, and Shiffrin (2002) found that words composed of rare letters were remembered better than words composed of common letters. A fundamental problem for this class of explanation is that some recognition memory studies have found that the word-frequency effect is actually non-monotonic when measured over a sufficiently large range of frequencies: very low-frequency words and very high-frequency words are remembered better than words with intermediate frequencies (Hemmer & Criss, 2013; Rao & Proctor, 1984; Zechmeister, Curt, & Sebastian, 1978). A non-monotonic word-frequency effect has also been reported for free recall tasks (Lohnas & Kahana, 2013). Thus, feature distinctiveness may account for the low-frequency advantage, whereas need probability may account for the high-frequency advantage.

Reconstructive memory

Sims, Jacobs, and Knill (2012) considered the optimal encoder design for a channel with a Gaussian information source and a squared error distortion function.[4] They assumed that the encoder stores a trace of the stimulus with Gaussian-distributed noise, and the noise variance is a free parameter that can be adjusted to minimize distortion subject to the channel's information rate constraint

[4] See also C. Bates and Jacobs (2020), who analyzed rate-distortion trade-offs in reconstructive memory using a more expressive family of deep neural network channels.

(R, measured in bits). These assumptions are essentially the same as those described earlier for reconstructive memory tasks with Gaussian channels. The optimal encoding noise variance is given by:

$$\sigma^2 = \frac{\sigma_0^2}{2^{2R} - 1}, \tag{14}$$

where σ_0^2 is the variance of the information source.[5] The key implication of this result is that encoding noise should increase with stimulus variability (σ_0^2). Sims et al. (2012) found evidence consistent with this prediction in two change detection experiments.[6]

Another prediction of the rate-distortion framework is that encoding precision should increase for stimuli with high need probability. Evidence for this prediction comes from delayed estimation tasks in which subjects are asked to reconstruct a stimulus feature following a delay. Yoo, Klyszejko, Curtis, and Ma (2018) showed that the precision of spatial memories was greater when a precue indicated that an item was more likely to be tested. van den Berg and Ma (2018) found similar results when manipulating need probability for color memory, and have argued that set size effects in such tasks are fully mediated by need probability (see also Fougnie, Cormiea, Kanabar, & Alvarez, 2016).

The rate-distortion framework also sheds light on the ubiquitous scalar variability in magnitude estimation tasks: the standard deviation of stimulus estimates is a linear function of stimulus magnitude. One explanation for scalar variability is that encoding noise increases with stimulus magnitude (Petzschner, Glasauer, & Stephan, 2015; Shi, Church, & Meck, 2013). However, this begs the question *why* the encoder should have this property. The answer comes from an analysis of need probability, which shows that magnitude-dependent encoding precision arises when larger stimulus magnitudes are lower probability, and hence can be represented with lower precision (Piantadosi, 2016; Sun, Wang, Goyal, & Varshney, 2012).

Conclusions and outlook

The idea that memory might obey rational design principles has been explored from many different perspectives. The purpose of this chapter was to unify some of these perspectives within the framework of rate-distortion theory. I have shown how Bayesian approaches to memory, when combined with an appropriate measure of distortion, arise naturally within this framework. On the retrieval side, Bayes' rule prescribes the optimal decoder design for a number of common distortion functions. On the storage side, the optimal encoder uses probabilistic information to efficiently compress signals.

I have focused on cases where the information source is stationary, so that there is a single optimal channel design for a given task. However, it may be more realistic to assume that the information

[5] Sims et al. (2012) distinguished between the variance of the information source and perceptual noise variance, which I have collapsed together here for simplicity.

[6] Change detection is not actually a reconstructive memory task, and thus squared error is not the most natural choice for the distortion function. C. J. Bates, Lerch, Sims, and Jacobs (2019) developed a model that uses the Hamming distortion, a more natural fit with change detection tasks.

source changes over time. Models like the one developed by McClelland and Chappell (1998) can accommodate a changing information source by continually updating beliefs about the parameters governing the source. Some models posit that there are discrete change points where one information source is replaced by another. For example, Gershman, Radulescu, Norman, and Niv (2014) developed a model for reconstructive memory in which changes can be both gradual and abrupt. Consistent with the predictions of this model, reconstructions were biased only by the central tendency of stimuli that putatively belonged to the same information source as the target.

In the interest of brevity, I have focused on recognition and reconstructive memory tasks. However, the ideas presented here are much more general, and have been been applied in various forms to semantic memory (Griffiths, Steyvers, & Firl, 2007), change detection (Brady & Tenenbaum, 2013; van den Berg & Ma, 2018; Wilken & Ma, 2004), classical conditioning (Gershman, Monfils, Norman, & Niv, 2017), and motor adaptation (Kording, Tenenbaum, & Shadmehr, 2007). Even though these tasks presumably tax distinct neural systems, all of them must obey the fundamental constraints of rate-distortion theory, which hold even for systems that have not been optimized. The substantive theoretical claim, which has now received extensive empirical support, is that these systems do in fact resemble the optimal channel design.

An exciting direction for the future is to more directly probe the underlying neural systems for signatures of optimal channel design. Promising work in this direction has recently been pursued by a number of authors. For example, Gershman (2017) discussed how need probability is related to predictive representations in the hippocampus (cf. Stachenfeld, Botvinick, & Gershman, 2017), and could be implemented using spike timing-dependent plasticity (Brea, Gaál, Urbanczik, & Senn, 2016). Mattar and Daw (2018) have shown how this idea can be used to explain many aspects of hippocampal replay activity. Although prior work has mostly concentrated on the hippocampus, it seems likely, given their generality, that these principles apply to other systems as well.

Acknowledgments

I'm indebted to Michael Kahana and David Halpern for detailed comments on an earlier draft of this chapter. This work was supported by a Multi-University Research Initiative Grant (ONR/DoD N00014-17-1-2961).

References

Acerbi, L., Wolpert, D. M., & Vijayakumar, S. (2012). Internal representations of temporal statistics and feedback calibrate motor-sensory interval timing. *PLoS Computational Biology*, *8*, e1002771.

Allred, S. R., Crawford, L. E., Duffy, S., & Smith, J. (2016). Working memory and spatial judgments: Cognitive load increases the central tendency bias. *Psychonomic Bulletin & Review*, *23*, 1825–1831.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*, 703–719.

Anderson, J. R., & Schooler, L. J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

Anderson, R. B., Tweney, R. D., Rivardo, M., & Duncan, S. (1997). Need probability affects retention: A direct demonstration. *Memory & Cognition*, *25*, 867–872.

Ashourian, P., & Loewenstein, Y. (2011). Bayesian inference underlies the contraction bias in delayed comparison tasks. *PloS One*, *6*, e19551.

Bates, C., & Jacobs, R. (2020). Efficient data compression in perception and perceptual memory. *Psychological Review*, *127*, 891–917.

Bates, C. J., Lerch, R. A., Sims, C. R., & Jacobs, R. A. (2019). Adaptive allocation of human visual working memory capacity during statistical and categorical learning. *Journal of Vision*, *19*, 11–11.

Benjamin, A., Diaz, M., & Wee, S. (2009). Signal detection with criterion noise. *Psychological Review*, *116*, 84–115.

Berger, J. O. (1985). *Statistical decision theory and bayesian analysis*. Springer Science & Business Media.

Botvinick, M., & Bylsma, L. M. (2005). Regularization in short-term memory for serial order. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 351–358.

Brady, T. F., & Alvarez, G. A. (2011). Hierarchical encoding in visual working memory: Ensemble statistics bias memory for individual items. *Psychological Science*, *22*, 384–392.

Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*, 14325–14329.

Brady, T. F., & Tenenbaum, J. B. (2013). A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. *Psychological Review*, *120*, 85–109.

Brea, J., Gaál, A. T., Urbanczik, R., & Senn, W. (2016). Prospective coding by spiking neurons. *PLoS Computational Biology*, *12*, e1005003.

Brown, G., Neath, I., & Chater, N. (2007). A temporal ratio model of memory. *Psychological Review*, *114*, 539–576.

Crawford, L. E., Huttenlocher, J., & Engebretson, P. H. (2000). Category effects on estimates of stimuli: Perception or reconstruction? *Psychological Science*, *11*, 280–284.

Dasgupta, I., Schulz, E., & Gershman, S. J. (2017). Where do hypotheses come from? *Cognitive Psychology*, *96*, 1–25.

Egan, J. P. (1958). Recognition memory and the operating characteristic. *USAF Operational Applications Laboratory Technical Note*.

Fougnie, D., Cormiea, S. M., Kanabar, A., & Alvarez, G. A. (2016). Strategic trade-offs between quantity and quality in working memory. *Journal of Experimental Psychology: Human Perception and Performance*, *42*, 1231–1240.

Gershman, S. J. (2017). Predicting the past, remembering the future. *Current Opinion in Behavioral Sciences*, *17*, 7–13.

Gershman, S. J., Monfils, M.-H., Norman, K. A., & Niv, Y. (2017). The computational nature of memory modification. *Elife*, *6*, e23763.

Gershman, S. J., Radulescu, A., Norman, K. A., & Niv, Y. (2014). Statistical computations underlying the dynamics of memory updating. *PLoS Computational Biology*, *10*, e1003939.

Glanzer, M., Adams, J. K., Iverson, G. J., & Kim, K. (1993). The regularities of recognition memory. *Psychological Review*, *100*, 546–567.

Glanzer, M., & Bowles, N. (1976). Analysis of the word-frequency effect in recognition memory. *Journal of Experimental Psychology Human Learning & Memory*, *2*, 21–31.

Glanzer, M., Hilford, A., & Maloney, L. T. (2009). Likelihood ratio decisions in memory: Three implied regularities. *Psychonomic Bulletin & Review*, *16*, 431–455.

Gorman, A. (1961). Recognition memory for nouns as a function of abstractness and frequency. *Journal of Experimental Psychology*, *61*, 23–29.

Green, D., & Swets, J. (1974). *Signal detection theory and psychophysics.* : RF Krieger.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. *Psychological Science*, *18*, 1069–1076.

Han, S., & Dobbins, I. G. (2009). Regulating recognition decisions through incremental reinforcement learning. *Psychonomic Bulletin & Review*, *16*, 469–474.

Heathcote, A. (2003). Item recognition memory and the receiver operating characteristic. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1210–1230.

Hemmer, P., & Criss, A. (2013). The shape of things to come: Evaluating word frequency as a continuous variable in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1947–1952.

Hemmer, P., & Steyvers, M. (2009). A Bayesian account of reconstructive memory. *Topics in Cognitive Science*, *1*, 189–202.

Hintzman, D. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*, 528–551.

Hintzman, D., Grandy, C., & Gold, E. (1981). Memory for frequency: A comparison of two multiple-trace theories. *Journal of Experimental Psychology Human Learning & Memory*, *7*, 231–240.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology*, *46*, 269–299.

Huttenlocher, J., Hedges, L. V., & Duncan, S. (1991). Categories and particulars: Prototype effects in estimating spatial location. *Psychological Review*, *98*, 352–376.

Huttenlocher, J., Hedges, L. V., & Vevea, J. L. (2000). Why do categories affect stimulus judgment? *Journal of Experimental Psychology: General*, *129*(2), 220–241.

Jazayeri, M., & Shadlen, M. N. (2010). Temporal context calibrates interval timing. *Nature Neuroscience*, *13*, 1020.

Kantner, J., & Lindsay, D. S. (2010). Can corrective feedback improve recognition memory? *Memory & Cognition*, *38*, 389–406.

Kording, K. P., Tenenbaum, J. B., & Shadmehr, R. (2007). The dynamics of memory as a consequence of optimal adaptation to a changing body. *Nature Neuroscience*, *10*, 779.

Lohnas, L., & Kahana, M. (2013). Parametric effects of word frequency in memory for mixed frequency lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *39*, 1943–1946.

Malmberg, K. J., Steyvers, M., Stephens, J. D., & Shiffrin, R. M. (2002). Feature frequency effects in recognition memory. *Memory & Cognition*, *30*, 607–613.

Malmberg, K. J., & Xu, J. (2006). The influence of averaging and noisy decision strategies on the recognition memory roc. *Psychonomic Bulletin & Review*, *13*, 99–105.

Mattar, M. G., & Daw, N. D. (2018). Prioritized memory access explains planning and hippocampal replay. *Nature Neuroscience*, *21*, 1609–1617.

McClelland, J., & Chappell, M. (1998). Familiarity breeds differentiation: a subjective-likelihood approach to the effects of experience in recognition memory. *Psychological Review*, *105*(4), 724–760.

McGeoch, J. (1932). Forgetting and the law of disuse. *Psychological Review*, *39*, 352–370.

Mensink, G.-J., & Raaijmakers, J. (1988). A model for interference and forgetting. *Psychological Review*, *95*, 434–455.

Mickes, L., Wixted, J. T., & Wais, P. E. (2007). A direct test of the unequal-variance signal detection model of recognition memory. *Psychonomic Bulletin & Review*, *14*, 858–865.

Mueller, S. T., & Weidemann, C. T. (2008). Decision noise: An explanation for observed violations of signal detection theory. *Psychonomic Bulletin & Review*, *15*, 465–494.

Nadel, L., Samsonovich, A., Ryan, L., & Moscovitch, M. (2000). Multiple trace theory of human memory: computational, neuroimaging, and neuropsychological results. *Hippocampus*, *10*, 352–368.

Olkkonen, M., & Allred, S. R. (2014). Short-term memory affects color perception in context. *PloS One*, *9*, e86488.

Olkkonen, M., McCarthy, P. F., & Allred, S. R. (2014). The central tendency bias in color perception: Effects of internal and external noise. *Journal of Vision*, *14*, 5–5.

Petzschner, F. H., Glasauer, S., & Stephan, K. E. (2015). A Bayesian perspective on magnitude estimation. *Trends in Cognitive Sciences*, *19*, 285–293.

Piantadosi, S. T. (2016). A rational analysis of the approximate number system. *Psychonomic Bulletin & Review*, *23*, 877–886.

Rao, K., & Proctor, R. (1984). Study-phase processing and the word frequency effect in recognition memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 386–394.

Ratcliff, R., Clark, S., & Shiffrin, R. M. (1990). List-strength effect: I. data and discussion. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 163–178.

Ratcliff, R., McKoon, G., & Tindall, M. (1994). Empirical generality of data from recognition memory receiver-operating characteristic functions and implications for the global memory models. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*, 763–785.

Rhodes, M. G., & Jacoby, L. L. (2007). On the dynamic nature of response criterion in recognition memory: Effects of base rate, awareness, and feedback. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 305–320.

Sanborn, A. N., & Chater, N. (2016). Bayesian brains without probabilities. *Trends in Cognitive Sciences*, *20*, 883–893.

Savin, C., Dayan, P., & Lengyel, M. (2014). Optimal recall from bounded metaplastic synapses: predicting functional adaptations in hippocampal area CA3. *PLoS Computational Biology*, *10*, e1003489.

Scimeca, J. M., Katzman, P. L., & Badre, D. (2016). Striatal prediction errors support dynamic control of declarative memory decisions. *Nature Communications*, *7*, 13061.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, *27*, 379–423.

Shi, Z., Church, R. M., & Meck, W. H. (2013). Bayesian optimization of time perception. *Trends in Cognitive Sciences*, *17*, 556–564.

Shiffrin, R. M., Ratcliff, R., & Clark, S. (1990). List-strength effect: Ii. theoretical mechanisms.

Journal of Experimental Psychology: Learning, Memory, and Cognition, *16*, 179–195.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*, 145–166.

Sims, C. R. (2016). Rate-distortion theory and human perception. *Cognition*, *152*, 181–198.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*, 807–830.

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*, 1643–1653.

Sun, J. Z., Wang, G. I., Goyal, V. K., & Varshney, L. R. (2012). A framework for bayesian optimality of psychophysical laws. *Journal of Mathematical Psychology*, *56*, 495–501.

Thomas, E., & Legge, D. (1970). Probability matching as a basis for detection and recognition decisions. *Psychological Review*, *77*, 65–72.

van den Berg, R., & Ma, W. J. (2018). A resource-rational theory of set size effects in human visual working memory. *ELife*, *7*, e34963.

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? optimal decisions from very few samples. *Cognitive Science*, *38*, 599–637.

Whitlow, J., & Estes, W. (1979). Judgments of relative frequency in relation to shifts of event frequencies: Evidence for a limited-capacity model. *Journal of Experimental Psychology Human Learning & Memory*, *5*, 395–408.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of Vision*, *4*, 11–11.

Yonelinas, A. (1999). The contribution of recollection and familiarity to recognition and source-memory judgments: A formal dual-process model and an analysis of receiver operating characteristics. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *25*, 1415–1434.

Yoo, A. H., Klyszejko, Z., Curtis, C. E., & Ma, W. J. (2018). Strategic allocation of working memory resource. *Scientific Reports*, *8*, 16162.

Zechmeister, E. B., Curt, C., & Sebastian, J. A. (1978). Errors in a recognition memory task are a u-shaped function of word frequency. *Bulletin of the Psychonomic Society*, *11*, 371–373.