No Coincidence, George: Processing Limits in Cognitive Function Reflect the Curse of Generalization

Steven M. Frankland[*], *Taylor W. Webb*[†], *Richard L. Lewis*[§∞], *and Jonathan D. Cohen*[∈∞]

[*]Program in Cognitive Science, Dartmouth College
[†] Microsoft Research
[§] Linguistics, and Cognitive Science and Department of Psychology and University of Michigan
[∈] Princeton Neuroscience Institute, Princeton University

[∞] Comparable and complementary contributions

Abstract

The striking constraints of some human cognitive processes stand in stark contrast to the near limitless capability of others. While we can acquire and flexibly use vast amounts of information, the amount we can process at any one time is often stiflingly limited. Here, we integrate ideas from information-theory, cognitive science, and neuroscience to offer a unified account of why processing is often so limited. We argue that this reflects a fundamental tradeoff between representational efficiency and processing efficiency. 'Representational efficiency' refers to how much and how compactly information is represented by an agent, that is directly related to its capacity for generalization. We distinguish this from 'processing efficiency', which refers to how many representations can be processed at the same time. We show that maximizing representational efficiency to optimize the capacity for generalization — a characteristically human cognitive strength — comes at the expense of surprisingly strict limits in processing capacity, an equally characteristic human weakness that has been observed in a variety of cognitive tasks. We refer to this as the "curse of generalization," and formulate this first in information theoretic form, and then demonstrate it in a neurally motivated model of a set of canonical cognitive tasks that have been used to demonstrate the strict limits in human processing capacity. We suggest that the tension between representational efficiency and processing efficiency imposes a fundamental constraint on information processing, that may provide a unified explanation for a wide range of psychological phenomena, from performance in the tasks on which we focus to representational learning and skill acquisition more broadly, as well as the performance of modern machine learning architectures that exhibit generalization capabilities comparable to humans.

Acknowledgements

The authors would like to thank the following individuals for valuable discussion and suggestions that contributed to the material presented in this article: Adel Ardalan, Tim Buschman, Declan Campbell, Nathaniel Daw, Zach Dulberg, Tom Griffiths, Alexander Ku, Sebastian Musslick**,** Marco Nurisso, Randall O'Reilly, Giovanni Petri, Alex Petrov and Jake Russin. The work reported in this article was supported by a Vannevar Bush Award sponsored by ONR to JDC.

Table of Contents

Introduction

Although humans exhibit striking information processing capabilities, they can often also be strikingly limited. Most famously, humans can actively maintain only a handful of items in mind at one time, an observation familiar from both everyday experience and classic laboratory experiments. For example, it has long been recognized that we are limited to remembering only about 7 randomly presented digits at a given time (Miller, 1956) – and even fewer objects in a visual display (Luck & Vogel, 1997; Cowan, 2001). Remarkably, however, the causes of this limit remain poorly understood. While the resources necessary to carry out human cognitive function are clearly finite — e.g., storage space (Amit, 1988), processing time (Cheyette & Piantadosi, 2020), and the precision of neuronal signaling (Wilken & Ma, 2004; Bays, Husain, & Ma, 2017) — it is unclear why limits in these resources alone should impose such severe cognitive constraints, especially given the tremendous volume of information that we can store and retrieve from longer term forms of memory (Brady et al. 2008). Furthermore, it is unclear why these constraints are so similar across diverse forms of representation and types of cognitive processes. This diversity was the focus of George Miller's classic article (1956), in which he pointed out that the limit to the number of items — seven plus or minus two — that can be held in immediate memory (now more commonly referred to as "short term" or "working memory") is remarkably similar to limits in the amount of information — about 2.5 bits — that can be processed in other tasks, such as the ability to match a tone to one of several reference tones (Pollack, 1952; Gravetter & Lockhead, 1973; Nizami, 2010), or to reliably estimate the number of items in a visual display (Kaufman et al. 1948; Mandler & Shebo, 1984; Cheyette & Piantadosi, 2020). The similarity of these constraints intrigued Miller. However, upon careful consideration, he was perplexed by several factors: the disparity of types of information and tasks involved; the fact that seven "items" could in some cases carry substantially more than 2.5 bits of information; and the lack of any principles or processing mechanisms that could explain the relationship between these constraints. Accordingly, he could not convince himself that the prevalence of the "magical number seven" was more than a coincidence. Here, we reconsider Miller's conclusion, by identifying a fundamental principle of information processing that puts representational efficiency and processing efficiency in tension. This tension, coupled with the value of representational efficiency for generalization, can explain the severe processing limits observed across disparate cognitive domains — limits that fall consistently in the range of Miller's "magical number seven."

Overview

Guided by principles of information theory, and building on progress in understanding the cognitive processes and underlying neural mechanisms responsible for representational learning and memory, we assume that human cognition is optimized to maximize *representational efficiency* — that is, the number of possible states of the world that an agent can represent for a given set of representational resources (quantified, in information theoretic terms, as code length) — that undergirds our ability for flexible generalization. However, the representational codes that best achieve this, by representing similar items with correspondingly similar codes, necessarily compromises accuracy, and this in turn imposes severe limits on *processing efficiency* — that is, the number of independent representations that an agent can process *at one time (*for a given code length*).* Here, we argue that this tradeoff between generalization and accuracy can explain the severe constraints in cognitive capacities that are observed ubiquitously when humans are asked to process novel stimuli, including the tasks that Miller considered. Below, we provide a brief overview of these constructs, and their relationship to one another, followed by an illustrative example. We then provide a formal information theoretic treatment that identifies the principles involved, followed by a set of mechanistic models that exemplify these principles in the tasks on which Miller focused.

Representational efficiency. Broadly, we assume that agents seek to optimize representational capacity — that is, the number of states of the world that they can represent. In *principle*, this can be maximized by exhaustively representing every possible state of the world with a distinct code. However, in practice, this is of course impossible: Agents have limited time for learning, limited resources for representation and computation, and even evolution cannot anticipate all possible eventualities. That is, agents face the famous *curse of dimensionality* (Bellman, 1957), that underlies the search for efficient learning algorithms in statistics and modern machine learning. Accordingly, we assume that, in the service of maximizing capacity, agents seek to maximize *representational efficiency* — that is, the number of states that can be represented with a fixed set of representational resources, closely related to classic work on efficient coding in psychology and neuroscience (e.g., Barlow, 1961; Linsker, 1988; Sims, 2016, 2018; Stocker & Simoncelli, 2006; Tishby et al. 2000; Friston, 2011).

Given that the world has structure, representational efficiency can be increased, and thus the curse of dimensionality mitigated, by acquiring representational codes that align with that structure. Cognitive scientists have long highlighted two ways in which natural systems exploit structure in the world to maximize representational efficiency: (i) The use of *semantic* codes that preserve the similarity structure in the world; and (ii) *compositionality*, the representation of

novel states (i.e., ones not previously experienced) through the re-combination of existing codes (Fodor & Pylyshyn, 1988; Smolensky, 1990). Together, these support generalization: semantic codes allow states that are similar along perceptually and/or behaviorally-relevant dimensions to elicit similar interpretations and/or responses (e.g., Shepherd, 1987), while compositional codes allow the system to represent never-before seen states through the composition of familiar elements (Fodor & Pylyshyn, 1988). Together, semantically and compositionally structured representations, acquired from states of the world that have been experienced, can be extended to represent the broader class of states that have not yet been experienced, and in this respect can be thought of as representationally efficient. This is a critical factor underlying the characteristic flexibility of human cognitive function and, most notably, the ability to generalize.

Processing efficiency. While the acquisition of structured codes helps maximize representational efficiency, we argue here that this comes at the cost of *processing efficiency* — that is, how many independent states can be represented and kept distinct *at the same time.*[1] This is because semantically and compositionally structured representations necessarily rely on similarity to support generalization, and the use of such codes to represent states that are otherwise independent of one another introduces correlations among them. That, in turn, makes them less distinguishable, and thus subject to potential confusion and interference, degrading performance when individual items must be kept distinct (e.g., identified), irrespective of their similarity. This tension, that we formalize further on in terms of *decoding error versus generalization error*, is fundamental to theories of memory, and most notably the theory of Complementary Learning Systems (McClelland et al., 1995). In this article, however, we focus on the consequences that it has for another closely related tension: between the value of generalization, that is afforded by *representational* efficiency, and the effectiveness with which representations of independent items can be processed at the same time — that is, *processing* efficiency.

The tension between representational and processing efficiency can be mitigated in one of two ways: by adjusting the representations themselves, or how they are processed. The former can occur through forms of representational learning that transpire over longer time frames (e.g., consolidation, automatization or chunking — Anderson, 1983; Miller, 1956; McClelland,

[1] By "same time," we mean the period during which a given set of representations are in use. In information theoretic terms, this is often referred to simply as a "use" and, in the psychological literature, variously as an "episode," "trial," or "context." From an implementational perspective, this corresponds to the period during which a given set of representations are simultaneously processed (e.g., over which a probability distribution is defined in a mathematical model, or is co-activated in a neural network model). Optimizing this quantity corresponds to the one addressed by multiplexing in communication systems, which seeks to maximize the rate of communication (i.e., number of independent messages that can be communicated at the same time). To our knowledge, however, there is no general formal treatment, in information theoretic terms, of the tradeoff between coding (representational) efficiency and multiplexing (processing) efficiency, or at least no such treatment that addresses this issue within the context of cognitive science and/or neuroscience.

McNaughton & O'Reilly 1995; Musslick et al., 2023; Shiffrin & Schneider, 1977). However, these forms of learning cannot account for the rapid, flexible, and effective processing of novel states of which people are often capable, and that require the use of structured representations that generalize. We assume that other mechanisms must be engaged to overcome the potential for interference that the use of such structured, generalizable representations introduces, and argue that these impose strict constraints on how many representations can be processed at once — that is, they extract an inexorable cost in processing efficiency. We refer to this as "*the curse of generalization",* that complements the familiar curse of dimensionality. Mitigating one invokes the other. While information theory has been used extensively in cognitive science and neuroscience to study the efficiency of coding, it has not generally taken account of representational structure (i.e., compositionality or semanticity), nor has it been used to address processing efficiency (though see Musslick et al., 2023; Petri et al., 2024).

Here we consider how the tension between representational efficiency and processing efficiency — that is, between the curse of dimensionality and the curse of generalization — can be cast in information theoretic terms, and in so doing help explain the pattern of constraints in human cognitive function that so perplexed George Miller. Accordingly, we focus primarily on the tasks that Miller considered, involving perceptual processing and immediate memory. However, in the Discussion, we suggest that this tension may provide a unifying explanation that extends to similarly restrictive constraints observed in other domains of cognitive function, such as the organization of long-term memory, as well as multitasking capability and cognitive control. That is, we suggest the tension between representational efficiency and processing efficiency reflects a fundamental tradeoff that can explain the capacity limits of human cognitive function and, more generally, shapes the envelope of information processing performance within which any system (whether biological or artificial) must operate — a relationship that we suggest might be dubbed "*Miller's Law."*

Illustrative example

To make the problem concrete, consider the visual short term (or working) memory task diagrammed in Figure 1A, that is used paradigmatically to study capacity limits in human information processing. In this task, colored squares appear at various positions in a display, followed by a delay, and the participant's task is to determine whether the color of an item at a particular location has changed. Performance on this task degrades precipitously when the display contains more than 4 or 5 items (see Figure 7C) — that is, within the constraints of the magical number 7 (in this case, minus two).

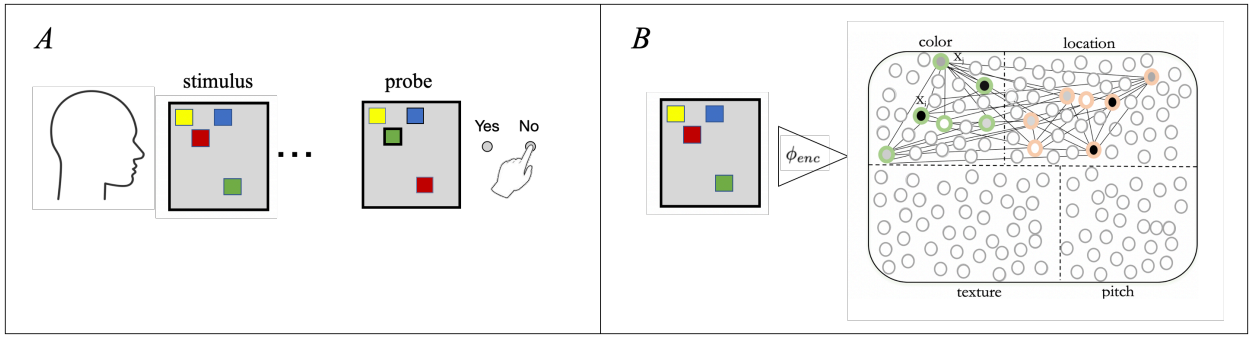

Figure 1. **Illustrative example: visual working memory. (A)** Participant observes a visual display with colored squares and, following a delay, is shown the same display or one in which one of the feature values has been changed at a probed location, and asked to respond whether that item is the same or different from the original display. **(B)** Representation of the information in the display in the form of a compositional code. Each item in the display is represented through the rapid formation of associations between feature values, and the display is represented through the full set of associations.

Figure 1B shows one possible scheme for how such a task might be performed. Each node corresponds to a coding element (comparable to a bit in a binary code), with a pattern of these used to represent a feature value along one of the task-relevant dimensions.[2] Critically, there are *no* nodes (i.e., codes) for *combinations* of feature values across dimensions (i.e, no code is uniquely assigned to the specific combination of *color* and *location* associated with a particular item). That is, the representations are *compositional:* there are codes for every feature value along each dimension, each object is represented as a combination of these by associating ("binding") the codes corresponding to the particular feature values of that object along each dimension (e.g., its color and its location), and the display is represented as the set of associations representing each of the objects in the display. As suggested above, this provides considerable flexibility: by representing each object as a combination of codes along different dimension, *any possible* combination of feature values along those dimensions can be represented; and, similarly, representing scenes through the combination of objects allows any possible combination of objects to be represented. This scheme can also be *acquired* much more efficiently, since it involves acquiring only the codes for feature values along each dimension (which scales multiplicatively), rather than a code for every possible combination (which scales exponentially), thus mitigating the curse of dimensionality noted above.

Importantly, however, compositional coding introduces the risk of interference and confusion (for example, the ability to identify an object by its color if its color is shared by another object). This is exacerbated by the use of semantically structured representations, in which similar feature values are represented similarly along a given dimension (for example, the ability to

[2] In this example, the nodes for feature values along a given dimension are independent of one another; however, as suggested above, nodes representing values close to one another may be more similar than ones representing values more distant from one another (that is, they may capture semantic structure along a given dimension, the consequences of which are discussed below).

reliably and specifically identify the color of an object as red, if this is coded similarly to orange). This is because similarity of representation introduces, by construction, correlations among representations, which makes them less distinguishable (i.e., more likely to be confused), so that even items with different feature values (e.g., red *vs.* orange) may now be confused.[3] That is, the efficiency and flexibility of using structured representations for generalization invokes the curse of generalization, limiting the reliability with which items can be distinguished from one another and, consequently, how many can be represented at the same time. Here we consider how the tension between representational efficiency and processing efficiency — that is, between the curse of dimensionality and the curse of generalization — can be cast in information theoretic terms, and used to explain the pattern of constraints in human cognitive function that so perplexed George Miller.

Argument Summary

We make the argument in three steps. First, we formalize the tradeoff between accuracy and generalization. This aligns closely with the bias-variance tradeoff widely discussed in machine learning, and the tension between pattern separation and pattern completion in theories of learning in cognitive science and neuroscience. Here, we draw upon information theory to formulate this in terms of a tradeoff between *decoding error* and *generalization error*. We provide an analysis of this tradeoff, showing that increases in representational efficiency, that minimize generalization error, come at the expense of decoding error (i.e., diminished accuracy in identifying a specific item). Second, we show that this effect on decoding error is amplified as the number of items that must be actively represented grows, radically restricting processing efficiency (i.e., the number of items that can be accurately decoded at the same time). Finally, we point out that these are exactly the conditions elicited by tasks in which strict limits to processing capacity are observed (such as the tasks considered by Miller): On the one hand, they invariably use novel stimuli (i.e., involving arbitrary combinations of features), and thus demand the use of codes that can support generalization. On the other hand, they evaluate performance in terms of accuracy of identification — a capacity that, as demonstrated.[4]

---

[3] This is sometimes discussed as the tension between "pattern separation" and "pattern completion" faced by any form of content-addressable memory (e.g., McClelland et al., 2005), and can be formalized in terms of Bayesian inference as a process of latent cause inference (e.g., Gershman et al., 2017), in which it must be decided whether an item that does not fully match the existing contents of memory should be treated as a partial cue to retrieve the closest entry (i.e., used for generalization), and/or treated as a new item to be encoded (or used to modify existing memories; that is, used for identification).

[4] The interest in these tasks is precisely because they probe mechanisms, such as working memory, thought to be important for the processing of novel stimuli. Although such stimuli are *individually* infrequent (at least in the agent's experience), they are common as a *class* in realistically rich environments. Generalization makes it possible to process such stimuli, though at the cost of being able to process only a small number at a time. Here, we focus on this tradeoff in the context of the tasks that Miller considered, involving immediate memory and perceptual processing. However, in the Discussion, we suggest that this tradeoff extends to a much wider range of tasks, including ones involving long-term memory, as well as multitasking and cognitive control.

Information Theoretic Formulation

To cast the problem in information-theoretic terms, we assume that the knowledge required to perform a task is represented by codes for the relevant information. That is, observed states of the world are converted to internal representations (encoded), that are accessed for processing (in channels) to generate (transmit) a task-relevant inference and/or action (decoded) in response. Here, we are interested in how the structure of the internal representational codes impact how much information from the original source can be transmitted to generate the task-relevant response(s) at a given time. We use code length (in bits) to quantify the representational resources that the agent has available, and we focus on two critical factors: how many states of the world can be represented with those resources (that we refer to as *representational efficiency*), and how many distinct items within such states can be processed at a given time (that we refer to as *processing efficiency*)? To formalize these questions, we first outline our assumptions about the structure of the world, and define different forms of coding structure that can be used to represent states of the world. We then formalize what we mean by representational and processing efficiency, in turn, and consider how coding structure impacts these. More specifically, we show how semantic and compositional codes that support generalization dramatically increase code overlap (i.e., decrease minimum code distance), which leads to a dramatic increase in decoding error and a concomitant decrease in processing efficiency.

Structure of the World

We assume that there is dimensional structure in the world that our perceptual and representational apparatus have evolved to exploit, and that states of the world can be described in terms of this structure. That is, we assume that this structure in the world can be captured by describing the world as points in a multidimensional vector space, with each point described by a set of vectors, each of which represents a single feature value along one of the dimensions (including location and time),[5] and with dimensions that are orthogonal to one another so that, over the entire space, the feature values along any given dimension vary independently of those along any other. The dimensions comprise a basis set for the space, and any point in the space can be compositionally coded as a set of the feature values along each dimension. Furthermore, we define two types of hierarchically-related subspaces that are

[5] In principle, features might be represented at any level of precision (or, in information theoretic terms, code length) along any given dimension. Although precision will impact representational and/or processing efficiency individually (we consider this in the Discussion) importantly, the tension between them obtains for *any given level* of precision. In the abstract formulations we consider here, we fix precision (i.e., code length) both within and across dimensions, while in the simulations of human behavioral performance we draw empirical empirical evidence concerning the resolution of features along different dimensions.

relevant to the information processing demands of physical agents: *states*, that are, for every point in physical space (i.e., along the three dimensions of position) and a given point (or set of points) in time, the features values along all other dimensions; and *items* within a state (e.g., objects within a display, or a display within the broader scope of the state), that are comprised of a subset of points in physical space and/or time that share covariance structure across other feature dimensions (e.g., color, shape, etc.) that is independent of any other such subset. For example, an item in Figure 1 (i.e., one of the colored shapes) can be defined as the locations in the display that all comprise the same shape (square) and share a given color (e.g., yellow), as shown in Figure 2B.[6] Critically, we assume that all of the feature values of each item, and all the items in a state are represented using the *same* set of codes.

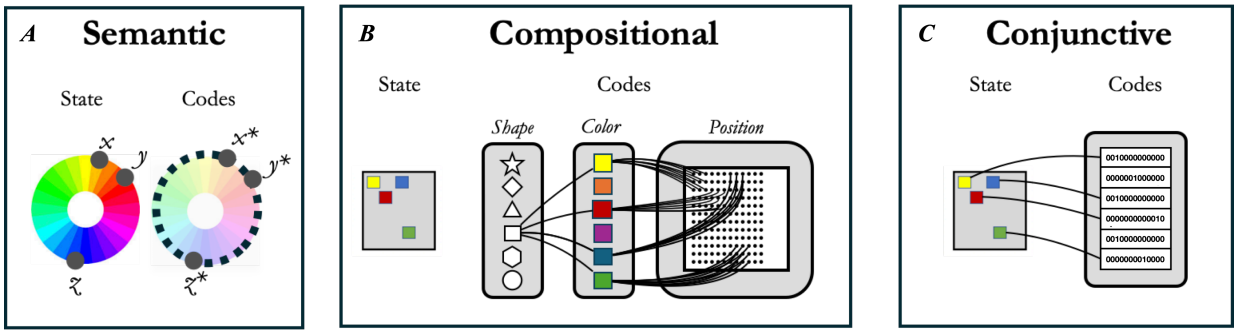

Figure 2. **Types of codes. (A)** *Semantic codes* represent feature values that are similar along a given dimension (e.g., red *x* and orange *y*) with codes that are themselves more similar (i.e., comparably close to one another, such as x* and y*) than they are to others (e.g., blue *z*). This supports *generalization* (Shepard, 1987), by allowing similar feature values (e.g., red and orange) to be processed in similar ways (e.g., if it is known that red berries are poisonous, then it might be advisable to avoid eating orange ones, but perhaps less so for blue ones). However, for the same reason, if there is any noise in processing, it makes distinguishing items less reliable if they have feature values that are similar (e.g., *x* and *y*) than more different (e.g., *x* and *z*). **(B)** *Compositional codes* represent states of the world as compositions of feature values along each dimension (e.g., *shape*: square, *color*: red, *x-position*: 3, *y-position*: 3, etc.). This supports the *flexibility* to process novel stimuli, by allowing any state to be represented whether or not it has previously been experienced (Fodor & Pylyshyn, 1988). However, if there is no additional mechanism that independently binds the feature values of distinct items (shown here as lines connecting feature values across dimensions), then it is impossible to determine which feature values belong to which items. **(C)** Conjunctive codes assign a distinct code to each distinct item that, in the limit, is equidistant from all other codes. While this preserves the unique identify of each item, it is an inefficient form of coding (see text).

Compositional structure. Given the hierarchical structure outlined above, we can consider two levels of *compositionality* within the representation of a given state: compositionality of the representation of each item in terms of its *feature values,* that amounts to the concatenation of its features along each dimension (i.e., for each of the points comprising its location), and

[6] The level of resolution (or precision) at which items are represented is of course an important consideration that, in information theoretic terms, can be considered in terms of code length (for example, each item (colored square) in Figure 2B could be represented, at a lower level of resolution, as occupying a single shared position, such as "upper left"). As explained further below, we hold code length constant in our analyses, in order to isolate the effects of *correlations* in the code, and show that these are qualitatively preserved over reasonable assumptions about code length. In the Discussion we consider how these may interact with code length, and how this may relate to work that links capacity limits to a tradeoff between precision and load (Wilken & Ma, 2004; Ma, Hussain & Bays, 2014)

reflecting the covariation of its feature values along those dimensions; and the compositionality of *items* within a given state (e.g., objects in a display), that amounts to the simultaneous representation of (i.e., superposition of the codes for) feature values along each dimension belonging to the different items in that state, and reflecting the covariation of those items in physical space (i.e., the locations spanned by the display) and/or in time (i.e., their co-occurrence during the duration of the display). Accordingly, to the extent that the identities of individual items need to be kept distinct within a state and/or different states need to be kept distinct, then there needs to be some mechanism of associating (binding) the feature values belonging to each particular item and/or the items within a state. This is because compositional representations reflect only the first order statistics of the codes involved (i.e., their marginal frequencies), and not higher order statistics need to represent correlations. For items, this means only the (frequency with which) feature values are present in the state are represented. This is sufficient to identify the coherent covariation of feature values that defines an individual item, but insufficient to keep these distinct for different co-occurring items; the latter requires some associative mechanism for encoding higher order statistics among feature values. The same applies to higher level conjunctions, such as among items within an display.

Semantic structure. We assume that the world also exhibits *semantic* structure (see Figure 2A). In the abstract model presented in the next section, we operationalize this as the extent to which there is metric structure along each feature dimension, such that feature values can be more or less similar to (i.e., distant from) other feature values along the same dimension. In the subsequent, neurally-inspired simulations, we use empirical estimates of similarity structure among feature values along a dimension to implement semantic structure.

The work we present points out that tasks used to study constraints in human information processing capacity are usually designed to probe item identification in settings that demand compositional representation — that is, the encoding of states that are novel compositions of items, which may themselves be novel compositions of feature values. This demand reflects the *representational efficiency* of compositional and semantic coding, that we consider next; however, it comes at the cost of *processing efficiency*, that we consider further below.

Representational Efficiency

We define this as the representational *capacity* of an agent relative to its representational *resources.* Capacity is defined as the maximum of the mutual information between the agent's representations and those of all potential states of the word it may encounter. Representational resources are quantified as the average code length (in bits) used to achieve a given level of

representational capacity. That is, representational efficiency is the representational capacity (mutual information) for a given level of representational resources (code length).

If the world had no structure, then the *only* way to represent it would be by pairing each state with a unique code. That is, there would be no opportunity for increasing efficiency, and the agent would obligately face the curse of dimensionality, manifest in the number of samples (and hence time) required to learn such a code.[7] Fortunately, however, the world has structure, and thus the curse of dimensionality can be mitigated through more efficient coding.

Probabilistic vs. similarity structure. Traditional work on efficient coding in information theory (e.g., Barlow, 1961; Stocker & Simoncelli, 2006), and its application to learning and representation (Friston , 2011; Linsker, 1988; Tishby et al., 2001), addresses one particular form of structure: the *probability distribution* over different states of the world. Accordingly, a fundamental tenet of information theory is that efficiency of coding can be maximized by optimizing code length to reflect probabilistic structure: the length of a code assigned to a state should be proportional to its surprisal (the negative log probability of its occurrence), so that shorter codes are used to represent more frequent states. This minimizes the average code length for a given level of mutual information between the codes and potential states of the world; that is, it maximizes representational efficiency, by minimizing the representational resources needed for a given representational capacity.

The idea that the frequency of a state should impact how it is represented also plays a central role in theories of representation based on statistical learning. While these do not formulate the relationship specifically in terms of code length, they do address how factors such as passive experience, replay, and practice — mediated by psychological processes such as consolidation (e.g., McClelland et al. 1995, 2016) and automatization (Anderson, 1983; Shiffrin & Schneider, 1977; Musslick et al., 2023) — can lead information that is needed more frequently to be represented more compactly. The same may be true for longer term processes, such as evolution and/or early development. In the Discussion we consider how this relationship between frequency and representational structure might be considered within an information theoretic framework. However, our primary focus in this article is on how *infrequent* states are processed.

Given our assumptions about the structure of the world outlined above, infrequent states are ones comprised of novel or unfamiliar combinations of items, and/or items made up of similarly unpredictable combinations of feature values. However, while individual instances of such

[7] This amounts to the Coupon Collector's problem, which requires, in expectation, $n * log(n)$ samples to solve.

states are infrequent, as a class they can be frequent, especially in high dimensional, non-stationary environments, and the effectiveness with which they can be processed is critical to adaptation. Furthermore, these are precisely the kinds of states used in tasks that showcase limits to human information processing (e.g., displays containing novel combinations of colors, shapes and positions, such as in Figure 1A). At the same time, they also showcase the flexibility of human information processing — the ability to rely on existing codes to represent novel or unfamiliar states.

Theories of optimal coding do not generally address this capacity for generalization: they prescribe how to assign codes to *newly* encountered states, but not how to best represent them in terms of *existing* codes. As noted above, Shepard's (1978) Universal Law of Generalization prescribes the use of structured codes, that reflect the similarity structure of the world, and not just their probabilistic structure. To isolate the effects of such structure, independent of probabilistic structure, the work we report here makes the simplifying, but focusing assumption that the states of interest are equiprobable and represented with codes of equal length. We then ask: What forms of structure in the code maximize representational efficiency — that is, maximize representational capacity for a given fixed representational resource (code length)? We refer to this as representational efficiency to distinguish it from *coding* efficiency, which in standard analyses is tied to frequency of occurrence.

One possible coding scheme is to pair each unique state of the world with its own code. We refer to these as conjunctive codes, because they are arbitrary mappings to states that reflect only the *conjunction* of feature values or items that comprise states, and bear no systematic relation to their constituent parts or to one another. However, under the assumption that the world has compositional and/or semantic structure, it is inefficient to represent the world in this way. As noted above, the samples required to learn to represent every possible state of the world grows exponentially with the number of feature values and dimensions in the world. In contrast, the number of samples needed to acquire a strictly compositional code grows linearly with the number of feature values and dimensions. Thus, for a given world, compositional representations of its states can be acquired and represented more efficiently than conjunctive ones; and, once acquired, they can be used more flexibly to represent any possible — but as yet unexperienced — state. Similarly, semanticity allows responses to novel states to be similar to those that share similar feature values along the relevant dimensions. Together, these maximize generalization — the ability to accurately represent and/or respond appropriately to states that have not previously been encountered, based on existing representations of ones that are *similar* (Shepard, 1987). However, they come at the cost of the ability to reliably identify *a specific* state, which is the criterion for performance in standard information theoretical

analyses (e.g., the mutual information between the transmitted and received code). This is because introducing similarity structure increases overlap among codes, making them harder to distinguish when that is required.[8] We formalize this tension between accuracy and generalization in information theoretic terms below.

Decoding vs. generalization error. The effects of representational structure on accuracy versus generalization can be quantified by evaluating performance with respect to *decoding error* versus *generalization error*. The former corresponds to the standard construct in information theory: an error occurs whenever the code received is not *exactly the one* transmitted. To reduce decoding error, codes should be as far apart from one another as possible. The Singleton Bound (Singleton, 1964)[9] places a lower bound on the best (largest) minimum distance achievable between codes given a fixed code length and desired number of unique codes in the "codebook;" in our setting the codebook size is a measure of representational capacity (at some point in the agent's learning history). Codes that achieve the Singleton Bound are referred to as *maximum distance separable* (MDS) codes, and are approximately equidistant from each other (MacWilliams & Sloan, 1977).

In standard treatments, where similarity structure is not considered, MDS codes are optimal. However, codes that have similarity structure — such as those that preserve the structure in the world — will fall short of the Singleton Bound, since some codes will be closer to each other than others, by construction. This, in turn, should increase decoding error. Critically, however, this can be contrasted with *generalization* error, for which cost is measured as some function of the distance between the transmitted and received codes (unlike standard coding theory, in which code distance has no impact on error cost). Below, we use a simple distance-dependent error function based on Shepard's (1987) analysis of generalization to examine the tradeoff between decoding error and generalization error. However, since coding distance is the mediating factor, first we consider the impact that semantic and compositional structure have on the minimum distance between codes with respect to the Singleton Bound.

Structure and minimum coding distance. Critically, even modest amounts of structure precipitously reduce the minimum code distance, driving it well below the Singleton Bound. This can be seen in Figure 3, which shows the effect of compositional and semantic structure on the minimum distance between codes (using Hamming distance, a standard metric used in

---

[8] In machine learning, this tension is often discussed in terms of the bias-variance tradeoff, or V-C theory (Vapnik & Chervonenkis, 1968; Vapnik 2000), which refers to the tension between use of structure in the data to bias learning and inference in order to make them more efficient, but that can impair accuracy and make learning less efficient if the bias is misaligned with the structure of new data (i.e., there is variance in the data not accounted for by the previously learned structure).

[9] This also known as the Joshibound (Joshi, 1958).

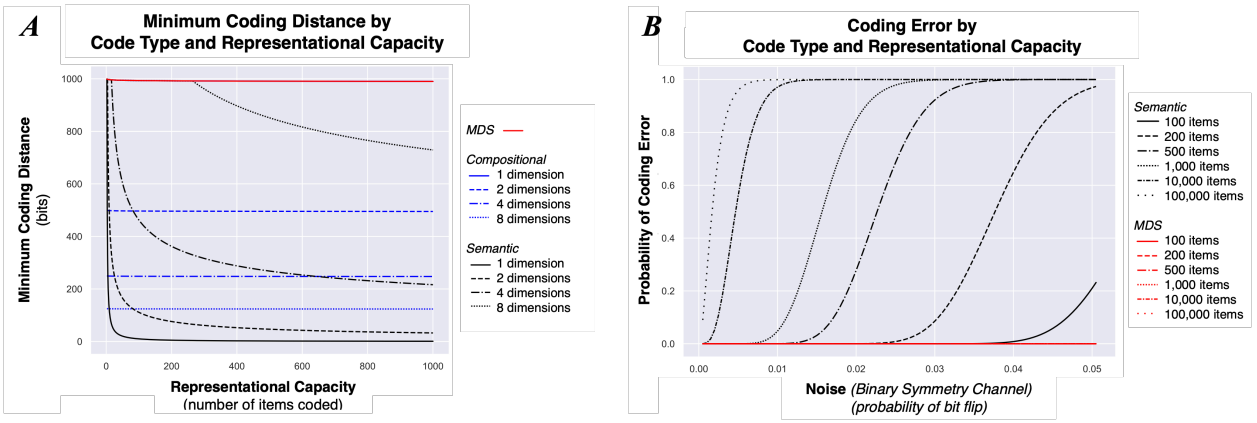

Figure 3. **Coding structure, distance and error. (A)** Effect of representational capacity on minimum coding distance for different forms structure. For simplicity, we assume distributed binary codes. Plots show minimum code distance computed analytically (see equations in text) over all codes for a given coding scheme and different numbers of items using a fixed code length (1000 bits). **MDS codes** (red line): codes given the maximum possible distance from one another (corresponding to conjunctive codes), that implement the Singleton bound. Note that the minimum distance drops numerically but slowly as the number of items increases. **Compositional codes** (blue lines): codes formed by assigning each item a single feature value along each of the specified number of dimensions *(d),* each of which has $\sqrt[d]{N}$ feature values (where *N* is the number of items) that are equidistant from one another along dimensions and orthogonal across dimensions. Note that only the number of dimensions matters: since feature values are orthogonal to one another, the only correlations come from shared feature values across items, and the minimum distance is determined by the maximum number of shared feature values which grows with the number of dimensions (see text). **Compositional and semantic codes** (black lines): same as compositional, but with feature values that are evenly spaced along each dimension (see text), but remain orthogonal across dimensions). **(B)** Probability of decoding error as a function of noise (*p*) for one-dimensional semantic (black) vs. MDS (red) codes at different representational capacities (note: there are no differences for MDS codes, so all lines are superimposed).

information theory) compared to MDS codes (i.e., that implement the Singleton Bound), as a function of number of states represented in the code book (i.e., representational capacity) for a fixed code length (i.e., representational resource) of 1000 bits. Note that the maximum possible distance between any pair of codes is the length of the code (i.e., for codes that are complements of one another), but this can be achieved with only two codes. As representational capacity increases, the minimal distance possible between codes necessarily decreases, reaching the limit of 1 bit at full representational capacity (i.e., all possible codes are used to represent states), in which case *any* single-bit error will produce a decoding error. The Singleton Bound identifies the furthest apart that two codes can be for a given representational capacity:

$$K - \log_2(N) + 1 \qquad (1)$$

where *K* is the size of the code (here in bits) and *N* is the number of codes in the codebook.

The logarithmic nature of this relationship means that, for the low end of representational capacities (i.e., number of codes used to represent states), increasing this should have negligible impact on minimum coding distance. This is evident in Figure 3, which focuses on the

range of 1 to 1000 states (out of the 2^{1000} possible for a code length of 1000): there is a negligible reduction in minimum coding distance for MDS codes over this range (red line). In contrast, the structured codes (blue and black lines) impose a much sharper drop in minimum coding distance, even for this range in which representational capacity is a small fraction of the full space of states that could be represented. Both compositional and semantic structure exhibit qualitatively similar effects.

Compositional structure. To consider the effect of compositionality separately from semantic structure (blue lines in Figure 3), we factored the code into *D* dimensions, and divided the code length (*K* 1000) evenly among dimensions (with $\sqrt[d]{N}$ feature values per dimension, where *N* is the number of items), without assigning any semantic (metric) structure to the values along each dimensions (that is, all codes were equidistant from one another along each dimension). The case of a single dimension (solid blue line) is identical to the MDS code (red line), as it should be. This closely approximates the Singleton Bound, as noted above. However, as *D* increases, the minimum coding distance rapidly decreases. This is because the minimum coding distance is determined by codes that have the same value along *D*–1 dimensions, differing only in the remaining dimension, and the similarity of such codes increases (i.e., their distance decreases) as the number of dimensions they share increases, given by:

$$\frac{K}{F} - \log_2(N^{\frac{1}{D}}) \qquad (2)$$

In other words, compositionality decreases the minimum coding distance by increasing the potential for codes to share the same feature value along multiple dimensions — that is, by increasing the opportunity for similarity structure *over* dimensions. Next, we examined the extent to which this interacted with semanticity (i.e., similarity structure *within* dimensions).

Semantic structure. To evaluate the effect of semantic structure within dimensions, we again constructed dimensions in which the code length (*K*=1000) was distributed evenly among them. In this case, however, we distributed the values of codes along each dimension systematically along each dimension, by assigning two of them as far apart as possible (*K/D*), and then evenly spacing all the others between them. For a representational capacity of *N,* the minimum coding distance over the full space increases with *D as:*

$$\frac{K}{N^{\frac{1}{D}} - 1} \qquad (3)$$

For the limiting case of a single dimension (solid black line), the effect is dramatic: simply placing codes on a line reduces the minimum coding distance to 1 bit, which is not surprising since, by construction, that is the distance between each code and the one next to it. Interestingly, however, unlike the effects of compositionality on in its own, increasing *D* for semantically structured dimensions *increases* the minimum coding distance (black dashed and/or dotted lines). This is because dimensions are assumed to be orthogonal, so that feature values related along a given dimension can be reassigned to different independent dimensions as the number of those grows, and therefore states can be assigned increasingly distinct codes. This is made clearest by considering the limit of *D* = *K,* in which each state is assigned a unique (i.e., conjunctive) code along its own dimension, and thus is maximally distant from all other codes. Importantly, however, this is at the expense of generalization. That is, increasing the dimensionality of semantically structured codes increases their distinctiveness, but compromises generalization.[10] We quantify this tension below. However, it worth noting a qualitative effect here, that may help explain the strongly asymmetric relationship between decoding error and generalization error that we quantify below: The increase in minimum coding distance that comes with increasing dimensional structure for semantic codes is a relatively weak effect. For example, while increasing the number of dimensions to 8 (dotted line) affords codes that approach the Singleton Bound, this is only up to a representational capacity of about 300, beyond which the minimum coding distance drops rapidly. 300 codes is a negligible fraction of the full number of codes the system can represent ($3/2^{998}$). For a given code length and representational capacity, it is possible to compute the number of semantic dimensions at which the minimum distance approaches the Singleton Bound. For example, for even a modest representational capacity of 10,000 items, 14-dimensional spaces are required for 1000-bit codes—but the number of feature values per dimension approaches the degenerate case of two (see Supplemental Information).

Minimal coding distance and decoding error. Under virtually any noise model, the precipitous reductions in minimum coding distance as a function of structure shown in Figure 3 will produce a correspondingly dramatic increase in decoding error relative to that afforded by the Singleton Bound. For example, under the simple *Binary Symmetric Channel* (BSC) noise model — in which there is some small probability *p* that a bit will flip — the probability of decoding error as a function of *p* and minimum code distance *d* is given by:

$$P_e = \sum_{k=t+1}^{K} \binom{K}{k} p^k (1-p)^{K-k} \qquad (4)$$

[10] This accords with the general notion that generalization involves abstraction, which in turn involves dimension reduction.

where $t = \left\lfloor \frac{d-1}{2} \right\rfloor$ is the maximum number of bit flips that can be corrected without confusing two codes, $\binom{K}{k}$ is the number of ways to choose *k* bits out of *K,* and $p^k(1-p)^{K-k}$ is the probability of having exactly *k* bit errors.

Figure 3B shows the probability of decoding error as a function of *p* for MDS codes and two-dimensional semantic codes at several codebook sizes. Note the sharply rising error rates relative to that of MDS codes, which are not discernible on the graph for these parameters.

It is important to note that others since Miller have sought to identify the kinds of error that can explain human capacity limits in information theoretic terms. For example, SIms (2016) used Rate Distortion Theory (RDT) to explore the kinds of information loss that can best account for patterns of performance in the tasks that Miller considered. Here, we focus on the tension *between* two forms of loss — decoding error and generalization error — and how this may explain capacity limits. In the Discussion, we consider how this relates to Sims' findings, and RDT more generally. At the same time, it is also important to note that the tradeoff between decoding and generalization is similar to ones that have been long recognized in other settings and forms. For example, it can be viewed as underlying a similar tension between pattern separation and pattern completion that is central to Complementary Learning System Theory (McClelland et al., 1995), cast there in terms of how partial information is used for the purposes of identification (pattern separation) versus generalization (pattern completion). The same tension has also been addressed in machine learning in terms of the bias-variance tradeoff (see note 7). Here, we cast this explicitly in information theoretic terms, both to address its relationship to processing efficiency, that we turn to next, as well as for generality of application to other domains, such as automatic versus controlled processing and multitasking (e.g., Musslick et al., 2023; Petri et al., in press) or "in context" versus weight-based learning in neural networks and machine learning (Chan et al., 2022) that we consider in the Discussion.

Processing Efficiency

The tension between decoding error and generalization error is most clearly evident under conditions that, on the one hand require generalization for adequate performance but, on the other, evaluate performance in terms of accuracy of identification (i.e., decoding error). The requirement for generalization is greatest when processing novel stimuli, and thus must rely on structured (compositional and/or semantic) codes to represent. As elaborated above, this should invoke the curse of generalization, and an attendant increase in decoding error.

Furthermore, this should be greatly exacerbated when *multiple* novel items must be represented *at the same time* — that is, it has a direct impact on processing efficiency. This is evident in the Illustrative Example presented above, where representing novel stimuli (e.g., a yellow square in the upper right corner of a display, next to a green square diagonally just below and to the right of it, etc.) requires the use of compositional coding, both for items in terms of their features, and the display in terms of those items. However, this poses a problem: without some mechanism for associating each point in the display, and similarly each item, with its corresponding set of features, it would be impossible to determine which particular features belong to a given item — for example, to distinguish it from the case in which the green square was in the upper right corner, and the yellow square was below it. That is, the use of compositional coding restricts the number of items that can be represented and identified at the same time. This has been referred to as the "Binding Problem" in cognitive science (e.g., Treisman & Gelade, 1980), that highlights the curse of compositionality, classically observed as a cost in the ability to accurately identify the features of a given item in a display with many. As noted earlier, this is because strictly compositional codes (i.e., on their own, without binding) carry information only about the fist order statistics of features in a state (which are present or absent), and not higher order statistics (i.e., their associations with one another, or patterns of covariation). Furthermore, this is exacerbated by semantic coding, in which the similarity of codes for similar features makes the ability to distinguish such codes more sensitive to perturbation (e.g., if the squares had subtly different shades of yellow). The constraints that structured codes place on processing efficiency can be mitigated in one of two ways.

Code modification. One way to avert the cost of decoding error, in accord with the central tenet of RDT, is to augment the code under pressure from the environment. This can be done either by adding conjunctive representations for frequently encountered items (to mitigate the problem of compositionality) and/or increasing precision by representing finer distinctions among features along relevant dimensions (to mitigate the problem of semanticity). While both of these can occur, as noted above, through representational learning over longer time frames (e.g., through consolidation, automatization, or "chunking"), such longterm adaptation can't explain the flexibility people exhibit in representing and processing *novel* states that, by definition, have not had a chance to exert pressure on the code. However, it can be explained by mechanisms of short term adaptation, either by *rapidly* binding features along different dimensions of a compositional code (i.e., to represent higher order statistics), and/or dynamically modify the precision with which features are represented along a given dimension (to diminish the impact of semanticity). The former is the function ascribed to episodic memory by Complementary Learning Systems Theory (McClelland et al., 1995). This aligns closely with binding mechanisms we consider in the work presented below, which indicates that rapid

binding can only partially mitigate the constraints on processing efficiency imposed by compositionality — owing in large measure to the additional effects of semanticity. In principle, the latter could be mitigated by modifying precision, however it has been argued that this too is subject to constraints (e.g., Wilken & Ma, 2004; Bays & Husain, 2008). While we do not focus on the latter in this article, how it interacts with compositionality, mechanisms of rapid binding, and processing efficiency remain important questions for future research that we consider in the Discussion.

Serial processing. Another way to overcome the constraints imposed by compositionality on the simultaneous representation of multiple items is to serialize processing. For example, the classic binding problem can be averted by sequentially processing one item at a time, focusing on its features in isolation of others. A large and longstanding body of empirical evidence suggests that this is exactly what people do (e.g., Shiffrin & Schneider, 1977; Treisman & Gelade, 1980), and has been variously interpreted as the function of attention (Treisman & Kahneman, 1984) and cognitive control (Botvinick et al., 2001). For instance, in the example above, focusing on one (set of) location(s) at at time, the covariance of red and square can be kept separate from the covariance of green and triangle. However, this comes at the cost of an increase in time needed to process all the items in the display — that is, at the cost of processing efficiency.

Summary of Information Theoretic Formulation

The effectiveness of a set of codes in a given setting — that is, when performing a given task or set of tasks — can be quantified as the mutual information between the representation of a given state (e.g., the stimuli in the visual display of Figure 1A together with the probe used to indicate which should be reported on a given trial) and the correct response(s) for that state.[11] The use of conjunctive codes for each possible state satisfies this need, and would allow processing efficiency to grow linearly with the number of states and or items that must be processed (e.g., objects in a display), insofar as all can be processed at the same time (i.e., in parallel). However, not only does this face the curse of dimensionality but, to the extent there is structure in the world, it is an inefficient form of representation. Accordingly, agents that are resource constrained (whether in computational power and/or time) face pressure to acquire and use structured representations — in the form of compositionality and semanticity — that

[11] Here, for simplicity, we assume that there is a "correct" response in any given setting, such as the accurate reporting of the memoranda in a memory experiment, or pressing the button assigned to a particular stimulus in a sensorimotor task. However, this can be generalized to "optimal" responses in settings where different options may be associated with different values and/or costs. Furthermore, whereas we focus here on tasks that require a single response, the approach can be readily generalized to circumstances in which a single task demands more than one response and/or multiple tasks each demanding a different response, that we consider in the General Discussion.

optimize representational efficiency. At the same time, “there is no free lunch” (Wolpert & MacCready, 1997): adding representational structure, in the form of compositionality and semanticity, introduces correlations into the code. While this can be accommodated over the longer term through forms of representational learning that may preserve processing efficiency, it is not possible to do so over the short term in order to deal with novel states — precisely the conditions under which representational structure is needed for generalization. That structure comes at the cost of processing efficiency. We refer to this fundamental tension between *representational* and *processing* efficiency as “Miller’s law.”

In the sections that follow, we provide an example of how the tension between representational and processing efficiency can be quantified; first in abstract form, and then in the form of a neural network model parameterized with psychophysically plausible properties, that we use to simulate the tasks on which Miller (1956) focused. We show that: i) while it is possible to partially mitigate the effects of representational structure on processing efficiency, even modest amounts of structure have a strikingly restrictive effect; and ii) over parameters that optimize representational efficiency, the restriction in processing efficiency manifests as a capacity limit that falls consistently around 2.5 bits; that is, restricting the number of items that can be processed to approximately 7, the number Miller considered so magical. Finally, in the Discussion, we consider the extent to which this principle explains similarly restrictive constraints observed in the capacity of other cognitive functions, such as multitasking and cognitive control; how these can be overcome by processes such as consolidation, automatization, or chunking; and how these constructs relate to issues in computer science (such as interpreted versus compiled procedures) and machine learning (such as tradeoffs between serial and parallel processing in distributed systems, and “in context” vs. weight-based learning in neural networks).

Results

Abstract Model

We begin with a simple abstract formulation of the problem, and show that whereas compositionally and semantically structured codes can be used to optimize representational efficiency and generalization, they severely constrain the number of individual items that can be processed at once. To demonstrate this, we consider a simple two dimensional world that has both compositional and semantic structure. We formalize compositionality by assuming that the feature values along the two dimensions are uncorrelated; and we formalize semanticity by treating each feature value along a given dimension as the center of a symmetric exponential

function, such that the similarity between two features is an exponentially decaying function of the distance between them (and has no similarity with features along other dimensions). The latter is grounded in Shepard's (1987) Universal Law of Generalization, and work showing that this can be derived from principles of efficient coding (Sims, 2018). Here, these forms of structure are reflected in the correlations among codes, that captures the proximity relationship of features along each dimension and orthogonality across them (see example in Figure 4). We assigned 256 feature values to each dimension, and constructed 65,536 items from all combinations of one feature from each dimension. Each item was represented by a concatenation of its two features, implementing a simple form of binding that, in principle, could avert the curse of generalization. However, we show that processing capacity was nevertheless severely constrained by the compositional and semantic structure of the codes.

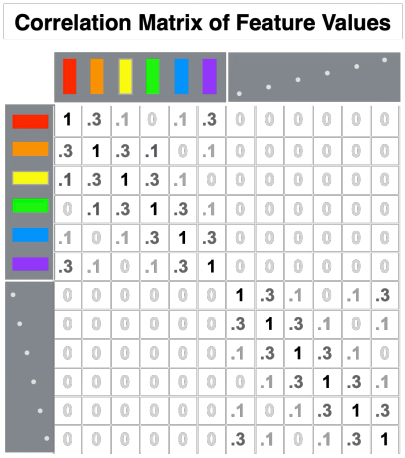

***Figure 4*. Compositional and Semantic Structure.** Correlations among features in a two dimensional space (labeled here, for expository purposes, as colors and spatial position), in which features are uncorrelated across dimensions (compositionality) and correlated within each dimension as an exponential function of their proximity along a circle.

We conducted two tests in this environment: i) a *similarity* test, in which we quantified generalization error to assess the ability to select which of two items a probe is most similar to along a given dimension; and i) an *identification* test, in which we quantified decoding error to assess the ability to identify a probe among a set of distractors. We modeled both tests as a probabilistic retrieval of items from a set given a single-dimensional probe, where the probability of retrieval of each item in the set was a normalization of the values in the correlation matrix set by the exponential similarity functions, and perturbed by noise (see Supplemental Information).

In the similarity test, we randomly selected test sets of two items (items sampled without replacement) from all possible items in the environment, and then sampled a third item as the probe uniformly from the set of all possible items [1, 256], and assessed the agent's ability to pick which of the two items in the test set was most similar to the probe. Because the probe is

one-dimensional, the most similar item is easily computed as the one with a value closest to the probe's value along the relevant dimension. This closest-match task assesses how well the representations would support generalization to novel features.

In the identification test, we constructed test sets of various sizes (sampling items without replacement), then randomly selected one of the items in the set as the probe and assessed the ability to identify the probe by reporting its feature along one dimension given its feature along the other. In both tests, we measured performance as the mutual information between the index of the correct response (i.e., an integer from 1… *n* for a set size of n) and the index of the agent's response. We computed this mutual information via a numerical simulation that estimated the joint probability distribution over the correct responses *X* and agent responses *Y,* then computing mutual information as:

$$I(X;Y) = \sum_{x,y} p(x,y) \bullet log \frac{p(x,y)}{p(x)p(y)} \quad (5)$$

$$= H(X) + H(Y) - H(X,Y) \quad (6)$$

where *H(X)* is the entropy of the marginal distribution of correct responses, *H(Y)* is the entropy of agent's responses, and *H(X,Y)* is the entropy of the joint distribution.

We examined the influence of two factors on these tests: *semantic structure* and *set size*. For both tests, we assessed the influence of semantic structure by manipulating the decay rate (spread) of the exponential distribution of correlation values over features within each dimension. This provided a measure of generalization error and, accordingly, *representational capacity* — that is, the extent to which semantic structure allowed a response to be selected that was as similar to the probe as possible. For the identification test, we also manipulated the number of items in the test set. This provided a measure of the extent to which decoding error increased as a function of the number of items being represented at once and, accordingly, *processing capacity*. The representational codes in this model are simply scalars and distances in the semantic space that were specified directly by the exponential gradient; the model thus abstracts away from vector codes and their lengths. Since code length was implicitly fixed in all cases, these measures of representational and processing capacity directly indexed the corresponding forms of efficiency.

Figure 5 shows the effects of both factors — semanticity (shown as thumbnail distributions of the exponential function for representative points), and test set size in the identification task, (shown as different colored lines) — on performance in each of the two tests, both in terms of mutual information (Panel A) and proportion of maximum available information (Panel B). The

plots illustrate the tension between optimally tuning semantic structure (spread of the exponential function) to maximize performance on the similarity task, and minimizing this structure to maximize performance on the identification task, by narrowing the spread as much as possible in order to separate perfect matches from non-matches. We consider each of these effects in more detail below.

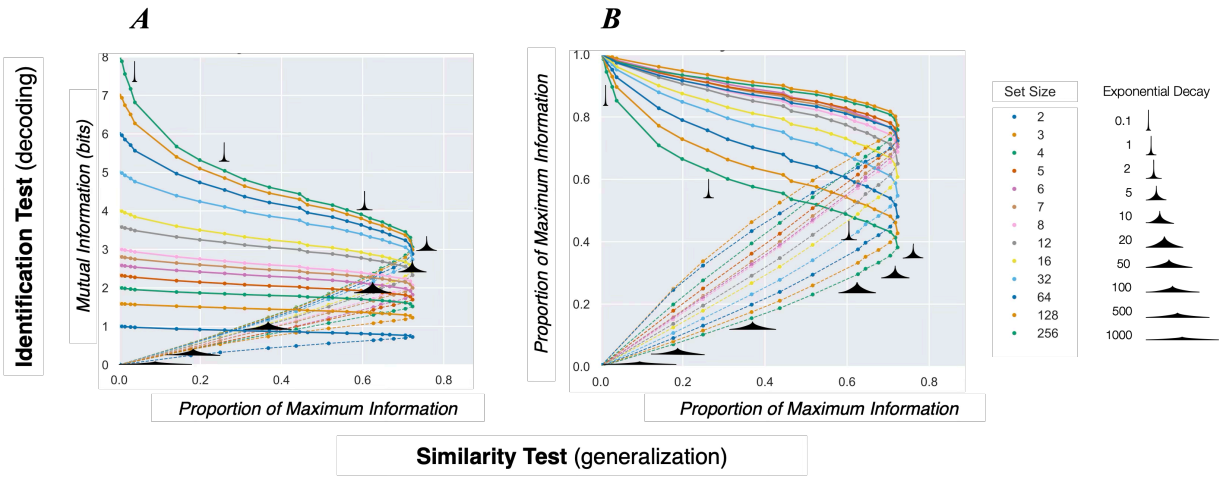

Figure 5. Tradeoff between representational capacity and processing capacity. Plots of information transmission in a similarity test (abscissas) used to probe representational capacity (generalization error) and an identification test (ordinates) used to probe processing capacity (decoding error); see text for details of test implementations. Points show different values of similarity among nearby codes (exponential spread of correlations shown in example thumbnails), and colored lines show different number of items in the test set (set size) used in the identification test. **(A)** *Mutual information* between the correct response and agent response (see text). For the similarity test, the maximum mutual information possible is 1 bit because the task required selecting which of two items most closely matched a probe. For the identification test with set size *n*, the maximum mutual information possible is log(n) bits; e.g. 8 bits for set size 256 (dark green). **(B)** *Proportion of maximum information,* normalizing for the total amount of information that varied with set size in the identification task. In both plots, solid lines designate a pareto front in the tradeoff between representational and processing capacity, and dotted lines show regions in which degradation of information occurs over both tests. Note that comparisons of capacity directly reflect relative *efficiency*, since code length was fixed (representational efficiency) and all items in a test set were always processed at the same time (processing efficiency).

Semantic structure (exponential decay) had a non-monotonic effect on generalization error in the similarity test: At the lowest levels (narrowest spread) all representations approached orthogonality, and at the highest levels (widest spread) representations became indistinguishable, both of which degraded performance on the similarity matching task. However, at intermediate levels, representations captured similarity structure within the relevant dimension, supporting good performance. The optimum (at an exponential rate of approximately 4.0) yielded a representational capacity close to the maximum possible.

In contrast, in the identification task, increasing semantic structure had a monotonic and dramatically degrading effect on performance. This reflects the tension between generalization error and decoding error considered above. Critically, it interacted with set size, with the relative impact of structure on decoding error increasing as the number of items (distractors) in the set

increased: With minimal semantic structure, performance was optimal for all set sizes (in accord with the Singleton Bound), increasing with set size up to 8 bits for the largest set size 256). However, as semantic structure increased, performance decreased — dramatically in the case of the largest set sizes — and converged on a range with an upper bound of about 3 bits for *all* set sizes. The precipitous loss in processing capacity is seen most clearly in Figure 5B, which shows performance as a proportion of the maximum information available in the task.

Note that the effect of set size for even minimal amounts of semantic structure suggests that compositionality played an important role in compromising processing capacity. It should also be noted that the exponential function used to implement similarity structure is sharply peaked relative to the spread of other possible functions, so that Figure 5 may present a conservative view of the impact of semantic structure on processing capacity. This is consistent with an analysis using a Gaussian function (see Figure S1 in Supplemental Information), as well as the results of simulations using empirically motivated codes that we report blow. Finally, note that the impact of structured codes were observed despite the fact that pairs of feature values along each dimension were conjunctively bound to one another (i.e., concatenated) *independently* for each item. That is, *compositional and semantic* representations impose a limit on processing capacity, irrespective of the ability to conjunctively bind such representations in context.

In summary, the effects of structured representations, and their interaction with set size, define a pareto front in the relationship between representational capacity (indexed by generalization error) and processing capacity (indexed by decoding error), along which they trade off. This pareto front exhibits a strong asymmetry, with substantially greater room to improve representational capacity relative to processing capacity than the reverse. This may explain why processing capacity is so consistently and severely constrained in tasks that demand generalization.

One might ask whether these observations extend to more complex environments and natural agents. The free parameters of the analysis were the size of the environment (including the number of dimensions and feature values along each), and the form of structure used for representations. In the next section, we show that simulations using empirically-derived parameters and neurally-plausible processing mechanisms exhibit strikingly similar effects, This suggests that the effects observed in Figure 5 reflect a fundamental underlying relationship between representational capacity and processing capacity in information processing systems.

Mechanistic Model

We implemented a mechanistic model to simulate behavioral performance in the tasks considered by Miller (1956), comprised of embedding, binding, memory, and response processes (Figure 6). The model implemented compositionally and semantically structured codes that conformed to theoretical considerations outlined above (utility for generalization), and that were informed by empirical data concerning neural coding. It also included a binding mechanism for rapidly associating feature values across dimensions to form a conjunctive representation of each item in a display. While these conjunctive representations allowed the model to process more than a single item at a time, we show that, like the abstract model presented above, it was still subject to the strict constraints on processing capacity imposed by the use of compositional and semantic codes.

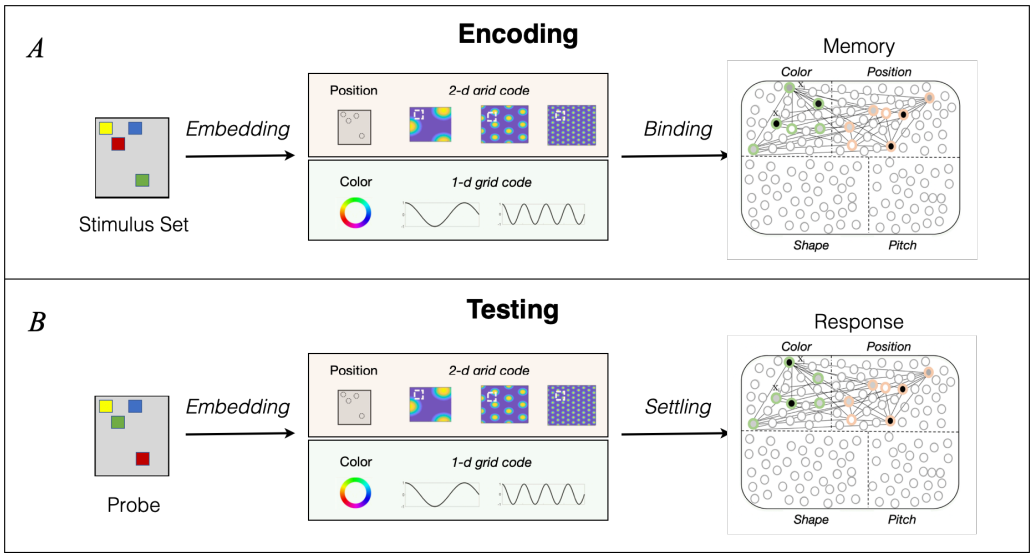

Figure 6. Schematic of MEME model. (A) Stimulus sets were represented (encoded) using semantic (grid) codes for feature values separately along each dimension (compositionally) that distinguished the items in the display (color and position), with those feature values represented in the model as patterns over nodes in a recurrent network. The network used Hebbian learning to associate the feature values belonging to each item. **(B)** The model was tested by presenting the probe display (here with color exchanged between two items), encoding it in the same way, and allowing the network to settle into a stable state, at which point the feature of the

Neural Network Implementation: The MEME model

Embedding using structured representations. The model used compositional representations over semantically structured feature dimensions (Figure 6A) similar to grid-like codes observed in medial entorhinal cortex (Hafting et al. 2005; Dordek, 2016; Stachenfeld et al. 2017; Wei et al. 2015). For example, an object's location was coded as a pattern of activity over location nodes representing sine waves of different frequencies and phases (Bicanski & Burgess, 2019) using an empirically observed scaling of frequencies of $\sqrt{e}$ (Fiete, Burak, &

Brookings, 2008). Thus, as in the abstract model above, similar states were represented by similar codes. The specific choice of codes was motivated by both theoretical and empirical considerations. Theoretically, they exemplify representational efficiency (Fiete, Burak, & Brookings, 2008; Wei et al., 2015; Stachenfeld, 2017; Chandra et al., 2023), minimizing the number of variables necessary to cover the space (Wei et al., 2015) while promoting generalization (Whittington et al. 2018; Frankland et al. 2019; Mondral et al. 2024). Empirically, grid-like codes have been found in a variety of species, including humans, to represent spatial as well as non-spatial dimensions such as sound (Aronov et al. 2017), olfactory stimuli (Horner et al. 2018), 2D visual arrays (Bicanski & Burgess, 2019), and abstract conceptual dimensions (Constantinescu et al. 2016). Here, we assumed that features such as color and number could be reasonably and usefully represented in this way as well.

Binding to create online, context-specific conjunctive representations. Grid cells in the model projected to a memory buffer in the form of a simple recurrent neural network (Hopfield, 1982), with nodes corresponding to the coding elements (e.g., grid cells) in the embedding layer, that could be used to rapidly bind the codes (along each dimension) belonging to each item in the stimulus set through Hebbian learning.[12] The network had *L* binary nodes, where *L* was the number of bits necessary to efficiently code for a stimulus domain (e.g., total number of grid cells needed to represent all the feature values along a given dimension). Each item in a display (*x*) was defined by a combination of feature values along several dimensions in the encoding layer (such as color and spatial location). Thus, every possible item could be represented as a unique pattern of activity (*Z)* over the nodes of the network. We assumed the representation of every item also included a feature that was associated with the current task context *c,* that was shared by all items in that context (e.g., trial), and with a distinct representation over the context nodes for each context in the experiment. Thus, Z_c was the representation of *x* in context *c.*

Finally, we assumed that only codes for items required to perform the current trial of the task were actively represented, together with their corresponding responses. Each item was represented as a conjunction of the codes for its task-relevant feature values along each dimension together with the current task context (x_c) and the corresponding response (Z_c) These conjunctions were dynamically generated for each context *c* (e.g., each trial of a task), and stored in the associative memory.

[12] Here, we focus on an implementation using a simple recurrent neural network and Hebbian learning for rapid associative binding through modifications of connection strengths (i.e., in "weight space"); in Supplementary Information we show that similar effects are observed if, instead, tensor product representations are used to bind representations as unique patterns of activity (i.e., in "state space") that might reflect use of working memory rather than episodic memory as the store, highlighting the generality of the coding and information processing principles involved.

Processing capacity. In information theoretic terms, the embedding layer and associative memory constituted a *processing channel (*that is, a mechanism for encoding information about the display and task demand, and "transmitting" this to the selected response). A trial (or context *c)* constituted a *use* of this channel, during which one more more items were encoded and used to generate a response. We sought to evaluate the processing capacity of this channel, by quantifying how many items could be stored simultaneously while maintaining their identity. The conjunctive representations *z* were formed by updating the weights among the nodes in the network corresponding to each item to be stored, so as to minimize an energy function (i.e., activation of the nodes corresponding to the feature values of *x* and the context *c*; see Figure 1B). Hebb's rule for weight updates implements this by capturing which feature values vary (are relevant) and co-vary (are related to one another) in the observed data, thus insuring the *m*aximum *e*ntropy estimate of the statistics over the distribution of $z(x_c)$, while *m*inimizing the *e*nergy of each representation (Mackay, 1991; Amit 1988, respectively). Accordingly, we refer to this as the MEME model. The updating of weights was specific to and independent for each *c*. In summary, the set of conjunctive representations *z* needed to represent the specific set of stimuli *x* that occurred in a given context *c* were formed over the compositional codes (nodes of the network) by binding the nodes representing the relevant feature values of each item using the weights of the Hebbian network, and then resetting those weights for each new context.[13] While the compositional codes used to represent the feature values along each dimension were assumed to have developed over a long time frame, the weight updates used to bind those feature values to form the conjunctive representations for each item occurred on the time-scale of experimental variation — that is, within a specific context *c* — and were specific to that context. This could be a single trial, as in working memory tasks, or over the set of trials in an experimental condition, as in the perceptual judgment and numerical estimation tasks that we consider below.

Under the assumptions above, the upper bound on the processing capacity of the system (that is, how many conjunctive representations of items (x_c) can be simultaneously and distinguishably stored) is defined by the well-known capacity constraints of a Hopfield network (Hopfield, 1982; McEliece et al. 1987; Sompolinsky, Amit, Guttfreund, 1985; Amit, 1989). This has a critical point (p), based on the number of items x_c that have been encoded in z_c, (N) at

[13] Note that this differs from the standard use of Hebbian learning, such as its use in models of episodic memory (e.g., Norman & O'Reilly, 2003), in which weight adjustments are allowed to accumulate across contexts experienced by the agent. In the Discussion we consider how the use of context-constrained weight updates for associative binding may arise from an interaction between durable representations in episodic memory and other psychological functions that mediate context-dependent processing (such as attention and cognitive control; e.g., Giallanza et al., 2024) and related constructs in machine learning (such as the use of external memory; Graves et al. ,2024). In Supplemental Information, we consider an alternative mechanism for associative binding, that uses activity patterns based on tensor product representations (Smolensky, 1990) rather than weight updates, providing a mechanism by which such bindings may be temporarily represented for a given context in working memory, and then replaced in the next context (consistent with our model), rather than represented more durably in episodic memory.

which the representation for any given item is dramatically less likely to be retrieved. When the set size *)/L* is less than *p*, the equilibrium states of the network are likely to correspond to previously experienced representations. When *N/L* is greater than *p*, they may reflect statistical summaries of those representations, rather than the identity of the representations themselves. That is, the processing capacity is limited by *p.*

Critically, however, *p* assumes that all of the stored memories are independent of one another. If the codes used to construct the memory are *not* independent (and thus the conjunctive representations are less distinguishable), then the processing capacity will be further reduced. In either case, retrieval accuracy degrades, as does the time to retrieve any one of these at a given level of accuracy. In the simulations of cognitive tasks presented below, we consider these as costs in the following optimization problem: Given a set of compositional codes, what is the maximum number of conjunctions among them that can be assigned in a given context *c* (i.e., that maximizes its processing capacity), while minimizing degradation in performance. We show that this obeys remarkably similar constraints to those observed in the abstract formulations presented above.

Behavioral Performance

We applied the model outlined above to the three tasks highlighted by Miller (1956): short-term (or "working") memory, absolute perceptual identification, and numerosity judgment. Though these three tasks operate over different time-scales and different domains of information, in each case the task-relevant input-output mappings ($x_c \rightarrow y_c$) require mediating conjunctive codes (z_c) that, we assume, were not represented in the system's existing set of codes. Instead, they must be constructed compositionally, as combinations of existing codes, such as representations of color and location, or reference tones and their ordinal labels. Here we provide a brief summary of the empirical findings, and our models' account of them. The Supplementary Information provides greater detail concerning model implementation and simulations of the experimental task.

Immediate Memory

Figure 7 shows the model for a classic cued change detection paradigm, variants of which are widely used in the study of visual short term memory (e.g., Luck & Vogel, 1997; Wilken & Ma, 2004; Bays & Husain, 2008; Sims, Knill, & Jacobs, 2012; Luck & Vogel, 2013; Bays, 2015), that we use to evaluate the capacity of what Miller referred to as immediate (and now more commonly referred to as working) memory. In its simplest form, the task involves a brief

presentation of a visual display containing a number of items (e.g., shapes of different colors at various locations), followed by a delay (~1000 ms), and then a second version of the display in which a target item is cued, and the participant must respond by indicating whether the color of that item has changed (see Figure 1A). Both the number of items in the display ("set size") and, critically, the combination of feature values for each item are varied randomly across trials. Memory performance is ubiquitously found to decline as a nonlinear function of set size, with precipitous declines typically in the range of 3 or 4 items for simple combinations of color and shape (Luck & Vogel, 1997).

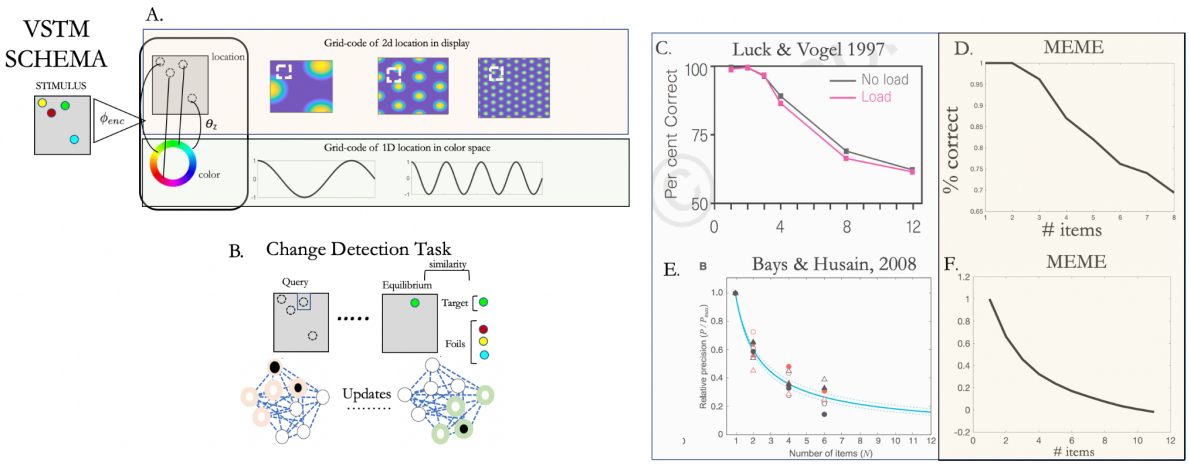

Figure 7. Visual Short Term Memory (VSTM) phenomena. **(A)** We assume features of a visual stimulus are factored into separate representational streams and re-combined. Locations are represented by grid-like codes of the 2D array and colors are grid-like codes of the 1D color space. The stimulus-specific weights (W) reflect *what color was where* in a particular image. **(B)** We present the network with a location code and allowing it to evolve until a local energy minimum. Changes are reported on 50% of change trials in which the correct color is not within a Hamming distance of 0.05 from the target. **(C & D).** The model's performance closely tracks Luck & Vogel's (1997) observation of qualitative change in performance at ~3 items **(E,F)**. In this framework, representational precision also decreases as an approximate power law, as observed in Bays & Husain (2008).

To simulate performance of this task, each item in the visual display was embedded in the model as a binary pattern over the nodes coding for each task-relevant stimulus dimension. Here, the nodes coded for *spatial location* and *color* (Figure 7A). On each trial, the network's weights were updated to associate *what was where* on that trial (Figure 7B). Thus, these weights served both to represent and store the set of bindings on that trial. We then tested the network's ability to use this information by presenting it with the location code for the target item as input, and allowing its activity to evolve until equilibrium (based on the input and the bindings encoded in its weights), and quantified performance using the Hamming distance between the equilibrium state and the correct representation of the color at the cued position. The code with the shortest hamming distance to the settled state was selected as the network's response.

Figure 7C,D shows that, with location codes derived from empirical data on spatial acuity (Westheimer & Beard, 1998) and color discriminability (e.g., Long et al. 2006), MEME captures the form of human error rates as a function of set size: Nearly perfect performance up to 3 to 4 items (the "capacity-limit"), followed by a dramatic decline. Furthermore, as observed in more recent work (e.g., Wilken & Ma, 2004; Alavarez & Cavanagh, 2004; Bays & Husain, 2008; Sims et al., 2012), the representational precision of these stimulus features decreases as a power law-function of set size: The more items that were present in the image, the greater the Hamming distance of the equilibrium state from the target (Figure 7E-F). The model thus captures phenomena central to both slot and resource models.

Absolute Identification

We used the same model architecture to address capacity limits observed in the absolute perceptual judgment tasks considered by Miller (1956). In such tasks, participants were first presented with a reference set of auditory stimuli (e.g., tones) and associated ordinal rankings along a single underlying dimension (e.g., frequency). Then, throughout the experiment (usually lasting around an hour, though in some cases, weeks (See Shiffrin & Nosofsky, 1994), they were presented repeatedly with test stimuli drawn from the reference set in random order, and tested on the ability to identify the corresponding rank. For example, the reference set might have consisted of nine tones evenly spaced between 1,000 and 5,000 Hz and their corresponding ranks (e.g., 1 for 1000 Hz, 2 for 1500 Hz, etc.), for which the participant should have responded "2" to a test tone of 1500 Hz. The limit to the number of such unidimensional perceptual stimuli that humans can reliably identify is strikingly similar to the capacity limits in working memory — 7 plus or minus 2 (or about 2.5 bits; e.g., Pollack, 1952; 1953).

To simulate performance of this task, each item in the reference set was embedded in the model as a binary pattern over nodes coding separately for tones and integers, again representing each using a grid-like code comprised of frequencies and phases (Figure 8A). We derived the grid-like codes for tones based on the range sampled in Pollack (1952,1953) and a simplified uniform just noticeable difference (JND) of 0.05% over the range of 20 Hz to 20 kHz (Sek & Moore, 1995), as well as the same general grid-cell scaling properties (Stensola et al., 2012; Wei et al., 2015) used in the short term memory model. Also as in that model, the combination of tone and ordinal position for each item in the reference set was stored in the weights between the nodes representing the two relevant features. To evaluate the network, the grid code for the tone of a test stimulus was presented as input, activity states were allowed to evolve to equilibrium, and the representation of the integer code with the shortest Hamming distance to the settled state was chosen as the network's decision.

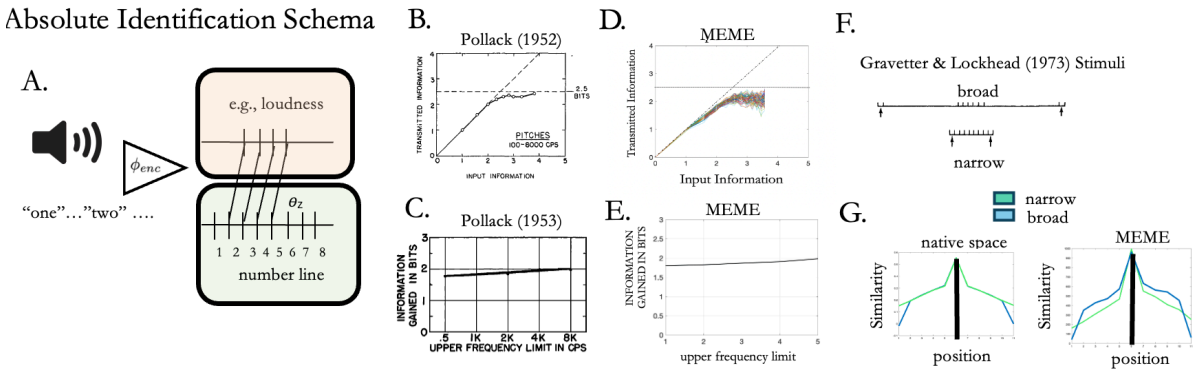

Figure 8. **(A)** Schema of Absolute Identification task. Subjects are presented with a set of stimuli along a target dimension (e.g. pitch) together with associated ranks. The task is to later report the corresponding rank when queried with a particular stimulus. To model these phenomena, we assume1D grid-like code for perceptual dimensions as well as the mental number line. **(B)** Pollack found a limit of ~2.5 bits of information for absolute identification of pitch, highlighted in Miller (1956). This limit is approximately scale invariant **(C)**, as increasing the absolute difference between pitches has little effect on discriminability. Our model (averaged over 1000 trials) reproduces the limit **(D)**, and the approximate scale-invariance effect **(E),** to the point of predicting the slight linear increase. **(F)** However, this depends on the relationship between sampling distribution and precision. For example, "broad sampling" to include stimuli near the min and max of the perceivable range (Gravetter & Lockhead, 1973) reduces precision on the middle items, relative to narrow sampling. **(G)** Our model predicts this pattern, as broad sampling introduces low-frequency redundancies into the codes, causing increased errors.

Once again, the model exhibited the empirically observed capacity limit of approximately 2.5 bits (Figures 8B and 8D). Furthermore, the model also reproduced the observation by Pollack (1953) that, when items are sampled uniformly within a given range, the capacity limit is largely invariant to the scale of that range (Figures 8C and 8E). The network exhibits this approximate scale invariance because its computational dynamics are influenced only by the coding elements that vary and co-vary *within-context*: As the sampling range is gradually increased, the number of within-context variables that vary and co-vary may increase, leading to a small increase in capacity. However, given the geometric progression of the frequencies (Wei et al. 2015), the difference in L is small and largely offset by the coarseness (imprecision) of the low-frequency variables. Importantly, however, Gravetter & Lockhead (1973) also observed that, when samples are *not* uniformly distributed, precision is greater for stimuli that are closest to others ("narrow" condition) than ones that are further ("broad"). Figure 8, panels F and G show that the model reproduces this effect, which has been observed in a variety stimulus domains (Braida & Durlach, 1972, Rouder, 2001). In the model, the broad sampling context causes additional within-context variation in low-frequency grid-codes, which are inherently redundant with respect to nearby stimuli. Statistically, broad sampling decreases precision for the same reason that increasing set size decreases precision in short term (working) memory tasks (Wilken & Ma, 2004; Bays & Husain, 2008): both introduce correlations in the code, and these correlations increase the error rate for a fixed code length.

Numerical Estimation

Finally, following Miller (1956), we "cannot leave this general area" (pg. 90) without considering limitations in the ability to judge the number of items in a visual display. There, participants show fast and nearly perfect numerosity reports when the number of items is small — referred to as the "subitizing" range — but increasingly slow and error-prone reports as that number increases past some limit (Kaufman et al., 1948; Mandler & Shebo, 1982). Early results put the limit at about 6 (Kaufman et al., 1948), in the range of Miller's 7 plus or minus 2, though decades of subsequent work has found a limit often closer to 3 or 4 (Mandler & Shebo, 1982; Trick & Pylyshyn, 1993; Revkin et al., 2008; Cheyette & Piantadosi, 2020), with variation depending on features of the experiment (e.g., Revkin, 2008; Cheyenne & Piantadosi, 2020).

To simulate performance of this task, items in the display were embedded in the model as binary patterns over nodes once again using a grid-like code comprised of frequencies and phase, in this case to code for position and number (Figure 9A). The position codes were parameterized to cover the size and resolution of the 2D spatial display, and the integer codes were chosen to span the range of numerosities sampled (e.g., 1-12), and that constituted the range of possible responses in the task.

Note that this task differed from the preceding two tasks in that all of the information necessary to perform the task was available from the display when the participant was required

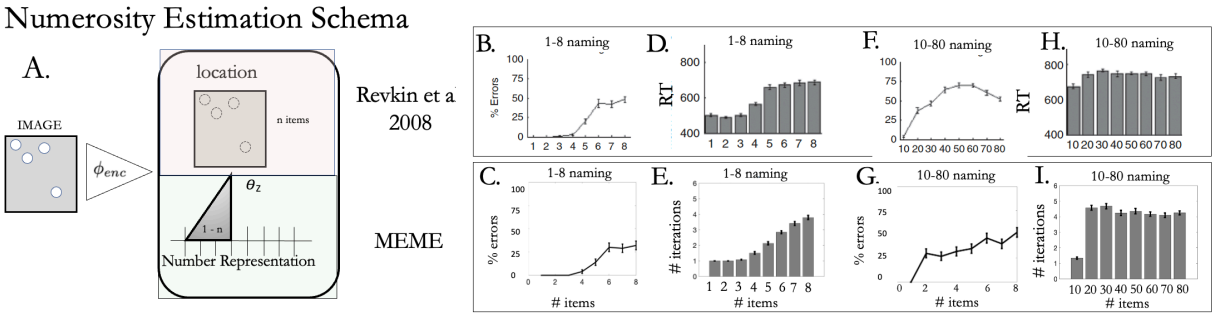

Figure 9. Numerosity estimation. **(A)** Stimuli are visual displays consisting of sets of objects of size N, which varies across trials. Subjects' task is to report the cardinality of the set, and error rates and response times are collected. To model this task, we assume that agents factor the stimulus into a spatial representation (grid code of 2D space) and a number line representation that must then be related ("how many?"). To evaluate the model, the network is queried with location codes, and allowed to evolve until equilibrium to identify the most similar number-code. **(B).** Example human behavioral data from Revkin et al. (2008) showing standard subitizing effect in error rates. **(C)** Like humans, the model shows nearly perfect performance when $N<=3$, but declining performance thereafter. (**D&E)** Likewise, the canonical human RT signature is qualitatively predicted by the number of iterations (settling time) through the Hopfield network. **(F,G,H,I).** Both human and model performance depends on the experimental context shaping the statistics in The weights encode the network's knowledge of this relationship. (1-8 naming left vs. 10-80 naming right). The capacity-limits stem from (1) the correlations in the spatial structure as *N* increases and (2) the decreasing precision in the number code with increasing cardinality (itself determined by correlational structure on a slower time scale, as in the absolute identification task).

to respond, without the memory requirements imposed by either of the other tasks considered above. Nevertheless, we assumed it required the same form of structured (i.e,. generalizable) codes and binding mechanisms to represent the arbitrary (i.e., novel) pairing of the number of items and their positions used in the task. Accordingly, the model's representation of the display was generated as follows: For each of the *N* items in the display, item $i \in [1 \rightarrow N]$ was selected randomly without replacement from the set of remaining *(N-i)* items, the code for its position was determined and added to the sum of the codes for the preceding *N-1* items, that sum was paired with the code for the number *i,* and the pair was stored (bound) in the weights of the associative memory. This provided the model with a set of numerosity representations for the display, each of which was the conjunction of a number code with a representation of the simultaneously embedded positions for the corresponding number of items. The model was tested by encoding the display as described above, then presenting it as input and allowing the network's activity state to settle, and then taking the representation of the number code with the shortest Hamming distance to the settled state as the network's response regarding the number of items in the display (Figure 9B).

Similar to the simulations of the other two tasks, the model exhibited an abrupt change in error rates for displays with 3 or 4 items, closely matching the empirical data for human participants (see Figure 9C,D). In addition to the error curves, we computed an estimate for response times (RTs) as the mean number of updates required to reach equilibrium. The qualitative form of RT as a function of set size closely tracked empirically observed RT curves (e.g., Kaufman, 1948; Mandler & Shebo, 1983; Trick & Pylyshyn, 1993; Revkin et al., 2008; Cheyette & Piantadosi, 2020): For displays of 3 items or less, RTs are almost flat, after which they increase monotonically (see Figure 9E-F). Furthermore, as with humans (Revkin, 2008), the error rate and RT functions depended on the range of numerosities used in the task. The observed functions, for both humans and the network, are markedly different when possible responses range from 10-80 in intervals of 10 than 1-8 in intervals of 1 (See Figure 9G-J). This pattern has previously been interpreted as inconsistent with single system accounts of subitizing that predict that numerosity estimation should follow a Weber-fraction, exhibiting a consistent logarithmic decrease in precision as a function of numerosity (e.g., Gallistel & Gelman, 1990). Our findings question that interpretation, showing that a single system — motivated by normative coding considerations — can capture this phenomenon. This extends previously proposed unified efficient coding model of numerosity estimation (Cheyenne & Piantadosi (2020), to also account for context-specific effects, and do so with a candidate mechanistic model.

Discussion

Summary

The constraints on perceptual processing and immediate memory capacity, famously referred to by Miller (1956) as the “magical number 7,” have largely remained a mystery. This might be considered even more perplexing, given that it seems to extend beyond tasks involving perception and memory, as considered by Miller, to ones involving other functions that have been equally foundational in the study of cognition and the brain, including attention, multitasking capability and cognitive control (Posner & Snyder, 1975; Shiffrin & Schneider, 1977; Treisman & Gelade, 1980). Here, we have offered a unified account of these constraints from an information theoretic perspective, which arise from a fundamental tradeoff between the value of structured codes for efficient acquisition and flexibility of generalization on the one hand — that is, the optimization of *representational efficiency* — and, on the other hand, the correlations among codes that this introduces, which limit the ability to simultaneously represent and process codes for multiple independent items — that is, that restrict *processing efficiency*. We have shown that: i) even modest amounts of structure, that improve representational efficiency, dramatically restrict processing efficiency; ii) this is observed for psychophysically and neurally informed types of structure and processing mechanisms; and iii) this account can provide a unified explanation of the restrictive constraints on performance observed across the range of tasks that puzzled Miller (1956).

In the remainder of this article we consider several questions raised by this account: i) How does it relate to the information theoretic treatment of capacity limits presented in Miller’s paper? ii) How does it relate to previous applications of information theory to learning and representation in cognitive science and neuroscience? iii) How does our implementation of binding relate to standard psychological and neural mechanisms thought to be responsible for associative memory? iv) To what extent can our account explain similar constraints in other cognitive capabilities, such as multitasking and cognitive control? v) How can these constraints be overcome? vi) How do our observations relate to related issues in statistics, machine learning, and the use of neural network architectures in artificial intelligence.

Relationship to Miller’s Analysis

Miller’s (1956) article was perhaps the most influential one to call attention to the severe limits on immediate memory, indelibly imprinted in most people’s minds — as it was in Miller’s — as the magic number 7. Oddly, it is also known mostly for that observation alone — though

he assiduously referred to those who originally made it (Hayes, 1952; Pollack, 1953) — rather than for the observation that consumed him, and motivated him to refer to the number as magical: the diversity of settings and forms in which it appears as a constraint, and the possibility “that there may be something deep and profound behind all these sevens, something that is calling out for us to discover it.“ Miller was inspired by the initial surge in applications of information theory to perception and memory following Shannon's seminal 1948 paper, and intrigued by the prospect of using information theoretic measures to solve the mystery of “all the sevens."

In his effort to solve the mystery, Miller focused on a set of tasks that had been used to assess the channel capacity of human cognition as an information transmission system and, accordingly, sought to quantify the capacity of memory and perceptual judgment in terms of the amount of information transferred, in terms of bits. He accepted this as a reasonable measure of capacity for perceptual processing (i.e., performance in the absolute judgement and numerosity tasks), but ran into trouble when applying it to immediate memory, where he noted that the number of bits transmitted seemed to vary widely depending on the items to be remembered. For example, decimal digits carry about about 3.3 bits each, so remembering 7 digits (e.g., a birth date) actually amounts to remembering 23 bits of information; and English words carry about 10 bits, so remembering 7 English words amounts to remembering ~70 bits, and so on. This led Miller to conclude that the constraint on immediate memory is not the amount of information conveyed by the memories, but rather the number of *familiar items* that must be remembered, which he referred to as “chunks.” Without a similar construct for perception, and without a formal characterization of a chunk, he expressed suspicion that the constraint of 2.5 bits for perception and the 7 chunks for immediate memory was “only a pernicious, Pythagorean coincidence.”

The work we present here suggests that the disparities Miller observed across tasks can be resolved by taking account of representational structure. As noted earlier, information theoretic accounts typically consider only the *probabilistic* structure of *independent* events, without considering the extent to which there may be *meaningful correlations* among these that can be captured by including structure in the codes to represent them. Much like Mathy & Feldman (2012) and Nassar et al. (2018), we suggest that such correlations provide a formal grounding for what Miller referred to as a “chunk” (and what we define as an *item*): a pattern of covariation over a set of feature values along one set of dimensions (e.g., color and shape) that is shared over a range of values along other dimensions (e.g., spatial or temporal). This embraces objects in a display (such as in Figure 1A), but also the digits of a birthday, or the letters of a word. Crucially, we show that the pervasiveness of the ~2.5 bit constraint can be explained by taking

account of how representational structure affects decoding error, and how this interacts with the number of items that must be processed at the same time.

From this perspective, Miller's chunks can be viewed as conjunctive (unstructured) codes dedicated to particular items, that have been been acquired through repeated exposure to the correlations among feature values that define those items. While such codes are not *representationally* efficient (i.e., with respect to learning and generalization), the lack of correlations among them increases their coding distance, allowing them to be used at the same time without risk of interference (decoding errors), and thus increases the rate with which information can be transmitted — that is, they afford *processing* efficiency. Conversely, when such conjunctive codes are not available — that is, when the items to be processed are comprised of novel combinations of feature values for which the correlations have not been learned — then structured codes must be used for generalization. However, the structure of such codes reduces their minimum coding distance, introducing the potential for interference (decoding errors) that limits processing efficiency. These are exactly the conditions imposed by the tasks on which Miller focused. Furthermore, the formal analyses and simulations we presented indicate that these conditions are subject to a remarkably consistent constraint of approximately ~2.5 bits in the *rate* of information processing, and that this can be traced to a strikingly strong asymmetry in the effects of coding structure on representational efficiency (generalization) versus processing efficiency (information rate): Even the modest amounts of structure required to promote generalization have draconian effects on processing efficiency.

In sum, inspired by Miller's information theoretic approach, and his observations of a provocative similarity in processing constraints among tasks that probed perception and memory, we arrive at a conclusion that differs from his, but one that we suspect he might have liked: The similarity in constraints does indeed reflect "something deep and profound behind all these sevens," a relationship between representation and processing in which gains in representational efficiency come at a disproportionate cost to processing efficiency, along the lines of:

$$(\textit{Representational efficiency})^k\ (\textit{Processing efficiency}) \propto c$$

which defines an envelope of performance within which any information processing system must operate — a principle that we suggest reflects Miller's Law. We have shown that, with even a modest premium placed on the efficiency of learning and flexibility of generalization afforded by representational efficiency, processing efficiency is constrained — in number of independent items that can be processed at once — to the number that so intrigued Miller.

However, if it is truly a general principle, it should apply not only to the particular phenomena that intrigued Miller but, more broadly, to a wealth of related data and models that have accrued in the nearly seven decades since his landmark article. In the remainder of this Discussion, we consider some of these connections.

Relationship to Rate Distortion Theory (RDT)

RDT is a major branch of traditional information theory that addresses coding efficiency, a central tenet of which is that it is possible to reduce signal distortion by a commensurate increase in code length (i.e., rate, in bits per symbol; Shannon, 1959). As noted earlier, an implication of the Singleton Bound is that similarity structure can be viewed as a form of distortion with respect to decoding error, insofar as it reduces the minimum distance between codes (see Figure 3) which, in turn, increases decoding error. RDT suggests that a commensurate increase in code length can compensate for this — a consideration that we turn to below, as a way of formalizing the effects of consolidation, automatization, and "chunking" noted above. However, as emphasized from the outset, we are interested here in cases for which code length is *fixed* — both for the purposes of analysis, and under the assumption that modification of code length requires time, and thus cannot contribute meaningfully to the rapid and flexible processing of novel states. Under the constraint of a fixed code length, the use of structured code extracts a cost in decoding error in exchange for the benefit of generalization, as captured by the expression above. From the perspective of RDT, this suggests that, using decoding error as the loss, it should be possible to quantify the increase in code length that would be required to compensate the distortion introduced by similarity structure. This is consistent with findings reported by Sims (2016), who used RDT to analyze the efficiency of coding (in terms of bits) with respect to a loss function akin to decoding error. He found that people consistently use less than optimally efficient codes in the immediate memory and perceptual judgement tasks of interest to Miller. We suggest that this can be explained by the use of structured codes required to support generalization for the processing of the arbitrary stimuli used in those tasks. This, in turn, is consistent with his finding that human performance was better fit with a loss function that takes account of distanced-based similarity (akin to generalization error). Here, we have made a direct connection between these two observations — in terms of the tradeoff between generalization error and decoding error — by showing that the use of structured codes directly and profoundly constraints processing efficiency.

Information Theoretic Accounts of Learning and Representation

Our work builds on a longstanding and influential tradition of theoretical work in cognitive science and neuroscience, inspired both by statistical physics and information theory (Barlow, 1961; Linsker, 1988; Friston, 2011) suggesting that the brain implements efficient coding to maximize information preservation, given an agent's resource constraints. Here, we argue that by extending this approach to take account of similarity in additional to probabilistic structure in optimizing representational capacity can help explain an apparent paradox in human cognitive function: on the one hand, the remarkable range of information over which it can operate, as well as the flexibility it exhibits in responding to novel information; and, on the other, the equally striking constraints in processing capacity it exhibits. Most previous work has focused on the former — that is, optimal coding that maximizes the ability of the system to come to accurately represent as much information as efficiently as possible. Furthermore, it has focused largely on probabilistic structure, wherein optimal coding is achieved by adapting code length to the frequency of events. The objective of such forms of optimization is to minimize *decoding error* — the frequency with which the selected code is not the *same* as the correct one. Here, we have pointed out that an equally important goal is representing the *similarity* structure of the world as efficiently as possible, the objective of which is to minimize *generalization* error — the *distance* of the selected code from the correct one. This follows from Shepard's (1987) Universal Law of Generalization, and relies on the use of structured codes. Critically, we show that minimizing generalization error is in direct tension with the traditional focus on minimizing decoding error, placing the optimization of representational efficiency — how many independent things can be represented — in tension with the optimization of processing efficiency — how many independent things can be represented *at once.* We have shown that this tradeoff can explain the conditions under which cognitive function is constrained — viz., when it relies on the use of *semantically* and *compositionally structured* codes to process novel items — and the remarkable severity of these constraints even when the amount of structure introduced to the codes is relatively modest.

Our account also aligns with analyses of mixed selectivity in neural coding, their use in optimizing the tradeoff between generalization and discrimination, and the implications this has for the relationship between neural representation and task performance (e.g., Barak et al., 2013). Our work provides an information theoretic interpretation of this relationship, and applies it to the performance of tasks involving perception and memory that have figured centrally in

theorizing about cognitive capacity. More generally, by extending the information theoretic account of optimal coding to address representational structure, and the constraints this imposes on processing efficiency, our work continues in the rich traditions of cognitive and brain science, by framing an understanding of capacity constraints in terms of the nature of the representations on which processing relies, and formulating this understanding in quantitative terms. Advances in both brain imaging and computational modeling offer the promise of exploiting such quantification (e.g., by measuring patterns of correlation among neural representations) to ground a theoretically rigorous and quantitatively precise understanding of the envelope of our capabilities in empirical data.

Our approach may also provide a useful point of contact with an influential theory concerning capacity constraints on working memory (Wilken & Ma, 2004; Bays & Hussain, 2008; Ma, Hussain & Bays, 2014), that explains these in terms of a tradeoff between representational precision and load. According to this theory, increasing the precision with which items are represented increases accuracy, but at the expense of the number of items that can be represented at the same time. In information theoretic terms, increasing precision can be cast as an increase in the code length used to represent an item which, for a given total code length, will limit the number of items that can be represented at a given time and for a given level of decoding error. This theory suggests that one source of the empirically observed envelope of constraints on precision and load may be the metabolic costs of the code. Our approach complements this theory, pointing to the *structure* of the code as another critical factor that, in addition to code length, can constrain load. Furthermore, our results suggest that the benefits of representational structure for generalization, and its strikingly restrictive effects on load, can provide a normative account of constraints, independently of code length (i.e., representational or metabolic resources). Nevertheless, understanding the interaction between these factors is clearly an important direction for future investigation.

Relationship to Neural Mechanism for Associative Memory

The ability to rapidly associate (bind) compositional codes to represent novel items plays a central role in our account, that we assume is supported by some form of associative memory. This raises two closely related questions: How do such associative mechanisms relate to traditional cognitive constructs such as episodic memory (EM) and working memory (WM), and what neural mechanisms might be responsible for these?

On the one hand, rapid associative binding is a hallmark of EM, that is widely believed to be supported by the hippocampus (e.g., McClelland et al., 1995), and perhaps other structures

such as the cerebellum (Musslick et al., 2023; Webb et al., 2024). Our mechanistic model using a recurrent neural network and Hebbian learning aligns with models of such structures (e.g., Hasselmo & Wyble, 1997; Norman & O'Reilly, 2003; Spens & Burgess, 2024). Furthermore, one of the most influential theories of EM function, the Temporal Context Model (TCM; Howard & Kahana, 2002), asserts that entries in EM are accompanied by a temporal code, usually assumed to be a slowly drifting signal, that makes it possible to identify when an event occurred and its temporal relation to other events. However, this also introduces characteristic patterns of memory errors, arising from correlations in the temporal code, that have received extensive empirical support (e.g., Sederberg et al., 2008). The presence of such structure in EM, along with both its advantages and disadvantages, are consistent with its role as a mechanism the for rapid associative binding of structured codes that is central to our account.

On the other hand, it has also been proposed that rapid associative binding may occur among representations actively maintained in WM, whether using similar Hebbian mechanisms (Oberauer et al., 2012), phase coupling among oscillating representations (Hommel, 2004; Roux et al. 2012; Jensen & Lisman, 1998), or through the online formation of tensor product representations (Smolensky, 1990) (as we describe in the Supplemental Information).

One fundamental difference between EM and WM is the durability of the associations formed among representations: a fundamental feature of EM is that such associations are stable, and remain accessible for considerable periods of time after they are formed (hours to years), a feature that is assumed to reflect their formation through modification of synaptic *weights*. In contrast, WM is generally assumed to rely on the transient maintenance of information in activity *states:* associations formed in WM are inaccessible once the corresponding representations are no longer active, presumably after short periods of time (minutes if not seconds; though see Stokes, 2015 for the possibility of "activity-silent working memory"). Accordingly, use of EM versus WM for binding should make different predictions about the influence of the associations formed on subsequent processing over different temporal intervals — a factor that has received some attention (Beukers et al., 2021), but remains an important direction for future research. More realistic is the possibility that both mechanisms play a role in binding, and that interactions between them contribute to the efficiency and flexibility of human cognitive function, a possibility that has begun to attract attention both in cognitive neuroscience (Giallanza et al., 2024) and machine learning (Graves et al., 2014). One particularly interesting possibility is that the rapid binding function, that may have evolved under the pressures for optimization discussed here, provided a platform for the emergence of symbolic computation in the brain, an idea that has also begun to attract

attention in cognitive neuroscience and machine learning (Webb et al., 2021; Webb et al., 2024).

Relevance to Language Processing

Language is perhaps the most widely recognized domain in which there is pressure for structured, compositional representations that support generalization (Fodor & Pylyshyn, 1988). Therefore, based on our considerations, this should be closely accompanied by extreme limits in capacity when multiple similar representational codes must be processed simultaneously. Such limits were first pointed out by George Miller in collaboration with Noam Chomsky (Chomsky & Miller, 1958): we seem to have striking limits in our ability to process syntactically well-formed sentences that require the maintenance and discrimination of items with highly similar structure, such as in self-embedded clauses (e.g., “the mouse the cat the dog bit chased hid”). Modern accounts suggest that the need to manage long-distance dependencies — syntactic, semantic and referential relations that may stretch across dozens of words or even sentences — requires that linguistic representations be actively maintained in working memory which, coupled with similarity-based interference among those linguistic representations, give rise to the severe capacity limits observed in language processing (Gibson, 1998, 2000; Lewis, 1996; Lewis et al., 2006).

The challenge of reliably handling long-distance dependencies in natural language led directly to the development of architectures such as the Transformer (Vaswani et al., 2017), and the large language models (LLMs) based on them. Such models provide a useful test of the extent to which the tension between representational efficiency and processing efficiency is a general principle. On the one hand, they have massive numbers of parameters, suggesting that processing resources should not be a source of constraint. On the other hand, they were designed to handle the online processing of long-distance dependencies, and trained under heavy pressure for generalization, suggesting that they should make use of structured codes and, on our account, that these factors should together induce strict constraints on processing. Remarkably, there is growing evidence that such models do indeed exhibit constraints in processing capacity similar to humans when tested under conditions that elicit these in humans: they have difficulty processing sentences involving self-embeddings (and their attention patterns reflect this; Ryu & Lewis, 2021), and with tasks requiring the simultaneous processing of multiple items composed of arbitrary combinations of features (Campbell et al., 2024). We will return to the latter where we consider the relevance of our work to artificial systems in greater detail further below.

Relevance to Automaticity and Control

In this article we have largely adhered to Miller's (1956) focus on capacity constraints in tasks that involve perception and/or memory, ones that we suggested above may extend to language processing. Another domain of cognitive function in which similarly restrictive constraints have been observed is in multitasking capability and the allocation of cognitive control. Here, the constraints are usually framed in terms of a distinction, foundational in cognitive psychology, between automatic and control-dependent modes of processing (Posner & Snyder, 1975; Shiffrin & Schneider, 1977). *Control-dependent* processing is characterized as highly flexible and general, but subject to interference and therefore limited to inefficient serial processing. *Automatic* processing, in contrast, is less prone to interference and can support efficient, parallel processing, but requires extensive practice to achieve and is more rigidly task-dedicated. The same distinction — sometimes referred to as "dual process" theory" — has been made in other fields using other other terms, such as system 1 and system 2 in behavioral economics (Sloman, 1996), model-free (habitual) and model-based (deliberative) in machine learning and neuroscience (Daw et al., 2005; Doya, 1999; Sutton & Barto, 1998); and in the lay terms of thinking fast and slow (Kahneman, 2011).

The constraints on control-dependent processing have traditionally been attributed to reliance on a centralized, general purpose mechanism that is responsible for the flexibility of human cognitive function, including its ability to respond effectively to novel situations (Norman & Shallice, 1986), but that is capacity-limited and restricted to the serial execution of processing, akin to the central processing unit (CPU) of a traditional computer (Pashler, 1988; Posner & Snyder, 1975; Shiffrin & Schneider, 1977). However, a longstanding alternative — "multiple resources theory" (Allport et al., 1982; Navon & Gopher, 1979) — suggests that processing is distributed, and that constraints arise not from reliance on a "central bottleneck" imposed by a capacity-limited control mechanism, but rather when different processes must compete to use the same set of shared *local* resources. This has received growing support from computational analyses (e.g., Meyer & Keiras, 2001; Feng et al., 2014; Musslick et al., 2023), in which "local resources" are sets of representations required to perform a given task; and interference arises when different tasks (e.g., naming the color versus shape of a stimulus), that rely on a shared set of representations (e.g., phonological codes), require different ones to be activated at the same time. This has lead to a formulation of the distinction between control-dependent versus automatic processing in terms of a tradeoff between, respectively: i) the shared use of compositional representations, that can be acquired efficiently and flexibly reconfigured to perform novel tasks, but are subject to interference that imposes the need for serial processing; and ii) the formation of separated, conjunctive representations that are less subject to

interference and therefore permit parallel execution (i.e., multitasking), but require time and effort to acquire and are more rigidly task-specific (Musslick et al., 2023).

This interpretation aligns directly with the informational theoretic approach we have taken here, in terms of the tradeoff between representational efficiency and processing efficiency: Control-dependent, serial processing can be viewed as an adaptation by the cognitive system in which *processing* efficiency is sacrificed in exchange for the *representational* efficiency of structured codes, that afford flexible generalization. From this perspective, we can view the constraints associated with control-dependent processing as the *purpose* (rather than a *shortcoming*) of control: the imposition of serial processing to avert the risk of interference (i.e., decoding errors) incurred by the correlation intrinsic to structured codes. This framing offers a unified account of capacity constraints, that spans perception, working memory, and tasks that rely on cognitive control. It also provides a formally rigorous, and potentially normative framework within which to consider how and when such constraints may be overcome, that we turn to next.

Overcoming Capacity Constraints

Conjunctive Codes: Recoding, Consolidation and Automatization

In this article we have followed a long tradition of information theoretic work in cognitive and neuroscience by focusing on the importance of representational efficiency and expanding it, here, to take account of the value that similarity structure has for learning and generalization at the expense of processing capacity. While the latter are brought into sharp relief by tasks that involve novel or unpredictable stimuli — including those on which Miller focused — at the same time there are many conditions under which people exhibit remarkable parallel processing capabilities, as discussed above, both in perception (e.g., identifying a face in a crowd) as well as control and action (e.g., talking and driving). We assume that this relies on the use of conjunctive codes that have been acquired over the course of evolution, early in development, or through repeated experience, and that are not subject to the interference of structured codes. When the agent has access to such codes, identifying and using them can be an effective strategy for increasing processing capacity, a process that Miller referred to as “chunking” (e.g., representing a string of digits as a single date, or a string of letters as a single word). Such recoding is a well known mnemonic device that can produce dramatic increases in capacity, such as remembering an arbitrary string of 100 digits (Ericsson & Simon, 1980).

Importantly, with time and effort, people can also acquire new representations of this sort, using longer term learning mechanisms to generate purpose-specific conjunctive codes, by

forming enduring associations among existing codes that co-occur with sufficient frequency. This is what drives the process of automatization that occurs with practice in skill acquisition (Cohen et al., 1990; Logan, 1982; Musslick et al., 2023; Shiffrin & Schneider, 1997). The framework we have described provides a formally rigorous, normative approach to understanding this canonical trajectory, from capacity-limited, control-dependent performance early in training, that exploits the flexibility of structured codes, to more efficient, automatic, parallel processing and multitasking capability that comes from the formation of conjunctive codes with time and practice. It also bears striking parallels to observations and theory regarding representational learning, which is often framed in terms of statistical learning, where the same principles may apply.

For example, learning patterns of correlation among features has been used to explain the semantic structure of representations acquired by neural networks (Hinton, 1996; 2013; McClelland & Rogers, 2003; Rumelhart & Todd, 1993; Saxe et al., 2019), and the function of *consolidation* in the context of Complementary Learning Systems theory (Kumaran et al., 2013; McClelland et al., 1995). Consolidation refers to the use of associations rapidly formed in episodic memory (among previously *unassociated* representations in semantic memory) to forge new associations among those representations in semantic memory, while preserving existing structure as much as possible. Insofar as the initial, rapidly formed associations in episodic memory are between structured forms of representations in semantic memory, this can be considered homologous to the function of rapid binding among structured codes we have considered in this article (though see note 10), and subject to the same limitations. For example, while the storage capacity of episodic memory is generally treated as unlimited, the number of independent sets of associations (e.g., among features of an item) that can be either stored or retrieved at a given time are constrained by the same factors of compositionality (the binding problem) and semanticity (proactive interference) as those we have considered in this article. Similarly, the transition from the flexible but arbitrary associations formed in episodic memory to the richer, domain-specific semantic codes generated over time through replay and/or rehearsal, can be considered homologous to the transition from control-dependent to automatic forms of processing that come with practice in the domain of skill acquisition.

In summary, the information theoretic framework we have described may provide a unified, formally rigorous approach to understanding the effects of recoding (“chunking”), as well as the trajectories in representational learning and skill acquisition, in terms of the formation of purpose-specific conjunctive codes that reduce correlations with (i.e., increase distance from) other codes with which they may interfere. The frequency-dependent nature of the learning mechanisms involved accord with the general principle of optimal coding in information theory,

in which likelihood should impact the structure of the code. This has recently been applied to provide a normative account of dual process theory in terms of the minimum description length principle (Moskovitz et al., 2024). Here, we propose that this can be extended beyond the traditional assertion that frequency should impact code *length*, to the idea that frequency should impact code *distance,* by modifying or forming new codes for frequently encountered forms of correlation (e.g., items) that are further from those with which they can be confused. At the same time, there may be pressure to also consider their value for compositional coding: that is, the extent to which these codes can themselves be combined to represent novel items. In general, an important factor in consolidation and automatization may be the formation of dedicated representations for high frequency items, combinations of which are most useful for representing a broad class of low frequency items. This framing may provide a useful foundation for a more detailed, and normative understanding of how and when learning can be used to optimize the tradeoff between representational and processing efficiency.

Optimization of the Tradeoff between Representational and Processing Efficiency

Recognizing the normative value of frequency-dependent structuring of the code highlights the importance of the processes of consolidation and automatization discussed above: These processes allow people to actively manipulate the frequency with which they experience and process different information. That is, replay, rehearsal and practice all selectively increase exposure to some correlations over others, relative to their incidental likelihoods in the environment. This active manipulation of frequency suggests that optimization of the tradeoff between representational and processing efficiency involves at least three additional, closely related factors: the utility of restructuring the code, the time it takes to do so, and the temporal horizon over which that utility obtains. Characterizing the specific forms of utility that drive these processes is beyond the scope of this article. However, the general problem can be cast in the form of an intertemporal choice with respect to optimizing the rate of utility maximization. When confronted with a novel or unfamiliar task, should an agent: a) rely on the flexibility of currently available, representationally efficient codes, that can support immediate performance and thus maximize utility rate over the *short* term, relative to investing the time and effort required to restructure the code, but that incurs the opportunity cost of improving processing efficiency over the *longer* term; or b) invest in restructuring the code to improve processing efficiency, which will increase the rate of utility maximization over the *longer* term, but will take time and effort and thus incur opportunity costs over the *shorter* term. Recent work has begun to address this question in the context of skill acquisition and multitasking capability (e.g., Sagiv et al., 2018; Ravi et al., 2020; Petri et al., 2024). The current framework could be used to generalize this approach to perceptual and inference processes, which may have relevance not only for

understanding the profiles of human performance, but for the design of artificial systems, that we discuss next.

Relevance to Work in Machine Learning and Artificial Intelligence

Progress in machine learning has produced remarkably successful systems that exhibit flexibility of processing and generalization comparable to that of humans in tasks such as inference and analogical reasoning (e.g., Webb et al., 2023), suggesting that they have acquired structured codes required to support such generalization capabilities. According to our account, this predicts that, despite their massive resources (processing units and parameters) and the data to which they have access, these models should nevertheless be subject to constraints in processing efficiency similar to those observed for humans when called upon to process novel information.

Consistent with this idea, it has been observed that the distributional properties of the data on which transformers are trained is important in determining what information is stored in the structure of the weights, and what information relies on dynamic construction of activity states, often referred to as “in context” learning (Chan et al., 2022). Intriguingly, it was found that weight-based learning occurred for high frequency items involving stable correlations in the data, whereas large numbers of rarely occurring classes and/or items the meanings of which were dynamic rather than fixed across samples, tended to rely on “in context” learning. These observations align well with the idea that stable, frequently encountered forms of correlation will, with sufficient exposure, be assigned dedicated conjunctive codes; that is, they will become incorporated into the system’s set of existing semantic codes, with weights dedicated to the relevant correlations, comparable to the processes of consolidation and automatization discussed above. In contrast, items encountered less frequently (i.e., involving more “novel” correlations) will rely on dynamically constructed compositional representations formed from existing semantic codes; that is, the correlations will be represented using rapid binding mechanisms such as those implemented in the models described in this article. These observations also reinforce the prediction made by our analysis, that the amount of information transformers can process using “in context” learning should be limited in ways similar to humans when they are forced to process information involving novel or unpredictable correlations (as in the tasks discussed in this article).

As noted earlier, this may already be evident in state of the art machine learning models. For example, one recent study investigated the extent to which existing large visual language models (VLMs; such as GPT-4v) can reliably distinguish among simultaneously presented

items, using the types of tasks and stimuli that have been used to characterize capacity constraints in humans, including those we have considered in this article. That study found that VLMs consistently exhibited constraints in processing capacity as restrictive as those observed for humans forced to rely on rapid parallel processing (Campbell et al., 2024). This is consistent with the supposition that VLMs rely on the learning of structured representations for generalization, and the generality of the principle that such representations should be associated with restrictions in processing capacity.

Finally, in a separate line of work in machine learning, there have been extensive studies seeking ways to improve the generalization performance of neural networks. These have shown that training networks on several related tasks in ways that encourage the discovery of shared structure — referred to variously as “multitask training” (Baxter, 1995; Bengio at el. 2013; Caruana, 1997) or meta-learning (Finn et al., 2017; Hospedales et al., 2021; Santoro et al., 2016) — can substantially improve performance on novel tasks that share similar structure. Critically, however, networks in these paradigms are only ever trained or asked to perform *one task at a time.* Recently, the effects of such learning have been interpreted in terms of the acquisition of compositional representations, where it is shown that this leads to dramatic restrctions in *multitasking* capability — that is, the ability perform multiple tasks *at the same time* (Musslick et al., 2023) — an effect that is largely invariant to the size of the network (Petri et al., 2023). Again, this is consistent with the principle that, whereas structured representations support more efficient learning and generalization, this comes at the expense of processing efficiency (in this case, the ability to simultaneously perform multiple tasks). As noted above, recent work has begun to consider ways in which the tradeoff between generalization and processing capacity can be optimized, using both formal analysis of abstracted networks and simple task settings (Sagiv et al., 2018; Petri et al., in press), and using deep learning in the context of more complex and realistic tasks (Ravi et al., 2020).

These findings may have significance for the design of artificial systems. Most work to date, like traditional work in cognitive and neuroscience, has focused on representational efficiency: how artificial systems can achieve the efficiency of learning and the level of flexibility and generalization exhibited by humans. It is generally assumed that this will require imbuing models with more powerful inductive biases for discovering structure, whether in their training curricula (e.g., Marinescu et al., 2024) and/or their architecture (Webb et al., 2024), in order to optimize the bias-variance tradeoff (see note 7). The work we have presented suggests that one way to frame this is the need to form dedicated representations of high frequency items that can be combined (i.e., compositionally) to represent as many low frequency items as possible. However, while optimizing this tradeoff remains a major challenge for the design of artificial

systems, the work we have presented suggests that, even with complete success in this effort, the design problem faces another tradeoff: to the extent that inductive biases are found that maximize representational efficiency, this will carry a commensurate cost in *processing* efficiency. That is, it will be subject to the *curse of generalization*.

This tradeoff between flexible but inefficient processing and efficient but task-dedicated processing is comparable to a similar tradeoff faced in traditional computational architectures, between highly flexible components (interpreted procedures and general purpose processors) and purpose-dedicated ones (compiled procedures, drivers and custom chipsets). It is intriguing to consider the possibility that this reflects yet another expression of the information theoretic principles we considered in this article. Whether or not this is so, the tension is a fundamental one for any adaptive system that must function in the real world. While flexibility and generalization are clearly assets in naturalistically non-stationary and unpredictable environments, so is the need to simultaneously encode and meaningfully distinguish among multiple sources of input, and/or manage the simultaneous execution of multiple independent tasks — capabilities that are likely to be critical for artificial autonomous agents functioning in naturalistic environments. It is clear that the human brain has developed mechanisms to adjudicate this tradeoff and, as we discussed above, to adapt both its behavior and its structure over time to do so as a function of experience and need. We hope that our consideration of the information processing principles underlying this tradeoff will help advance the understanding of how this tradeoff is managed by the brain, as well as the design of artificial agents that are as effective in doing so.

Conclusion

In this article, we provided a formal characterization of a fundament tradeoff faced by any information processing system, between: representational efficiency — the efficiency with which an agent can acquire and generalize information about the world, advantaged by the use of structured (similarity-based) codes; and processing efficiency — the amount of information that can be processed at the same time, advantaged by conjunctive (maximally separated) codes. We presented an analysis of this tradeoff in an abstract, information theoretic form, and then demonstrated that the same principles are exhibited in a neurally-inspired mechanistic model of tasks that have been used to characterize human information processing capacity constraints in perception and memory. Finally, we argued that the same principles can be applied to other domains of cognitive function that exhibit similarly restrictive constraints on processing capacity, such as language, multitasking performance and control-dependent processing.

Our account follows a long tradition of work that assumes natural agents place a premium on representational efficiency, that is the pace at which generalizable codes can be acquired and the flexibility with which they can be used. While our analyses demonstrated the value of structured codes in an abstract formulation of the problem, and in simulations of relatively simple tasks that have been used in the laboratory, these likely underestimate their value in more realistic environments, given the statistical complexity of the natural world, and the value of responding flexibly and effectively within it. Nevertheless, processing efficiency can also be of considerable value, and it is clear that humans actively invest the effort to achieve this when the demand is sufficiently frequent and valuable, through processes such as consolidation and automatization. The information theoretic approach we have presented offers a unifying, formally rigorous framework for understanding these processes, as the formation of purpose-dedicated, conjunctive codes that reduces the correlation with (i.e., increases the distance from) other codes with which they may interfere.

Our analyses, inspired by George Miller's landmark observations, have lead us to a very different conclusion than the one he felt compelled to make: the consistency of capacity-limits in the range of the "magic number 7" is not a coincidence. Instead, we suggest that it is a fundamental property of a system optimally configured to represent as broad a range of states as possible, weighed against the value of responding to these as efficiently and accurately as possible. That is, it reflects a ubiquitous and inescapable tradeoff, that defines the envelope of achievable performance for any finite information processing agent — a constraint that we suggest should appropriately, if ironically, be thought of as Miller's Law.

Supplementary Information

Abstract Model

We simulated the *similarity* and *identification* tasks as follows. For each trial n items were sampled without replacement from the integers $\{1, 2 \ldots 256\}$ to constitute a test set S. For the identification task, one of these items was sampled as the *probe.* For the similarity test, a probe was sampled from the uniform distribution $U[1, 256]$. The probability that any item i would be retrieved as the response to probe p was proportional to an exponential *generalization gradient* parameterized with μ as in Shepard (1987):

$$g(i,p) = \exp\left(-2\frac{|\,i-p\,|}{\mu}\right)$$

This function returns 1 when i and p are equal and approaches 0 as i and p are separated, at a rate determined by μ. We refer to $g(i,p)$ as a *generalization score.* We introduced noise by sampling generalization scores $G_{i,p}$ from a beta distribution parameterized with a mode $\omega = g(i,p)$ and a concentration κ determining the amount of noise:

$$G_{i,p} \sim \text{Beta}(1+\omega\kappa, 1+(1-\omega)\kappa)$$

where the beta distribution is shown here mapping ω and κ into the standard $\text{Beta}(\alpha,\beta)$ parameterization.[14] The value of κ used for the plots shown in Figure 5 of the main text was 10^{15}. The noisy generalization scores for each of the n items in the set were normalized to produce a probability distribution over the n items, so the probability of retrieving item i from S given probe p is:

---

[14] *An unintentional "bug" that reaffirmed the principle.* Noise is critical to producing the tradeoff between generalization and identification error. Without noise, generalization error — like identification error — is minimized by minimizing μ, and there is no tradeoff between the two forms of error. However, the amount of noise required to produce the tradeoff is extremely small, and even introducing minimal amounts of noise quickly and dramatically induces the tradeoff. This can be shown in analyses that will be included in a future update of this document; these indicate that no practically realizable reduction in noise is sufficient to meaningfully diminish the severity of the tradeoff. That is, even modest increases in μ, that fall considerably short of optimizing generalization error, induce sufficient identification error to impose the radical constraints in processing efficiency observed empirically. Interestingly, simulations described in a previous version of this report demonstrated these effects, but did so without including an explicit noise term. Rather, the observed effects arose (inadvertently but instructively) due to the limits in the floating point precision of the simulations. Those made it impossible to distinguish between distances below some critical value, thus imposing effects comparable to noise, and thereby induced the tradeoff. This was confirmed by directly manipulating the floating point precision, which reduced the effects as precision was increased. While this effect of floating point precision was not intended or anticipated, it can be taken as a convincing affirmation of the generality of the principle: The same qualitative effects were observed when the simulations were run with the maximum precision on the most physically precise machine to which we had access. That is, no physically realizable level of precision could avert the tradeoff between generalization and identification error, and thus achieve any reasonable amount of generalization error without drastically impacting identification error and thus constraining processing efficiency.

$$P(i) = \frac{G_{i,p}}{\sum_{j \in S} G_{j,p}}$$

An $n \times n$ table of weighted counts was maintained tracking the co-occurrence of correct response indices and model response indices; for each simulated trial, the table was updated using the vector of probabilities *P*. After 1000 trials the table was used to estimate the joint probability distribution from which the mutual information performance scores were computed at each value of μ tested. 1000 trials at each level of μ was sufficient to produce the smooth curves shown in Figure 5.

Gaussian Similarity Function

The results shown in Figure 5 of the main text used the exponential decay function above to implement similarity structure in the code. This was inspired by the derivation of this in Shepard (1987). However, given the "peakedness" of this function, it may provide a conservative estimate of the impact that semantic structure can have on processing capacity, relative to ones that may be relevant to various forms of coding. This is suggested by Figure S1, which shows analyses using the same procedures described in the text, but here using a Gaussian rather than an exponential distribution to implement similarity among codes along each direction.

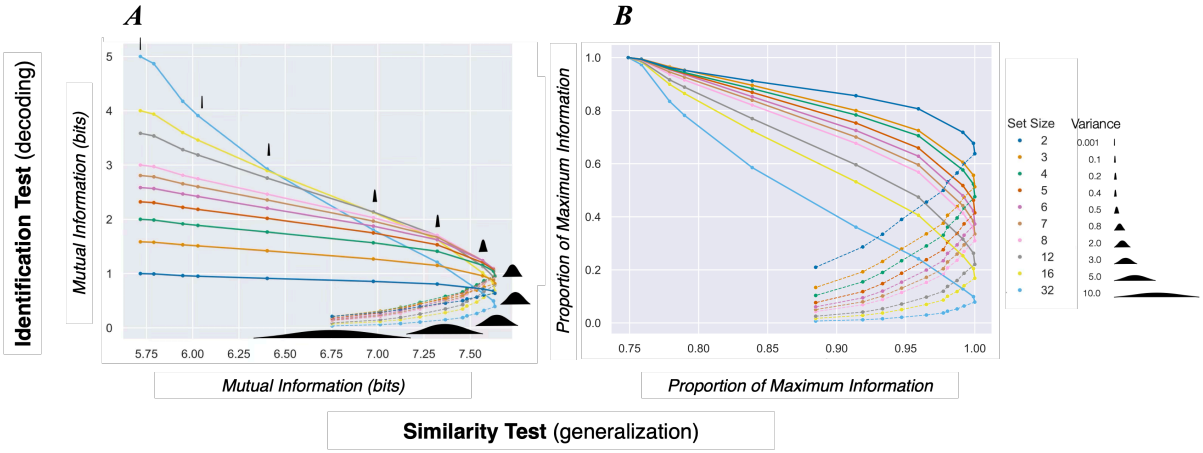

Figure S1. Tradeoff between representational capacity and processing capacity. Shows results of analysis using the same procedure described for Figure 5 in the main text. but using a Gaussian rather than an exponential distribution to implement similarity structure among codes along each dimension. Note that this has a more restricting influence on processing capacity than the exponential function used for Figure 5.

Mechanistic Models

Hopfield-Ising Network

We adopt the statistical framework of computation (i.e., encoding and decoding in the channel above) as minimization of an *energy function* (Hopfield, 1982; Hinton & Sejnjowski, 1983), influential in computational neuroscience (Hopfield, 1982; Friston et al. 2010; Gottlaub & Braun, 2021) as well as machine learning (Zemel et al. 1995; Hinton & Salaktudinov, 2007; Salakhutdinov, 2018; LeCun et al. 2006). The "energy" state of a network is a global measure of the agreement between the knowledge stored in the network's parameters (weights) and its current activity values. Here, we consider its utility as an abstract cognitive model, capable of explaining human experimental data in classic cognitive tasks.

As in other frameworks, our model assumes two time-scales of parameters: slow and fast (Hinton & Plaut, 1987; McClelland, O'Reilly, & McNaughton, 1995; Ba et al. 2018; Whittington et al. 2020). Given a stimulus (x*), a binary code Z|x* is inferred using an encoder *f*, which we imagine to be acquired over a slow ontogenetic or phylogenetic time-period. In the Hopfield network simulations, we specifically assume the outputs of the encoderZ(|x*) are efficient codes of the task-relevant stimulus variables (e.g., spatial location, color, or tone). To specify these codes we follow the coding scheme of Bicanski & Burgess (2019) adapted for both 1D and 2D domains: each state is encoded using frequencies at different phases, scaled at $\sqrt{e}$. These continuous grid-codes and binarized to allow the Hopfield network to operate over variables that can take values of (+1,-1).

We assume the agent is unable to store encodable/decodable entries in Z for all possible task-relevant states (e.g., colors at particular locations). Instead, they address this shortcoming by augmenting *f* with a fast-time scale encoder *f** that reflects the minimal statistics of Z|x*. For example, *what was where* in a particular display. We treat the outputs of the fast-time scale encoder as cases of rapid "variable-binding": for example, that a particular color was at a particular location on that trial. Although *how* neural circuits rapidly bind (Von der Malsburg, 1994; Singer & Gray, 1995; Doumas et al. 2008; Hayworth, 2012; Kriete, 2013; Maas et al. 2019) remains an outstanding question, here, we focus simply on the co-variance statistics induced by binding. The co-variance statistics of this "binding" are captured by the Hebbian weight matrix.

Eq S1. $W = Z \mid x^* \otimes Z \mid x^*$

To complete each task, partial stimulus information is available (Z'| x*), and the missing information must be inferred. We implement this process through standard asynchronous settling in a Hopfield network (Hopfield, 1982; Amit, 1988). Specifically, given the population of binary variables that have discrete-time (*t)* varying values (*a*) , computed as:

Eq S2. $a_i = \sum_{i \neq j} w_{ij} x_j + \theta$

update these individual variables according to

Eq S3. $a_i(t+1) = \begin{cases} 1, & if\ a_i(t) > 0. \\ -1, & otherwise \end{cases}$

That is, "asynchronously", until the energy function *(E)*

Eq S4. $E[S] = -\frac{1}{2}\sum_{i \neq j} w_{ij} s_i s_j + \theta.$

reaches a local minimum. In the absence of additional noise, dynamics will drive the network toward local energy minima. We further follow Mackay (1991) and add a sustained bias term to each node, proportional to the agent's assumed uncertainty about a task-relevant feature. For those features that are *observable* (e.g., the stimulus location at time of test) , a sustained bias on those coding elements devoted to that feature are included to prevent the network from "drifting away". This bias is defined for each coding element Z_i

Eq S5. $bias_i = Z_i(\frac{1}{2}\log(\frac{1-b_i}{b_i}))$

Where $b_i = 0$ if the feature Z_i partially codes for is currently observable (the "cue") and $b_i = 0.5$ if the feature Z_i partially codes for is is not observable, and must be inferred. If we treat the weights as an empirical prior, this is much like a Hierarchical Bayesian model, but without the explicit normalization in the denominator.

Finally, we note that our model follows in the tradition of Hopfield networks as non-ferromagnetic "spin-glass" models and assumes that the weight (interaction) matrix is *(a)* fully connected (rather than local as in the Lenz-Ising Model) and *(b)* symmetric (W_{ij}=W_{ji}). Symmetry is a sufficient condition for guaranteeing that the algorithm settles at local minima in the energy landscape (Hopfield, 1982; Amit, 1989). Moreover, for any fixed *N*, it's preferable to equip the system with all-to-all connections. All-to-all connections maximize the number of *possible* dependencies, while allowing the experience-dependent statistics to determine which in-context dependencies exist in *W*.

Tensor Product Variant

To highlight the representational generality of the principles, we consider a simple case in which the bindings are represented in a population of *activities* (Smolensky, 1990), rather than the network weights (Hopfield, 1982). Imagine an efficient *m* bit code for a task-relevant feature, (e.g., for color) (Z_f), and an *n*

bit code (for e.g. spatial positions) (Z_r) as in the Hopfield network formulation, but with a separate population of $(m+n)^2$ nodes that represent the bindings of each Z_r and Z_f as a tensor product ⓧ, such that each combination contains a unique node in TP, indexed by *i.* For each task context *, we compute the tensor product of the codes (Z_r, Z_f r for task-relevant features.

To perform the experimental participants' task, a control system "actively maintains" the TP representation over the temporal context window *, much like classical models of prefrontal cortical function (Braver et al., 1999; O'Reilly & Frank, 2001). Although we could imagine this control function optimized through reinforcement learning, here we simply stipulate by fiat that activities are maintained throughout the temporal context and are updated with each contextual change. Given partial information Z_f,, the network's task is to use the actively-maintained bindings in the TP layer to complete the missing information in Z_r. That is, as in the Hopfield model, we seek to *induce* the most-likely activity state in Z_r, given the knowledge in Z_f and the tensor product of Z_r and Z_f, maintained in a separate population of activities.

A population of Hebb-like gates (*g)* control the influence of TP* on Z. A gate between any of the $(m+n)^2$ TP nodes (*i)* and any of the m+n Z_k exists if TP_i is influenced by Z_k.. During the inference process, each g_k is closed by default. To align with the asynchronous updating procedure of the Hopfield network, we randomly sample nodes in Z without replacement throughout the settling process. The gates between the sampled nodes in TP that have co-varied with Z_k in * open, allowing those conjunctions induced by Z_k to project to Z at time *t*. We compare that projection against the information available in the content layer by inner product, as in the Hopfield rule. The node Z_k is updated as

Eq S6. $sign(g(\ Z_k(Z_r \otimes Z_f) + b))$

Eq S7 $= \begin{cases} 1 & \text{if } Z_k \times TP_k > 0 \\ 0 & \text{otherwise} \end{cases}$

The process is repeated over all content nodes until it finds a local energy minimum. Results for a simple activity-based (Tensor-Product) and weight-based (Hopfield) model with 24 nodes and similarity-preserving codes are presented in Figure S2. Results are qualitatively the same across implementations.

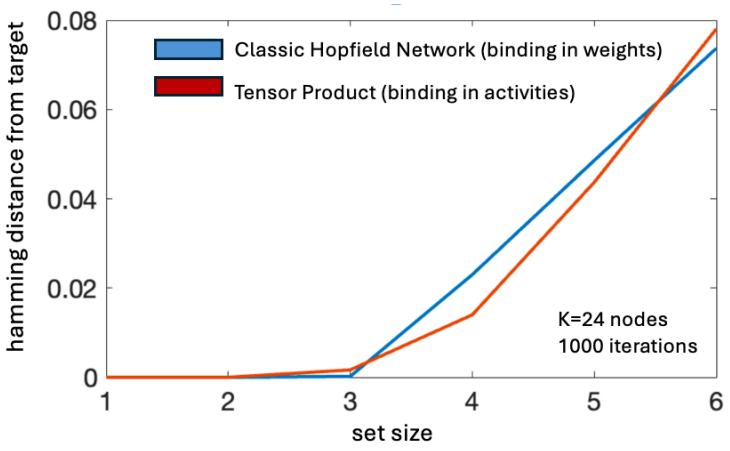

Figure S2. Comparison of average retrieval precision as a function of set size for neural networks that carry information about context-specific bindings in (a) their weights (classic Hopfield network, blue), versus (b) in the activities (tensor product, red) in a separate pool of conjunctive nodes. Models show qualitatively similar results.

References

Ackley, D. H., Hinton, G. E., & Sejnowski, T. J. (1985). A learning algorithm for Boltzmann machines. *Cognitive science*, *9*(1), 147-169.

Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological science*, *15*(2), 106-111.

Amit, D. J., Gutfreund, H., & Sompolinsky, H. (1985). Storing infinite numbers of patterns in a spin-glass model of neural networks. *Physical Review Letters*, *55*(14), 1530.

Amit, D. J. (1989). *Modeling brain function: The world of attractor neural networks*. Cambridge university press.

Anderson, J. R. (1983). *The Architecture of Cognition*. Psychology Press.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological review*, *61*(3), 183.

Baddeley, A. (1992). Working memory. *Science*, *255*(5044), 556-559.

Barak O, Rigotti M, Fusi S (2013). The sparseness of mixed selectivity neurons controls the generalization-discrimination trade-off. *Journal of Neuroscience, 33*(9):3844-56

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. *Sensory communication*, *1*(01).

Baxter, J. (1995). Learning internal representations. In *Proceedings of the Eighth Annual Conference on Computational Learning Theory,* 311–320.

Bays, P. M., & Husain, M. (2008). Dynamic shifts of limited working memory resources in human vision. *Science*, *321*(5890), 851-854.

Bell, A. J., & Sejnowski, T. J. (1995). An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, *7*(6), 1129-1159.

Bellman, R. (1957). *Dynamic Programming.* Princeton University Press, Princeton, NJ.

Bengio, Y., Courville, A., & Vincent, P. (2013). *Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence, 35*(8), 1798–1828.

Beukers, A. O., Buschman, T. J., Cohen, J. D., & Norman, K. A. (2021). Is activity silent working memory simply episodic memory?. Trends in cognitive sciences, 25(4), 284-293.

Bicanski, A., & Burgess, N. (2019). A computational model of visual recognition memory via grid cells. *Current Biology*, *29*(6), 979-990.

Braida, L. D., & Durlach, N. I. (1972). Intensity Perception. II. Resolution in One-Interval Paradigms. *The Journal of the Acoustical Society of America*, *51*(2B), 483-502.

Campbell, D.I., Rane, S., Giallanza, T., De Sabbata, N., Ghods, K., Joshi, A., Ku, A., Frankland, S.M., Griffiths, T. L., Cohen, J. D. & Webb, T. W. (2024). Understanding the limits of vision language models through the lens of the binding problem. *NeurIPS.* https://arxiv.org/abs/2411.00238.

Caruana, R. (1997). Multitask learning. Machine Learning, 28(1), 41–75.

Chan, S., Santoro, A., Lampinen, A., Wang, J., Singh, A., Richemond, P., McClelland, J. L. & Hill, F. (2022). Data distributional properties drive emergent in-context learning in transformers. *Advances in Neural Information Processing Systems, 35*, 18878-18891.

Chandra, S., Sharma, S., Chaudhuri, R., & Fiete, I. (2023). High-capacity flexible hippocampal associative and episodic memory enabled by prestructured"spatial"representations. bioRxiv, 2023-11.

Cheyette, S. J., & Piantadosi, S. T. (2020). A unified account of numerosity perception. *Nature human behaviour*, *4*(12), 1265-1272.

Chomsky, N., & Miller, G. A. (1958). *Finite state languages. Information and control, 1*(2), 91-112.

Cohen JD, Dunbar K & McClelland JL (1990). On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review, 97*(3), 332-361.

Cover, T. M., & Thomas, J.A. (1991). *Elements of information theory*. John Wiley & Sons.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, *24*(1), 87-114.

Daw, N. D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature neuroscience, 8*(12), 1704-1711.

Doya K. (1999). What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Networks*. 12, 961 – 974.

Dordek, Y., Soudry, D., Meir, R., & Derdikman, D. (2016). Extracting grid cell characteristics from place cell inputs using non-negative principal component analysis. *Elife*, *5*, e10094.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215.

Fiete, I. R., Burak, Y., & Brookings, T. (2008). What grid cells convey about rat location. *Journal of Neuroscience*, *28*(27), 6858-6871.

Finn, C., Abbeel, P., & Levine, S. (2017). Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning (PMLR)*, 1126-1135.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition*, *28*(1-2), 3-71.

Frankland, S. M., Webb, T. W., Petrov, A. A., O'Reilly, R. C., & Cohen, J. D. (2019). Extracting and utilizing abstract, structured representations for analogy. In Proceedings of the Annual Meeting of the Cognitive Science Society (Vol. 41).

Frankland, S. M., & Greene, J. D. (2020). Concepts and compositionality: in search of the brain's language of thought. *Annual review of psychology*, *71*, 273-303.

Frege, G. (1892). Über begriff und gegenstand. *Vierteljahrsschrift für wissenschaftliche Philosophie*, *16*(2).

Friston, K. (2010). The free-energy principle: a unified brain theory?. *Nature reviews neuroscience*, *11*(2), 127-138.

Gallistel, C. R., & Gelman, R. (1992). Preverbal and verbal counting and computation. *Cognition*, *44*(1-2), 43-74.

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

Giallanza, T., Campbell, D., & Cohen, J. D. (2024). Toward the emergence of intelligent control: Episodic generalization and optimization. Open Mind, 8, 688-722.

Gibson E (1998). Linguistic complexity: Locality of syntactic dependencies. *Cognition, 68*(1):1–76.

Gibson, E (2000). The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, Language, Brain: Papers from the First Mind Articulation Project Symposium*, 95–126.

Graves, A., Wayne, G., & Danihelka, I. (2014). Neural turing machines. arXiv preprint arXiv:1410.5401.

Gravetter, F., & Lockhead, G. R. (1973). Criterial range as a frame of reference for stimulus judgment. *Psychological Review*, *80*(3), 203.

Hafting, T., Fyhn, M., Molden, S., Moser, M. B., & Moser, E. I. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*, *436*(7052), 801-806.

Hasselmo, M. E., & Wyble, B. P. (1997). Free recall and recognition in a network model of the hippocampus: simulating effects of scopolamine on human memory function. Behavioural brain research, 89(1-2), 1-34.

Hayes, J. R. M. (1952). Memory span for several vocabularies as a function of vocabulary size. In *Quarterly Progress Report,* Cambridge, Mass.: Acoustics Laboratory, Massachusetts Institute of Technology.

Hayworth, K. J. (2012). Dynamically partitionable autoassociative networks as a solution to the neural binding problem. *Frontiers in computational neuroscience*, *6*, 73.

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society, 1,* 12.

Hinton, G. E., & Plaut, D. C. (1987, July). Using fast weights to deblur old memories. In *Proceedings of the 9th annual conference of the cognitive science society,* 177-186).

Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in cognitive sciences*, 8(11), 494-500.

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the national academy of sciences*, *79*(8), 2554-2558.

Hospedales, T., Antoniou, A., Micaelli, P., & Storkey, A. (2021). Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence, 44*(9), 5149-5169.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. Journal of mathematical psychology, 46(3), 269-299.

Jensen, O., & Lisman, J. E. (1998). An oscillatory short-term memory buffer model can account for data on the Sternberg task. *Journal of Neuroscience*, 18(24), 10688-10699.

Joshi, D.D (1958). A note on upper bounds for minimum distance codes", *Information and Control, 1* (3): 289–295.

Kahneman, D. (2011). *Thinking, fast and slow.* Farrar, Straus and Giroux.

Kaufman, E. L., Lord, M. W., Reese, T. W., & Volkmann, J. (1949). The discrimination of visual number. *The American journal of psychology*, *62*(4), 498-525.

Kohavi, Ron; Wolpert, David H. (1996). "Bias Plus Variance Decomposition for Zero-One Loss Functions". ICML. 96.

Kriete, T., Noelle, D. C., Cohen, J. D., & O'Reilly, R. C. (2013). Indirection and symbol-like processing in the prefrontal cortex and basal ganglia. *Proceedings of the National Academy of Sciences*, *110*(41), 16390-16395.

Krotov, D., & Hopfield, J. J. (2016). Dense associative memory for pattern recognition. *Advances in neural information processing systems*, *29*.

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in Cognitive Sciences, 20*(7), 512-534.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M., & Huang, F. (2006). A tutorial on energy-based learning. *Predicting structured data*, *1*(0).

Lewis, R. L. (1996). Interference in short-term memory: The magical number two (or three) in sentence processing. *Journal of Psycholinguistic Research, 25*(1), 93-115.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, *10*(10), 447-454.

Lind, J., Enquist, M., & Ghirlanda, S. (2015). Animal memory: A review of delayed matching-to-sample data. Behavioural processes, 117, 52-58.

Linsker, R. (1986). From basic network principles to neural architecture: Emergence of spatial-opponent cells. *Proceedings of the National Academy of Sciences*, *83*(19), 7508-7512.

Linsker, R. (1988). Self-organization in a perceptual network. *Computer*, *21*(3), 105-117.

Lisman, J. E., & Idiart, M. A. (1995). Storage of 7+/-2 short-term memories in oscillatory subcycles. *Science*, *267*(5203), 1512-1515.

Logan, G. D. (1985). Skill and automaticity: Relations, implications, and future directions. *Canadian Journal of Psychology/Revue Canadienne de Psychologie, 39*(2), 367.

Luck, S. J., & Vogel, E. K. (1997). The capacity of visual working memory for features and conjunctions. *Nature*, *390*(6657), 279-281.

Luxburg, Ulrike V.; Schölkopf, B. (2011). "Statistical learning theory: Models, concepts, and results". Handbook of the History of Logic. 10: Section 2.4.

Ma, W. J., Husain, M., & Bays, P. M. (2014). Changing concepts of working memory. *Nature neuroscience*, *17*(3), 347-356.

MacKay, D. J. C. (1991). Maximum entropy connections: Neural networks. In *Maximum entropy and Bayesian methods* (pp. 237-244). Springer, Dordrecht.

MacWilliams, F. J & Sloane, N. J. A. (1977). *The theory of error-correcting codes.* North-Holland.

Mandler, G., & Shebo, B. J. (1982). Subitizing: an analysis of its component processes. *Journal of experimental psychology: general*, *111*(1), 1.

Marcus, G. F. (2003). *The algebraic mind: Integrating connectionism and cognitive science*. MIT press.

Marinescu, I., McCoy, R. T., & Griffiths, T. L. (2024). Distilling Symbolic Priors for Concept Learning into Neural Networks. arXiv preprint arXiv:2402.07035

Mathy, F., & Feldman, J. (2012). What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition*, *122*(3), 346-362.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, *102*(3), 419.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature reviews neuroscience, 4*(4), 310-322.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological review*, *88*(5), 375.

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

Miller, A. (1963). Finitary models of language users. *Handbook of Mathematical Psychology,* 2.

Miller, E. K., & Cohen, J. D. (2001). An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, *24*(1), 167-202.

Moskovitz, T., Miller, K. J., Sahani, M., & Botvinick, M. M. (2024). Understanding dual process cognition via the minimum description length principle. *PLOS Computational Biology, 20*(10), e1012383.

Mondal, S. S., Frankland, S.M., Webb, T. W., & Cohen, J. D. (2023). Determinantal Point Process Attention Over Grid Codes Supports Out of Distribution Generalization. *eLife, 12*.

Musslick, S., Saxe, A., Hoskin, A. N., Sagiv, Y., Reichman, D., Petri, G., & Cohen, J. D. (2023). On the rational boundedness of cognitive control: Shared versus separated representations.

Nassar, M. R., Helmers, J. C., & Frank, M. J. (2018). Chunking as a rational strategy for lossy data compression in visual working memory. Psychological review, 125(4), 486.

Navon, D., & Gopher, D. (1979). On the economy of the human-processing system. *Psychological Review, 86(*3), 214

Nizami, L. (2010). Interpretation of absolute judgments using information theory. *Cybernetics & Human Knowing, 17*(1-2), 111-155

Norman, K. A., & O'Reilly, R. C. (2003). Modeling hippocampal and neocortical contributions to recognition memory: a complementary-learning-systems approach. *Psychological Review, 110*(4), 611.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behaviour. In: R. J. Davidson., G. E. Schwartz, & D. E. Shapiro (Eds.), *Consciousness and Self-Regulation* (pp. 1-14). New York: Plenum Press.

Oberauer, K., Lewandowsky, S., Farrell, S., Jarrold, C., & Greaves, M. (2012). Modeling working memory: An interference model of complex span. Psychonomic bulletin & review, 19, 779-819.

Oja, E. (1982). Simplified neuron model as a principal component analyzer. *Journal of mathematical biology*, *15*(3), 267-273.

Partee, B. (1995). Lexical semantics and compositionality. *An invitation to cognitive science*, *1*, 311-360.

Petri, G., Musslick, S., Dey, B., Özcimder, K., Ahmed, N. K., Willke, T. L. & Cohen J. D. (2021). Topological limits to parallel processing capability of network architectures. *Nature Physics, 17*(5), 646-651.

Petri, G., Musslick, S., & Cohen, J. D. (in press). An information-theoretic approach to reward rate optimization in the tradeoff between controlled and automatic processing in neural network architectures. *eLife.*

Pollack, I. (1952). The information of elementary auditory displays. *The Journal of the Acoustical Society of America*, *24*(6), 745-749.

Pollack, I. (1953). The information of elementary auditory displays. II. *The Journal of the Acoustical Society of America*, *25*(4), 765-769.

Pollack, I. (1953). The assimilation of sequentially encoded information. *American Journal of Psychology, 66*, 421-435.

Posner, M. I., & Snyder, C. R. (1975). Attention and Cognitive Control 1. In Information processing and cognition (pp. 55-85). Routledge.

Ravi, S., Musslick, S., Hamin, M., Willke, T. L., & Cohen, J. D. (2020). Navigating the trade-off between multi-task learning and learning to multitask in deep neural networks. arXiv preprint arXiv:2007.10527.

Revkin, S. K., Piazza, M., Izard, V., Cohen, L., & Dehaene, S. (2008). Does subitizing reflect numerical estimation?. *Psychological Science*, *19*(6), 607-614

Roskies, A. L. (1999). The binding problem. *Neuron*, *24*(1), 7-9.

Rouder, J. N. (2001). Absolute identification with simple and complex stimuli. *Psychological Science*, *12*(4), 318-322.

Roux, F., Wibral, M., Mohr, H. M., Singer, W., & Uhlhaas, P. J. (2012). Gamma-band activity in human prefrontal cortex codes for the number of relevant items maintained in working memory. *Journal of Neuroscience*, 32(36), 12411-12420.

Rumelhart, D. E. & Todd, P. M. (1993). Learning and connectionist representations. In Eds Meyer, D. E. & Kornblum, S. (Eds.): *Attention and Performance XIV: Synergies in Experimental Psychology, Artifical Intelligence, and Cognitive Neuroscience.* MIT Press, Cambridge, Massachusetts. 3–30.

Ryu, S. H., & Lewis, R. L. (2021). Accounting for agreement phenomena in sentence comprehension with transformer language models: Effects of similarity-based interference on surprisal and attention. *arXiv preprint arXiv:2104.12874*.

Sagiv Y., Musslick S., Niv Y. & Cohen J. D. (2018). Efficiency of learning vs. processing: Towards a normative theory of multitasking. *CogSci 2018: Proceedings of the Annual Meeting of the Cognitive Science Society.*

Sanger, T. D. (1989). Optimal unsupervised learning in a single-layer linear feedforward neural network. *Neural networks*, *2*(6), 459-473.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016, June). Meta-learning with memory-augmented neural networks. In *International Conference on Machine Learning (PMLR),* 1842-1850.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2019). A mathematical theory of semantic development in deep neural networks. *Proceedings of the National Academy of Sciences, 116*(23), 11537-11546.

Sek, A., & Moore, B. C. (1995). Frequency discrimination as a function of frequency, measured in several ways. The Journal of the Acoustical Society of America, 97(4), 2479-2486.

Sederberg, P. B., Howard, M. W., & Kahana, M. J. (2008). A context-based theory of recency and contiguity in free recall. Psychological review, 115(4), 893.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, *27*(3), 379-423.

Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record, 4*(142-163), 1.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317-1323.

Shiffrin, R.M. & Nosofsky, R.M. (1994). Seven plus or minus two: A commentary on capacity limitations. Psychological Review 101 (2):357-361.

Shiffrin, R. M., & Schneider, W. (1977). Controlled and automatic human information processing: II. Perceptual learning, automatic attending and a general theory. *Psychological review, 84*(2), 127.

Sims, C. R., Jacobs, R. A., & Knill, D. C. (2012). An ideal observer analysis of visual working memory. *Psychological Review*, *119*(4), 807.

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

Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science, 360*(6389), 652-656.

Singleton, R. C. (1965). Maximum Distance q-Nary Codes. *IEEE Transactions on Information Theory, (10)*2, 116–118.

Sloman, S. A. (1996). The empirical case for two systems of reasoning. *Psychological bulletin, 119*(1), 3.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, *46*(1-2), 159-216.

Spens, E., & Burgess, N. (2024). A generative model of memory construction and consolidation. Nature Human Behaviour, 8(3), 526-543.

Stachenfeld, K. L., Botvinick, M. M., & Gershman, S. J. (2017). The hippocampus as a predictive map. *Nature Neuroscience*, *20*(11), 1643-1653.

Stensola, H., Stensola, T., Solstad, T., Frøland, K., Moser, M. B., & Moser, E. I. (2012). The entorhinal grid map is discretized. *Nature*, *492*(7427), 72-78

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, *153*(3736), 652-654.

Stokes, M.G. (2015) 'Activity-silent' working memory in pre- frontal cortex: a dynamic coding framework. Trends Cognitive Science, 19, 394–405.

Sutton RS, Barto AG. 1998 *Reinforcement learning: An introduction (adaptive computation and machine learning)*. Cambridge, MA: MIT Press.

Szabó, Z. G. (2000). *Problems of compositionality*. Routledge.

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *arXiv preprint physics/0004057*.

Treisman, A. (1996). The binding problem. Current opinion in neurobiology, 6(2), 171-178.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. Cognitive psychology, 12(1), 97-136.

Trick, L. M., & Pylyshyn, Z. W. (1993). What enumeration studies can show us about spatial attention: evidence for limited capacity preattentive processing. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(2), 331.

Trick, L. M., & Pylyshyn, Z. W. (1994). Why are small and large numbers enumerated differently? A limited-capacity preattentive stage in vision. *Psychological review*, *101*(1), 80.

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

Vapnik, V. & Chervonenkis, A. (1968). On the uniform convergence of relative frequencies of events to their probabilities, *Doklady Academii Nauk USSR, 181*(4).

Vapnik, V. N. (2000). *The Nature of Statistical Learning Theory. Information Science and Statistics.* Springer-Verlag. ISBN 978-0-387-98780-4.

Vaswani, A. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

Von Der Malsburg, C. (1994). The correlation theory of brain function. In *Models of neural networks* (pp. 95-119). Springer, New York, NY.

Von der Malsburg, C. (1999). The what and why of binding: the modeler's perspective. *Neuron*, *24*(1), 95-104.

Webb T. W., Frankland S. M., Altabaa A., Segert S., Krishnamurthy K., Campbell D., Russin J., Giallanza T., Dulberg Z., Reilly R. O., Lafferty J. & Cohen J. D. (2024). The Relational bottleneck as an inductive bias for efficient abstraction. *Trends in Cognitive Science, 28(9)*:829-843

Webb, T., Holyoak, K. J., & Lu, H. (2023). Emergent analogical reasoning in large language models. *Nature Human Behaviour*, *7*(9), 1526-1541.

Webb TW, Sinha I & Cohen JD (2021). Emergent symbols through binding in external memory. *ICLR 2021: Proceedings of the International Conference on Learning Representations*. https://arxiv.org/abs/2012.14601.

Wei, X. X., Prentice, J., & Balasubramanian, V. (2015). A principle of economy predicts the functional architecture of grid cells. *Elife*, *4*, e08362.

Whittington, J. C., Muller, T. H., Mark, S., Chen, G., Barry, C., Burgess, N., & Behrens, T. E. (2020). The Tolman-Eichenbaum machine: unifying space and relational memory through generalization in the hippocampal formation. Cell, 183(5), 1249-1263.

Wilken, P., & Ma, W. J. (2004). A detection theory account of change detection. *Journal of vision*, *4*(12), 11-11.

Wolpert, David H., and William G. Macready. "No free lunch theorems for optimization." IEEE transactions on evolutionary computation 1.1 (1997): 67-82.