This is an Accepted Manuscript for ***Behavioral and Brain Sciences***. This Target Article has been accepted for publication but not yet copyedited or typeset, and so may be subject to change during the production process. The article is considered published and may be cited using its DOI 10.1017/S0140525X2610510X

Excess Capacity Learning

Marina Dubova[1,2,*,**] and Sabina J. Sloman[3,**]

[1]Cognitive Science Program, Indiana University, Bloomington, IN, United States
[2]Santa Fe Institute, Santa Fe, NM, United States
[3]Department of Computer Science, University of Manchester, United Kingdom, sabina.sloman@manchester.ac.uk

***Corresponding author:** mdubova@santafe.edu

******Equal contribution

Short abstract

How do humans learn from experience? Traditionally, cognitive scientists have assumed that discovering generalizable patterns requires that humans compress rich and noisy experiences. However, recent computer science results suggest otherwise—systems can learn by 'overfitting' and expanding all the details of their experiences. We offer a new perspective on learning based on a cognitive system's representational capacity, which can be *constrained* (forcing the system to compress details of past experiences), *sufficient* (to memorize past experiences), or *excess* (allowing the system to expand on the details of past experiences). This framework has implications for understanding learning across cognitive, clinical, and developmental contexts.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

Long abstract

We introduce a new framework for understanding how cognitive systems (e.g., humans) learn from experience, based on the concept of *representational capacity*—the relative amount of representational resources devoted to encoding past experiences. Most paradigms in cognitive science have operated under the assumption that these resources are constrained, forcing cognitive systems to compress rich and noisy experiences to effectively generalize to new situations. We leverage recent advances in computer science to outline the implications of learning with *excess capacity*, or applying even more representational resources than needed to perfectly memorize all the details of one's past experiences. In particular, we review evidence suggesting that excess capacity systems can exhibit many of the characteristics of human learning, such as the simultaneous ability to memorize individual experiences and generalize knowledge to new situations. We define and differentiate between *constrained* (not enough), *sufficient* (just enough), and *excess* (more than enough to perfectly capture all the details of one's past experiences) capacity. We derive empirical properties of learning in each of these capacity regimes, and compare these predictions to effects documented for human learning. We highlight the broad implications of this framework for advancing theoretical and empirical work across cognitive, clinical, and developmental psychology.

keywords: cognition, computer science, representational capacity, complexity, generalization, learning, memory, overparameterized learning

The code for reproducing the figures is available here: https://colab.research.google.com/drive/157VWxBxALXL0TY1lL3WVrIgwrkpjBt8o?usp=drive_link

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

1. Introduction

Many theories of human cognition posit that we construct internal representations that are "simpler" than the experiences we learn from. Arguments for learning with such *constrained* representational capacity are based on hypothesized biological and cognitive limitations as well as computational results, which suggest that constructing representations that discard the noisy details of individual experiences is necessary for effectively generalizing what one has learned to new situations (see Table 1; McClelland et al., 1995; Chater & Vitányi, 2003; Sherry & Schacter, 1987; Frankland et al., 2021). The premise of learning by compressing one's experiences is so ingrained in current cognitive theorizing that to date there is no systematic framework to test whether it holds in the context of a given cognitive process or situation.

This paper introduces such a framework. It articulates four distinct learning regimes based on the effective representational capacity used to learn from past experiences. We use our proposed *capacity framework* to derive a novel cognitive principle—learning with *excess*, rather than constrained, representational capacity—which we argue holds promise for understanding how humans learn about the world. The paper provides groundwork for future empirical study of whether and when minds may learn by allocating more representational resources than are required to encode past experiences.

The remainder of this section lays the conceptual and mathematical groundwork for the capacity framework, under which *constrained capacity learning* is conceptualized as one of multiple possible learning regimes. Using this framework, we characterize the *excess capacity learning* regime, in which systems construct representations that are more complex, or "richer," than the experiences themselves. In Section 2, we situate existing cognitive models in relation to these capacity regimes and highlight that the excess capacity regime remains relatively underexplored in cognitive science. In Section 3, we review recent research in computer science that shows that constrained capacity is not required for a system to generalize successfully: Instead, learning with excess representational capacity often results in even more successful generalization. In Section 4, we synthesize this and other related literature to identify behavioral patterns that can be used to empirically diagnose the representational capacity regime of a cognitive system or process. We provide examples of findings from human perceptual learning and categorization

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

that are consistent with excess capacity learning. In Section 5, we explore broader classes of phenomena in cognitive science that might benefit from the proposed capacity framework, emphasizing how the capacity framework is a fruitful paradigm for understanding human learning in domains including cognition, development, psychopathology, and others. In Section 6, we address common questions and criticisms. Finally, Section 7 discusses the limitations and future directions of this work.

Table 1. Sample quotes from the psychological literature illustrative of the paradigm of learning with constrained representational capacity, presented in chronological order. The right column introduces the key revisions of these ideas based on the excess capacity learning perspective developed in this article.

Author	**Quote**	**Excess Capacity Learning perspective**
Mach, discussing the principle of the economy of thought in "Knowledge and Error" (1905)	*The mind feels relieved whenever the new and unknown is recognized as a combination of the known, or the seemingly different is revealed as the same, or the number of sufficient leading ideas is reduced and they are arranged according to the principles of permanence and sufficient differentiation.*	Fully capturing the details of unique new experiences, or expanding the number of ideas rather than reducing them to a few familiar ones, can also support successful learning.
Koffka, describing Law of Pragnanz in "Principles of Gestalt Psychology", p. 110 (1935)	*[...] psychological organization will always be as "good" as the prevailing conditions allow. In this definition the term "good"... embraces such properties as regularity, symmetry, simplicity and others.*	"Good" (i.e., effective for an individual's goals) psychological organization may break the simplicity requirement.
Sherry & Schacter	*... preservation of the invariances and*	Preservation of invariances

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

(1987), discussing the evolutionary account of multiple memory systems	*variances among experiences are mutually incompatible tasks that can be best handled – and perhaps only handled – by functionally separate systems.*	(i.e., generalizable patterns) and variances (i.e., the unique details of each experience) can be handled by a single system.
McClelland & Goddard (1996), discussing the complementary learning systems approach	*… there is a trade-off between the need for information-preserving, structure extracting encoding of cortical traces and the need for effective storage and recall of arbitrary traces…*	The needs for storage of arbitrary traces and extracting structure in them are not necessarily at odds.
Chater (1999)	*The simplicity principle is normatively justified—patterns that support simple representations provide good explanations and predictions on the basis of which the agent can make decisions and actions.*	Complex representations can also lead to successful predictions that guide an agent's decisions and actions.
Gigerenzer (2008)	*Good decisions in a partly uncertain world require ignoring part of the available information (e.g., to foster robustness).*	Robust learned patterns and good decisions can be attained even on the basis of all situational information.
Sherman & Turk-Browne (2020)	*Encoding individual experiences (episodic memory) and extracting regularities across experiences (statistical learning) require fundamentally different kinds of learning, yet both have been linked to the same brain region, the hippocampus.*	Encoding individual experiences and extracting regularities across experiences do not require fundamentally different learning mechanisms.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

Frankland et al., (2021)	*… the tradeoff between accuracy and generalization… aligns closely with the bias-variance tradeoff widely discussed in machine learning... 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).*	There is no fundamental tradeoff between accuracy of decoding and generalization.

Glossary.

	Definition	**Computer science analogue**
Cognitive system	A system that learns to associate situations and outcomes. Could refer to an agent (e.g., human, animal, or robot) or a cognitive process (e.g., episodic memory).	Model $y = f(x, \theta)$ (e.g., $y = \theta_1 x_1 + \theta_2 x_2$)
Situations	The contexts in which the system anticipates outcomes (e.g., visual and auditory cues when driving; experimental stimuli in a category learning experiment).	Inputs x
Experiences	The previous situations and associated	Data points (x, y)

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

	outcomes from which the system learns (e.g., visual and auditory cues paired with the consequences of driving maneuvers; previously viewed stimuli and their categories).	
Representations	Transformations of situational information used to anticipate outcomes (e.g., a perception of "good" or "bad" driving conditions; weighted average of stimulus features).	Features of inputs (e.g., $[\theta_1 x_1, \theta_2 x_2]$)
Learned patterns	Outcomes anticipated on the basis of representations of possible situations.	Predictions $f(x, \theta)$ (e.g., $\theta_1 x_1 + \theta_2 x_2$)
Representational resources	Biological, cognitive, or computational resources that contribute to the system's expressivity.	Computational resources such as compute time and disk space
Expressivity	The amount of representations the system entertains to possibly express relationships between situations and their corresponding outcomes.	Number of parameters[1] (see other measures in Appendix A)
Capacity	The system's expressivity relative to experiences it learns from. Captures the representational resources available to the system after it accounts for its experiences.	Ratio between number of parameters and number of data points (see other measures in Appendix A)

[1]For general statistical models, expressivity is not necessarily a function of the number of parameters, and also depends on the model structure (f). However, in the polynomial regression examples we use as illustrations of the computational principles, expressivity *is* a direct function of the number of parameters (coefficients). See further discussion in Appendix A.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

1.1. The capacity framework

Here, we define the core concepts of our proposed capacity framework, the result of the translation of existing results in computer science into the cognitive scientific ontology. For interested readers, Section 1.2 provides the technical details underlying the existing research in computer science, and Section 1.3 outlines how the capacity regimes can be operationalized for computational and cognitive systems. We refer the reader to the Glossary for an exhaustive list of definitions and the terms we use to refer to and disambiguate between aspects of natural and artificial learning systems.

Consider a scenario where a cognitive system aims to learn from a set of past experiences. The system seeks to choose, as a basis for future behavior, one of a set of possible representations that best captures task-relevant structure from these experiences. For example, when making a category judgment, humans may systematically transform information about the features of a stimulus (*situation*) to anticipate its category (*outcome*). Similarly, when driving, adults use auditory and visual cues (*situation*) to anticipate the consequences of driving maneuvers (*outcome*) to decide on the speed and direction of the car. The transformations the system will apply (*representation* of the situation) depend on its past *experiences* in such tasks, from which the system acquires knowledge about which situational information is relevant to the outcomes and in what ways.

The size of the set of representations the system considers depends on its ability to extract and transform features of its past experiences. We use the terms *expressivity* and *capacity* to refer to the representational resources the system has available to construct descriptions of these experiences, in an absolute sense and relative to its amount of experiences, respectively. *Expressive* learning systems entertain a large number of possible representations, and can, in principle, faithfully describe a wide variety of possible past experiences. The *capacity* of a learning system is its expressivity relative to its amount of past experiences. Systems with more capacity can retain more detail and nuance of their experiences. A system's capacity can increase as a result of either increasing expressivity or of a decreasing number of experiences it learns from. The exact nature of this relationship depends on factors like the architecture of the learning system and the structure of its experiences.
We distinguish between the following capacity regimes, which are illustrated in Figure 1:

1. *Constrained capacity* systems are not expressive enough to memorize their past experiences and so must compress these experiences, effectively discarding details deemed irrelevant.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

2. *Sufficient capacity* systems are exactly as expressive as necessary to perfectly memorize their past experiences.
3. *Excess capacity* systems have more representational resources than necessary to memorize their past experiences. These systems have the capacity to extract detailed feature information and transform this information in many and sometimes convoluted ways. The representations of excess capacity systems tend to be more complex, i.e., "richer," than the experiences themselves.

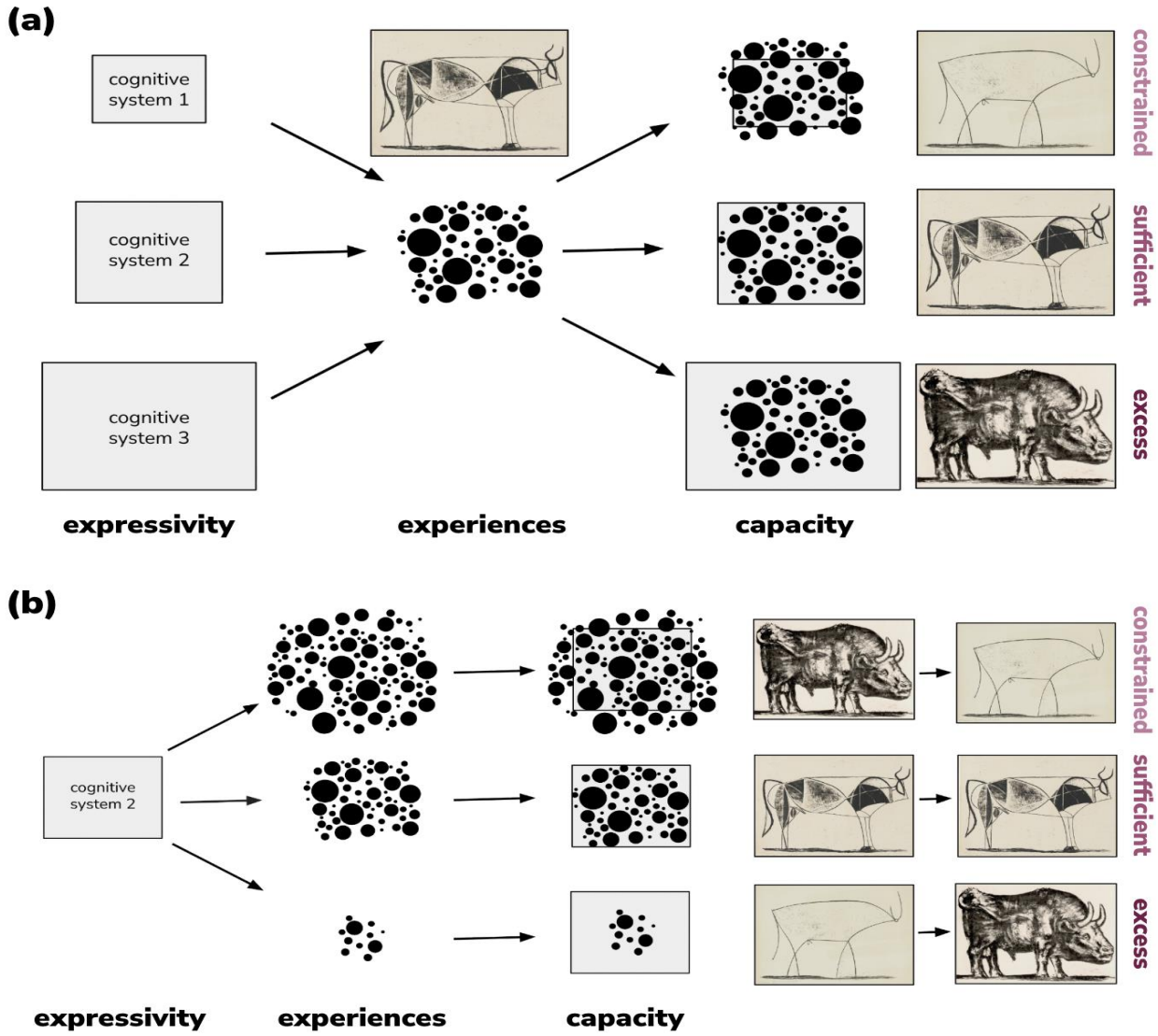


Figure 1. The relationship between expressivity and capacity in cognitive systems, illustrated through an analogy to painting. (a) Three cognitive systems with varying levels of expressivity process the same set of experiences under three distinct levels of expressivity. The resulting

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

representations can range from a simplified sketch of the experience (constrained capacity), an exact replication (sufficient capacity), to an enriched depiction with additional nuances (excess capacity). (b) A single cognitive system with fixed expressivity can encounter experiences that vary in number and complexity, leading the same system to learn in different capacity regimes. A rich, complex set of experiences (realist painting) may be simplified in the resulting representation (constrained capacity), a less detailed expressionist painting may be replicated exactly (sufficient capacity), and a simple minimalist sketch may be transformed into a more detailed and realistic representation (excess capacity). **Copyright disclaimer**: This image includes a portion of Pablo Picasso's *The Bull* (1945). It is used here under the doctrine of fair use for educational and scholarly purposes, to illustrate research ideas unrelated to the original artwork. No commercial use is intended, and these images are presented solely to enrich the academic discussion.

Constrained and sufficient capacity systems construct representations under constraints that are specified in advance of their experience (c.f., *restriction biases* in Section 1.2). Excess capacity systems, on the other hand, can represent their experiences in any of multiple ways that achieve memorization. Therefore, in order to select an ultimate solution, excess capacity learning requires a way of determining how (with the hindsight of its experience) the system will select between equally high-fidelity representations (Belkin, 2021; Belkin et al., 2019). As the research reviewed in Section 3 will show, excess capacity systems are often equipped with an intuitive, situation-independent decision heuristic that results in the selection of generalizing representations (c.f., *soft inductive bias* in Section 1.2). Our use of the term "excess capacity learning" refers to the combination of excess representational capacity with such a decision heuristic.

Box 1. Variable vs. constant capacity

Whether a system is characterized by constrained, sufficient, or excess capacity is conditional on its amount of past experiences: An excess capacity system may have more than enough expressivity to memorize a few experiences, but exhibit constrained capacity when the amount of experiences it acquires exceeds what it has the representational resources to express.

In addition to its capacity regime, a system is also characterized by whether and how its expressivity (the amount of representational resources it applies in the context of a given task) adjusts to changes in its experience. Systems with fixed expressivity can exhibit *variable* capacity, i.e., their

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

capacity depends on the amount of their experience (Figure 1b). Some systems, however, adjust their expressivity to the amount of their past experiences. Such systems can exhibit *constant* capacity (which may be constantly constrained, sufficient, or excess). In the context of artificial learning systems, common constant capacity systems include non-parametric models, such as Gaussian process regression models. Gaussian process regression models express patterns in their experiences as multivariate Gaussian distributions; the dimensionality (and in turn, expressivity) of these distributions is a direct and increasing function of the number of observed data points.[2]

In the context of cognitive models, a popular constant (sufficient) capacity model is the exemplar model of categorization. Like Gaussian process regression models, exemplar models represent new experiences as a weighted average of all past experiences. These models perfectly memorize only those data points they've encountered (exhibiting sufficient capacity for a given set of data points), but increase the amount of representational resources they employ as they acquire more data (exhibiting constant sufficient capacity). We discuss these models further in Section 6.12.

It is perhaps unrealistic to assume that natural cognitive systems adjust their representational resources to their experiences. Such scaling is in many cases computationally infeasible (the demands on such resources can grow quadratically or even cubically with the number of experiences; see footnote 2). Recognizing this limitation resonates with established cognitive traditions such as bounded rationality (Simon, 1991) and resource-rational analysis (Lieder & Griffiths, 2020). In this article, we maintain the premise of fixed representational resources, but extend this paradigm to include the possibility of fixed *excess* capacity—a regime in which systems operate with more representational resources than necessary for the task at hand, but are constrained in the sense that they cannot adjust these resources.

Unless otherwise specified, we generally assume the systems we refer to are operating with constant expressivity (variable capacity). This makes our discussion most relevant to systems whose representational resources in a given task do not adjust[3], e.g., human learners who allocate a fixed amount of memory to a category learning task. However, many of the behavioral signatures we discuss in Section 4 apply equally to constant capacity systems (see further discussion therein).

[2]More concretely, Gaussian process regression models demand more representational (in this case, computational) resources as the number of observations increases: The amount of compute time and memory required to build the set of its possible representations on the basis of n data points can be of $O(n^3)$ and $O(n^2)$, respectively.

[3]See in Section 5.3 possible application of the framework to systems with changing representational resources.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

1.2. The capacity framework: The computer scientist's perspective

The context of statistical models fitting data is a natural application of the capacity framework. A statistical model fit to a set of provided data points is an example of a cognitive system seeking patterns in its experiences. Understanding the phenomena that are present in statistical model fitting can yield insights into how cognitive systems learn from their experiences.

There is an emerging literature in computer science studying the effect a model's degree of freedom has on its behavior. This includes literature on "the theory of overparameterized machine learning" (Dar et al., 2021), "benign overfitting" (Bartlett et al., 2021), and "double descent" (Belkin et al., 2019). Throughout, we leverage these literatures to generate an understanding of the effects of capacity on a system's learning properties. These literatures draw on various operationalizations of a model's degree of freedom, that range from the number of parameters a model has (Belkin et al., 2020; Bartlett et al., 2021; Hastie et al., 2022) to the balance between the individual and aggregate size of ensembled decision trees (Belkin et al., 2020), among others. Appendix A provides a non-exhaustive list of measures from the computer science literature capturing a model's degree of freedom both in an absolute sense and relative to a given set of training data points (of expressivity and capacity, respectively).

Constrained capacity in the context of statistical models is due to *a priori* constraints on the model's expressions. These are known as *restriction biases* (Wilson, 2025), and arise from fixed aspects of the model's specification or architecture. Excess capacity models, on the other hand, can express a wide variety of possible data patterns. Preferences over representations emerge from algorithmic components of the learning process, which are often designed to satisfy the system's primary objective of fitting the available data points. In the context of statistical model fitting, such a decision heuristic is referred to as a *soft inductive bias* (Wilson, 2025). Unlike constrained capacity systems, excess capacity systems prioritize construction of a complete account of the available data points ahead of attention to their implicit preference for some solution properties over others.

In Sections 3 and 4, we will draw insights from formal investigations of the effects of the capacity of a statistical model to illustrate the ways and extent to which it can generalize to

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

unseen data points (Section 3) and derive empirical properties of learning in different capacity regimes (Section 4).

1.3. Diagnosing the capacity regime: Models

Here, we describe methods to assess the capacity regime of both statistical and cognitive models. In Section 2, we use the criteria discussed here to identify the capacity regime of some existing cognitive models.

While we discuss methods for diagnosing a *model's* capacity, cognitive scientists are often interested in the learning processes of natural cognitive systems, such as humans or animals. For these natural cognitive systems, we often don't have direct access to their experiences and whether these experiences were perfectly memorized in a given context. In Section 4, we discuss some ways in which memorization has been behaviorally assessed in humans. However, the lack of valid indicators of memorization in natural systems necessitates the development of additional empirical criteria to assess natural systems' capacity. In Section 4, we translate the extant computer science literature on the signatures of learning in different capacity regimes into behaviors that could be—and in many cases are—exhibited by natural systems.

As a toy example of measuring a model's capacity, refer to Figure 3, which shows the predictions of polynomial regression models fit to twenty-one data points. Here, the expressivity of the model increases with the number of parameters (coefficients) it has: Each additional parameter provides the model with an additional degree of freedom, expanding the set of patterns it can express. Constrained capacity polynomial regression models have fewer parameters (are of lower polynomial order) than the number of data points they are fit to, while excess capacity polynomial regression models have more parameters (are of higher polynomial order) than the number of data points.

To unpack this further, observe that the degree-one model shown in Figure 3a cannot express enough patterns to fully account for the data it sees, and so has constrained capacity: It is only guaranteed to perfectly memorize up to two arbitrary data points. The degree-twenty model shown in Figure 3b can perfectly memorize exactly twenty-one data points (once it encounters an

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

additional data point, it may be forced to tolerate some error), and so has sufficient capacity. The degree-one-thousand model can express more than enough patterns to fully memorize the data it sees, and so has excess capacity: It has multiple different expressions that could fit the data points perfectly. Figure 3c shows one of these possible expressions (the so-called "minimum-norm" solution; see additional discussion in Section 3).

In this example, each model's degree of expressivity is transparent (is directly related to the number of parameters it has). However, the relationship between the number of parameters and expressivity of a model is not so transparent in non-linear models. Parameter dependencies arise when parameters interact to produce outputs from inputs. Measures of expressivity and capacity for such cases are summarized in Appendix A.

For many cognitive models, these measures cannot be computed in practice. However, as we show in Section 2, one could directly identify if a cognitive model was operating with constrained capacity based on whether it perfectly memorized its experiences in the studied context.

2. The capacity framework reveals a gap in existing cognitive models

Figure 2 shows some existing cognitive model classes along with their corresponding capacity regimes. This figure highlights that in the cognitive science literature to date, successful learning has been predominantly viewed as a constrained capacity process, where cognitive systems' representations compress their experiences within a given context or task (Doroudi & Rastegar, 2023). This reflects the conventional wisdom that successfully generalizing to new situations requires discarding the details of individual experiences (Bates & Jacobs, 2020; Chater & Vitányi, 2003; Hampton, 1993; Landauer & Dumais, 1997; McClelland et al., 1995; Todd & Gigerenzer, 2012). Moreover, constrained capacity models are thought to reflect bottlenecks induced by limited biological and cognitive resources (Just & Varma, 2007; Lieder & Griffiths, 2020; Miller, 1956; Peterson & Peterson, 1959; Russell & Subramanian, 1994; Shiffrin & Nosofsky, 1994; Waugh & Norman, 1965).

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

On the other extreme, some cognitive models do not include any capacity limitations, proposing that a learner applies the same relative amount of representational resources regardless of their amount of experiences (Gershman & Daw, 2017; Nosofsky, 1988; Schulz, 2017). For example, the classic exemplar model posits that a cognitive system can perfectly memorize infinitely many experiences and use them to make predictions in new situations. In our discussion of the empirical signatures of learning in different capacity regimes in Sections 3, 4, and 5, we do not explicitly consider models that operate under the assumption of potentially infinite resources (see Box 1). We will return to a more detailed discussion of these models in Section 6.

Excess capacity models, which posit that a cognitive system applies more representational resources than needed to perfectly memorize their experiences, have been recently explored in computer science. Despite their growing popularity in computer science, excess capacity models are still underexplored as models of cognition. Notable exceptions include the work of Hasson et al. (2020), who advocate for an excess capacity interpretation of learning in neural systems, and applications of state-of-the-art convolutional neural networks as models of object recognition (Schrimpf et al., 2018; Yamins & DiCarlo, 2016). A key conceptual reason for the limited adoption of excess capacity models in cognitive science is the intuition that models capable of memorizing all experiences cannot simultaneously account for humans' ability to generalize from them (Table 1).

3. Decoupling memorization and generalization

A key challenge in cognitive science is understanding how cognitive systems learn from limited experiences and generalize this knowledge to novel situations. The ability to generalize has been commonly associated with constrained capacity and specifically the process of discarding the details of individual experiences (or "forgetting" of these details; Table 1). This is the idea we aim to challenge in this section. To understand the effects of capacity on generalization, we look to work examining statistical models with different numbers of flexible parameters available to fit a set of provided data points.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

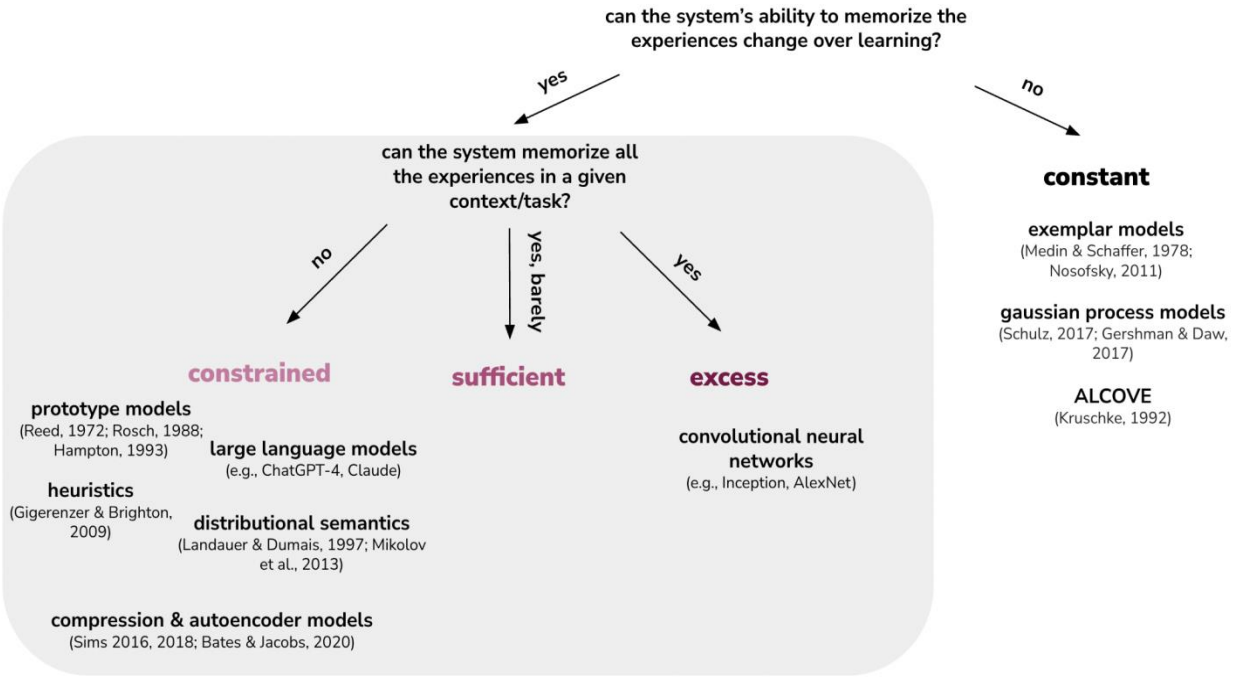


Figure 2. Example cognitive model classes and their corresponding capacity regimes. See Appendix B for details on how the capacity regime of each model was inferred. These assignments are provisional and may not encompass all applications of a given model class.

3.1. Bias-variance tradeoff

Statistical theory has long provided a compelling normative rationale for constrained capacity models of learning. These models, which avoid overfitting by focusing on fewer features and avoiding complex transformations, were thought to lead to better generalization performance (Hastie et al., 2009). This belief is grounded in the concepts of *bias* and *variance*: *Bias* refers to the error introduced by approximating a rich dataset with a simplified model, while *variance* refers to the error introduced by the model's sensitivity to noisy fluctuations in the data. Constrained capacity models typically exhibit higher bias ("oversimplify" their experiences), but benefit from lower variance, meaning they are less likely to overfit to noise in any particular sample of data.

Even in the presence of noise, constrained capacity models exhibit low variance and so often reliably capture generalizable patterns in data. Conversely, conventional wisdom holds that excess capacity models, which are more expressive and have more parameters, tend to have

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

lower bias but higher variance, i.e., are more prone to overfitting noise and therefore exhibit large generalization error. This trade-off between bias and variance has normatively grounded many "constrained capacity" accounts of the human ability to generalize. For example, in introducing the adaptive heuristic paradigm, Gigerenzer & Brighton (2009) note "*heuristics succeed... because they avoid overfitting the observations; ... by betting on lower variance rather than lower bias*" (see Appendix B for further discussion of the heuristics paradigm, a constrained capacity account of human decision-making).

The bias-variance trade-off is illustrated in Figures 3a and 3b, which show how the generalization ability of the class of polynomial regression models discussed in Section 1.3 changes as a function of the degree of the corresponding polynomial representation (i.e., the model's expressivity).[4] Figure 3a shows the predictions of a degree-one polynomial regression model. As discussed in Section 1.3, this model is *constrained* in the sense that it is not expressive enough (i.e., does not have enough parameters) to memorize the training data. Figure 3b shows the predictions of a degree-twenty polynomial regression model. As discussed in Section 1.3, this model has *sufficient* capacity: It is exactly expressive enough to perfectly memorize the training data. The dangers of overfitting can be seen by comparing Figure 3a to Figure 3b. Figure 3a shows that the predictions of the constrained capacity model capture an average trend in the data. As a consequence, models trained on twenty new data points drawn from the same distribution would make similar predictions on unseen inputs (have low variance; Somepalli et al., 2022). On the other hand, notice the erratic predictions of the model in Figure 3b in areas it has not seen. The sufficient capacity model shown in Figure 3b exhibits high dependence on the particular experiences it learns from (high variance): Were we to resample twenty new data points on which to train a new degree-twenty model, the new model's predictions would look very different from those shown in Figure 3b. Here, there is a trade-off between fitting the noise in the training data, and identifying a pattern in the data.

[4]This example was adapted from Nakkiran et al. (2021).

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

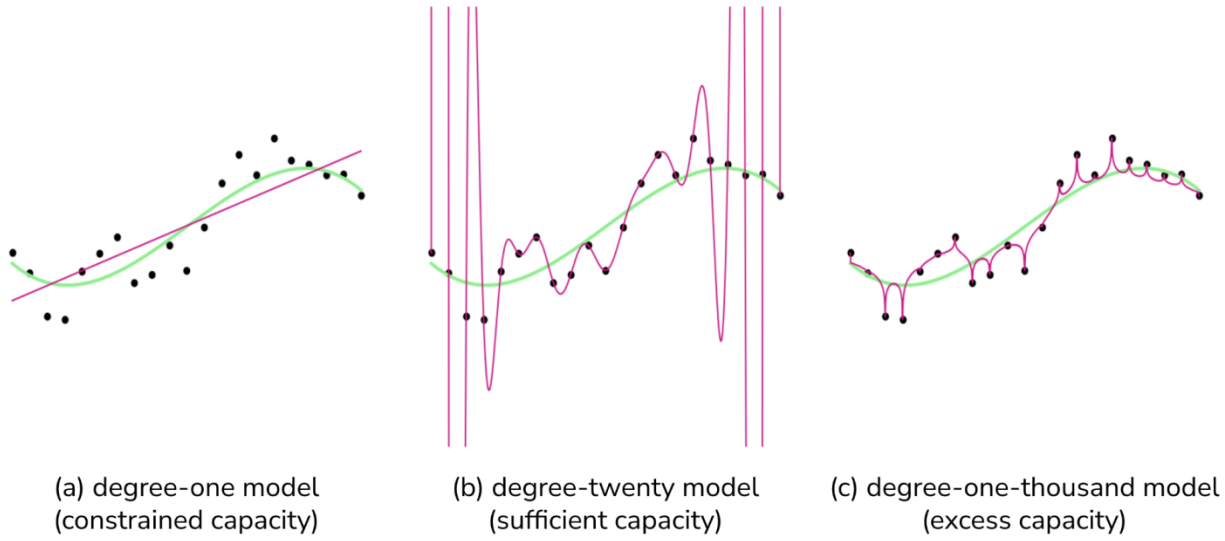


Figure 3. Polynomial regression models fit to data generated from a degree-three polynomial. Points: Training data points. Values on the *y*-axis are an additive combination of a degree-three polynomial function of values on the *x*-axis and Gaussian-distributed noise; in particular, $y = x - 0.5x^3 + \varepsilon$ where $\varepsilon \sim N(0,0.3)$. Green line: Mean of the data-generating function. Magenta line:

Predictions of the polynomial regression model with degree indicated by the corresponding caption and parameter values fit to the data points. Example is adapted from Nakkiran et al. (2021). The code for reproducing the figure is available at https://colab.research.google.com/drive/157VWxBxALXL0TY1lL3WVrIgwrkpjBt8o?usp=drive_link

3.2. Double descent of generalization error

Figure 3c shows the predictions of a degree-one-thousand polynomial regression model. This model has *excess* capacity: It is more than expressive enough (in this case, by virtue of having more than enough flexible parameters) to memorize the training data. Traditional statistical intuition would lead one to expect this to result in even more erratic predictions than those made by the degree-twenty model. Yet Figure 3c shows the opposite: While the degree-one-thousand model does fit the noise in the data, it does so in such a way that nevertheless appears to track the

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

underlying pattern in the data.[5] Like the constrained capacity model shown in Figure 3a, this model exhibits low variance in the sense that excess capacity models trained on twenty new data points drawn from the same distribution would make similar predictions on unseen inputs (Somepalli et al., 2022).

This illustrates the phenomenon that learning with excess capacity facilitates successful generalization, known as "benign overfitting" (Belkin, 2021; Dar et al., 2021b): In the constrained capacity regime, generalization error decreases and then increases as the model's expressivity causes it to overfit its data points, consistent with the bias-variance tradeoff. As the model's expressivity increases to the point that it is able to perfectly reconstruct its training data—i.e., it transitions to the excess capacity regime—generalization error again decreases, resulting in the so-called "double descent" of generalization error (Figure 4, Belkin, 2021).

In Figure 3, the best generalizing model would be the *correct* model—in this case, a degree-three polynomial. Under our definition, this is considered a constrained capacity model since it cannot perfectly fit all 20 training datapoints. This correct model would correspond to the dip in the U-curve of generalization error in the constrained capacity regime (see Figure 4). In other words, successful generalization can be achieved by building an internal model that faithfully captures the data-generating process. However, rarely do cognitive systems know the structure of the true data-generating process, and in such cases excess capacity models can be more successful at recovering generalizable patterns in the data than constrained and sufficient capacity alternatives (Dar et al., 2021; Hastie et al., 2022; N. Mallinar et al., 2022).

[5]Importantly, note that all three models are fit to the same twenty-one data points. With reference to discussion in Section 1, the cognitive systems (here, regression models) are each endowed with the same amount of experiences, but vary in terms of the expressivity they have to transform these experiences.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

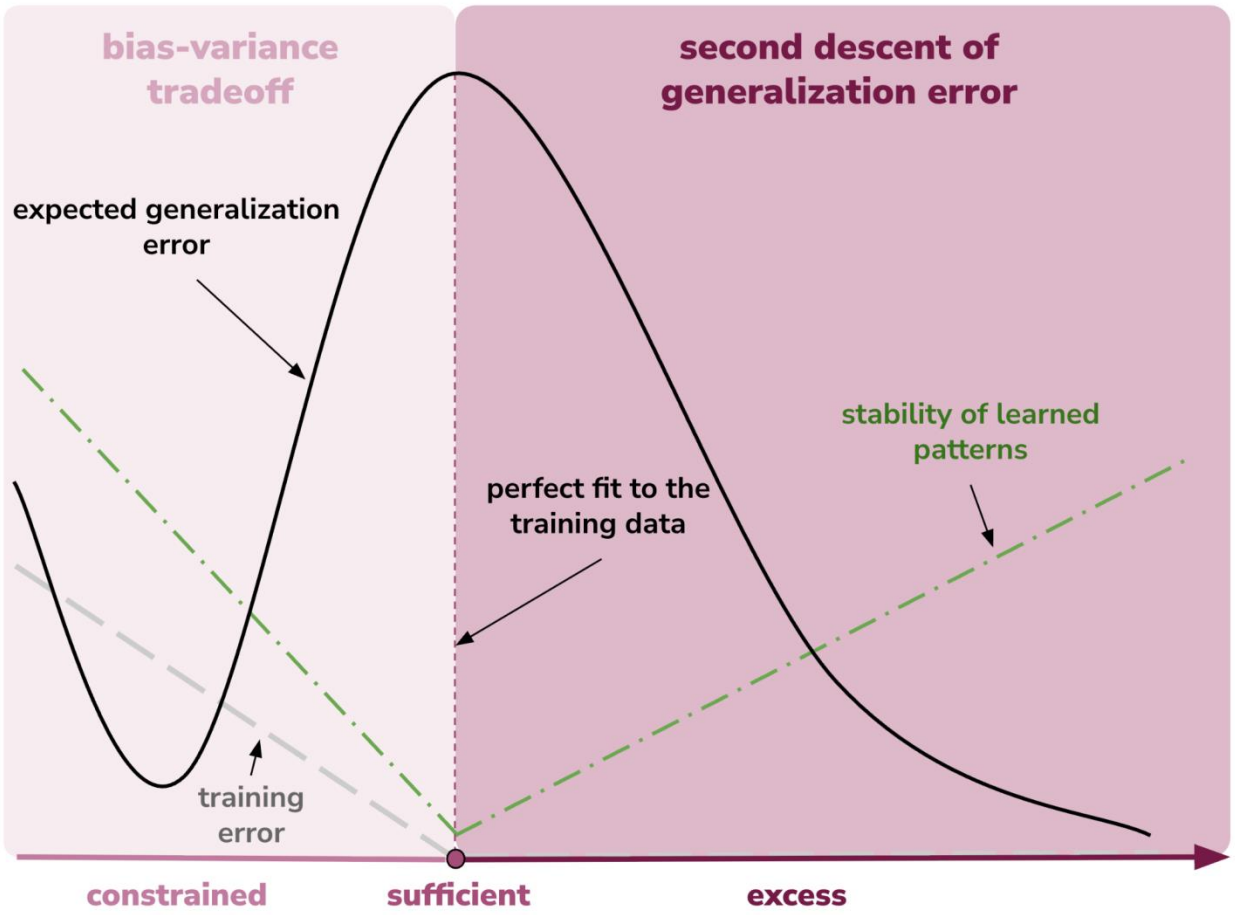


Figure 4. Conceptual illustration of the key learning properties in different capacity regimes. Memorization error (error when recalling one's past experiences; grey dashed line) decreases with increase in system's expressivity, eventually reaching zero for the sufficient and excess capacity regimes. Generalization error (error when making predictions in new situations; black solid line) follows a double-descent curve. In the constrained capacity regime, generalization error exhibits the U-curve characteristic of the bias-variance tradeoff: The solution that generalizes the best is the one that achieves the best balance between being expressive enough to capture important patterns in the data (reducing bias) and not being too expressive as to overfit the noise in the data (reducing variance). Generalization error peaks at the sufficient capacity regime which fits the data perfectly but very unreliably. Then, in the excess capacity regime, generalization error decreases with increasing expressivity (second descent). Systems with enough excess capacity identify solutions that exhibit both low bias and high stability of learned patterns (or low variance; green dashed line).

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

3.3. Why do excess capacity models generalize?

Understanding the mechanisms of the double descent phenomenon is an open area of inquiry. Notice that there are several versions of the degree-one-thousand model that could perfectly fit all twenty-one data points shown in Figure 3. As we discuss in Section 1.2, the model chooses one of these solutions according to an emergent decision heuristic.

Figures 5 and 6 visualize how the capacity of a system affects the process by which it chooses an ultimate solution. As shown in Figure 5, the constrained capacity system's initial solution set (which implies the set of possibly learned patterns) perfectly corresponds to only some of the experiences the system could have. The system chooses the solution that corresponds most closely to its experiences. If the restriction biases imposed on the system are helpful, its solution set corresponds more closely to these and future experiences. The sufficient capacity system's solution set perfectly corresponds to all possible experiences, and each set of experiences corresponds to a unique solution. Like the constrained capacity system, the sufficient capacity system chooses the solution that corresponds most closely (in its case, perfectly) to its experiences. The excess capacity system's solution set perfectly corresponds to all possible experiences, and each set of experiences corresponds to possibly multiple solutions. Unlike with the sufficient capacity system, the excess capacity system's decision-making process must, in addition to identifying the solutions that perfectly correspond to its experiences, incorporate a mechanism (soft inductive bias) to choose among these solutions.

In constrained capacity models, a preference for "simpler" solutions arises from the restriction biases imposed on the model: Complex solutions are simply not entertained by the model. In excess capacity models, a similar preference arises from the system's soft inductive biases: The preference for more stable learned patterns is an unintended consequence of architectural and algorithmic components of the learning process which are often designed to satisfy the system's primary objective of fitting the available data points (Belkin, 2021; Belkin et al., 2019; Geiger et al., 2019; Spigler et al., 2019; Wilson, 2025; Yang et al., 2020).

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

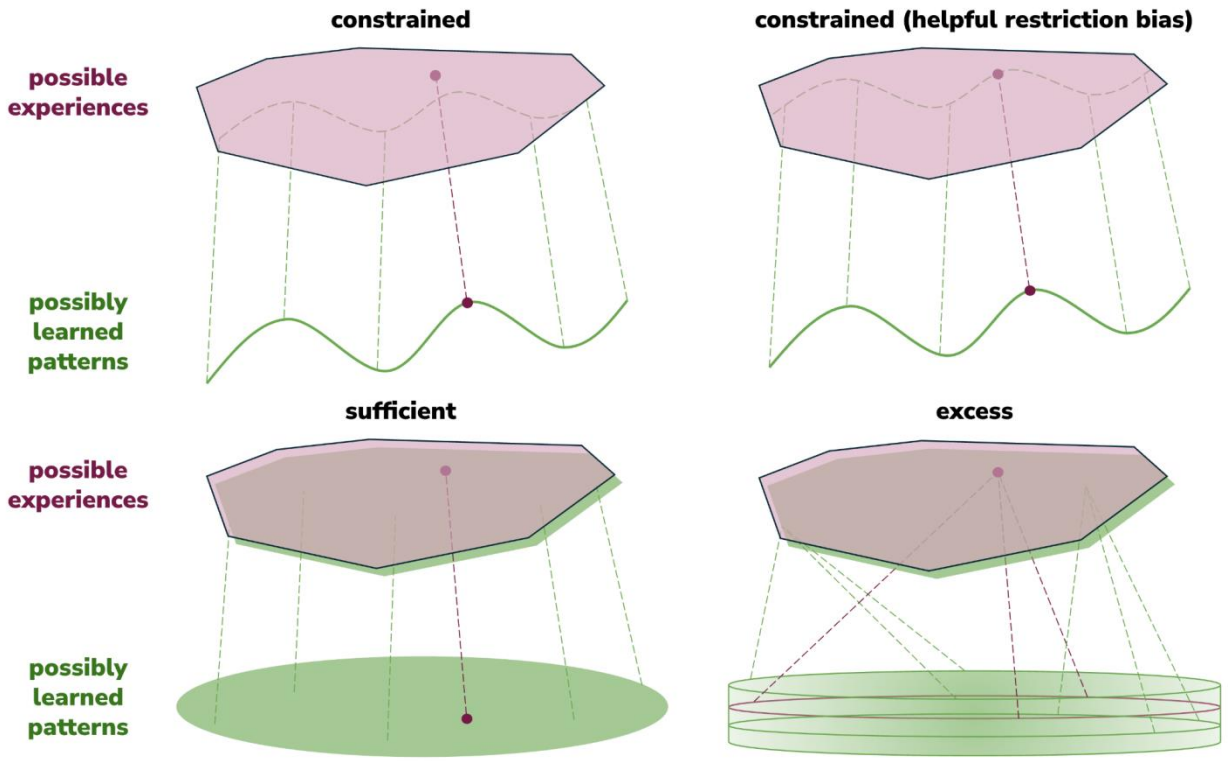


Figure 5. Visualization of how the capacity of a system (the relative size of the set of solutions it entertains) affects the process by which it chooses an ultimate solution. *Constrained:* In the case of *constrained* capacity, the patterns the system could possibly learn are less complex than the experiences they could possibly have. Most possible experiences do not correspond exactly to a member of the space of possibly learned patterns. The system maps its actual experiences (magenta dot) to the possibly learned pattern it most resembles, and selects the unique solution that produces that pattern. *Constrained (helpful restriction bias): Restriction biases* (Wilson, 2025) constrain the solution set and affect the correspondence between the space of possibly learned patterns and the system's actual experiences. Helpful restriction biases do so in a way that the system can learn patterns in its actual experiences. *Sufficient:* In the case of *sufficient* capacity, possibly learned patterns are just as complex as possible experiences. Every set of possible experiences corresponds exactly to a possibly learned pattern. *Excess:* In the case of *excess* capacity, possibly learned patterns are more complex than possible experiences. Every set of possible experiences corresponds exactly to multiple possibly learned patterns. The system's *soft inductive biases* (Wilson, 2025) guide how it selects amongst the solutions that produce patterns corresponding to its actual experiences (magenta elliptic curve). Shades of green in the space of possibly learned patterns indicate that the system's soft inductive biases are more likely to guide it towards some solutions than others. *Benign overfitting* refers to the case when the solution to which the system is guided by its soft inductive biases generalizes well.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

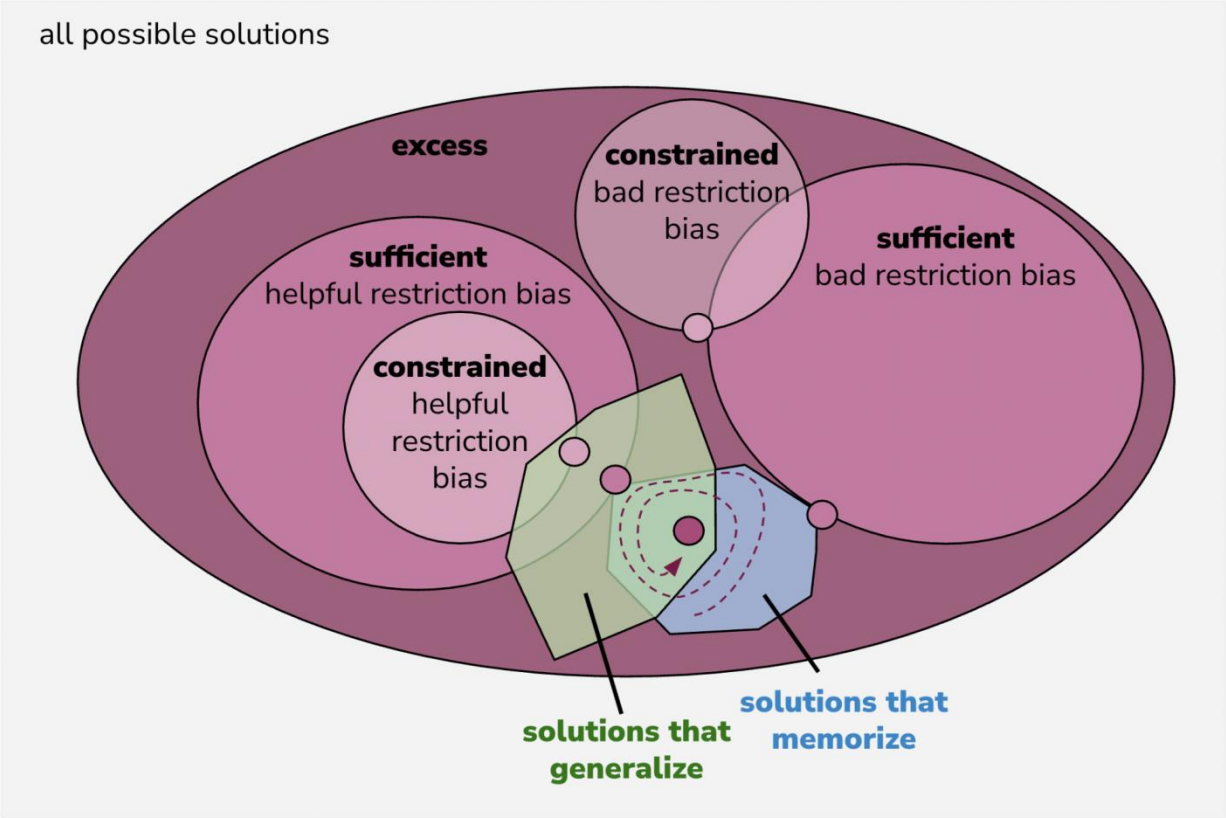


Figure 6. Visualization of how a system's ability to memorize and generalize arises from restriction or soft inductive biases. The system's capacity (the relative size of the set of solutions it entertains) is represented by the size of the magenta circles. The solution selected by each system is represented by the magenta points. Each system chooses a "best-fitting solution" as close as possible to the set of solutions that memorize. *Restriction biases* control the distance of the best-fitting solution to the set of solutions that generalize. *Constrained* capacity systems do not have the capacity to memorize, but can generalize with helpful restriction biases (e.g., imposition of a cubic form to fit the data shown in Figure 3). *Sufficient* capacity systems always memorize, and can generalize with helpful restriction biases (e.g., exemplar models that accurately predict category membership by representing new exemplars as weighted averages of past exemplars; see discussion in Section 6.12). *Excess* capacity systems always memorize, and can generalize with helpful *soft inductive biases*: These systems entertain a solution set large enough to encompass multiple solutions that memorize and that generalize. The magenta path indicates how the excess capacity system's *soft inductive biases* can guide its search within the space of solutions that memorize towards solutions that also generalize (region of overlap between the blue and green shapes).

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

Soft inductive biases emerge from components of learning processes frequently encountered in standard modeling practice. For example, common optimization procedures like (sometimes stochastic) gradient descent result in an implicit preference for solutions that do not assign excessive influence to individual features (Chizat et al., 2019; Du et al., 2019; Gunasekar et al., 2018), are not too influenced by the idiosyncrasies of a particular set of training data points (so-called "flat minima"; Blanc et al., 2020; Feng & Tu, 2021; Jastrzebski et al., 2020; Z. Li et al., 2021; C. Y. Liu & Flanigan, 2023; L. Wu et al., 2022; L. Wu & Ma, 2018; Z. Zhu et al., 2019), or collapse high-dimensional inputs into evenly-spaced partitions of possible outcomes (so-called "neural collapse"; Papyan et al., 2020; Farrell et al., 2022, 2023).

In the case of the polynomial regression model shown in Figure 3c, the chosen solution is the so-called minimum norm solution: the solution with the smallest parameter values among those that provide a complete account of the available data points.[6] The selection of the minimum norm solution (as opposed to other solutions that provide a complete account of the available data points) is a consequence of a straightforward adaptation of standard ordinary least squares to the overparameterized regime (Bartlett et al., 2021).[7]

4. Diagnosing the capacity regime: Humans

Diagnosing the capacity regime of a natural cognitive system involves a *capacity analysis*: an assessment of the degree to which the system exhibits properties characteristic of that capacity regime. In this section we derive some of the empirical properties of learning in different capacity regimes from the computer science literature and present findings from the psychological literature that illustrate each property.

In the constrained and sufficient capacity regimes, *complexity of the system's representations* tends to imply *instability of its learned patterns*: More complex representations tend to comprise patterns that are highly dependent on the specific experiences the system happens to learn from.

[6]More specifically, the chosen solution $\Theta = \text{argmin}\ \{||\Theta||_2 : X^TX\Theta = X^Ty\}$ where $||\Theta||_2$ is the L2 (Euclidean) norm of the vector of coefficients Θ.

[7]The minimum norm solution arises when the pseudo inverse replaces the inverse in the solution to the normal equations, i.e., when the solution $\Theta = (X^TX)^+X^Ty$ and $(X^TX)^+$ is the pseudo inverse of the normal matrix.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

A key principle of learning in the excess capacity regime is a reversal of this relationship. Excess capacity systems first apply their representational resources to construct representations of sufficient complexity to account for their past experiences, and then apply any excess representational resources to impose structure on the patterns implied by these representations. The tendency to produce complex representations alongside stable learned patterns is the primary characteristic property of excess capacity systems: Constrained capacity systems tend to produce simple representations and stable learned patterns, while sufficient capacity systems tend to produce complex representations and unstable learned patterns.

In addition to this principle, other properties associated with learning in different capacity regimes can help cognitive scientists diagnose the capacity regime of a cognitive system. We distinguish between 1) representational, 2) temporal, and 3) contextual properties. *Representational* properties reflect characteristics of the solutions selected by the system to capture the relationship between situations and outcomes. As mentioned above, while the solutions selected by excess capacity systems *internally* represent their past experiences in complex ways, these complex representations form the basis of *learned patterns* in outcomes and behavior that exhibit stability and smoothness. *Temporal* properties describe how the system's learned patterns evolve as the system accumulates experiences. *Contextual* properties capture how features of the environment affect the system's learned patterns.

These additional properties tie back to whether and when the cognitive system faces a tradeoff between the complexity of representation and the stability of its learned patterns. Representational properties show how the complexity of the representation can be decoupled from the stability of learned patterns if the system learns with excess representational capacity. Temporal properties characterize how the stability of learned patterns changes in excess capacity systems as the amount of their experiences grows. Contextual properties characterize the kinds of environments in which the trade-off between complexity and stability reemerges and can no longer be avoided by excess capacity learners.

Figure 7 lists a set of properties (left), the capacity regimes these properties are associated with (middle), and empirical predictions that could be used to test these properties in cognitive

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

psychology experiments (right). These properties are jointly diagnostic: The presence or absence of no single property can on its own definitively indicate a system's capacity regime; rather, diagnosing a system's capacity regime requires assessing which combination of properties a system exhibits. These properties are also provisional: Given the dynamic nature of research in computer science, we expect this list to evolve and be updated as new insights emerge.

For each property, we provide example models and studies from cognitive psychology that illustrate these properties in the context of human learning. Our goal is not to present decisive evidence or adjudicate between competing explanations, but to demonstrate how diverse findings can be interpreted through the capacity framework. Our main focus will be on highlighting evidence that points at the possibility of excess capacity learning in humans, as this regime has so far been largely overlooked. We present these examples as illustrative case studies rather than confirmatory tests, showing how the capacity framework can help to organize existing findings and predict as-yet unobserved behavioral phenomena.

	LEARNING PROPERTY	C	S	E	EMPIRICAL PREDICTION
representational	1. Retaining individual experiences		×	×	1. High recognition ability
	2. Constructing detailed and complex representations		×	×	2. "Irrelevant" or "superficial" details of the situations affect behavior
	3. Learned patterns are summaries of past experiences	×		×	3. Prototypical situations have an outsized influence on behavior
temporal	4. More experiences can hurt generalization			×	4. More experiences → less successful learning
	5. More experiences decrease smoothness of learned patterns			×	5. More experiences → new experiences are more distinguishable → new experiences might become more confusable with the old ones
contextual	6. Sensitivity to high noise environments		×	×	6. Generalization is negatively affected by noisy environments

Figure 7. Examples of the learning properties (left), the capacity regimes that they are associated with (middle: C—constrained, S—sufficient, E—excess), and the empirical predictions these properties lead to (right).

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

4.1 Ability to memorize individual experiences

Like sufficient capacity systems and unlike constrained capacity systems, excess capacity systems can by definition perfectly memorize their past experiences.

Examples from psychology. Like constrained capacity systems, in many contexts humans forget information, which can serve adaptive cognitive functions (Nørby, 2015). In other contexts, humans frequently exhibit a remarkable ability to memorize. Results from a variety of domains indicate that people store nearly all their past experiences to inform their predictions and behavior when encountering novel situations. In many tasks, people's responses to previously-seen visual stimuli tend to be faster and more accurate (Kolers, 1976; Shepard, 1967). Research on visual object perception shows that humans are often able to encode nearly all the visual items they encounter in high detail. For instance, Brady et al. (2008) asked participants to view 2500 images of individual objects for three seconds each. When asked to discriminate between these observed images and similar new images, participants were able to correctly identify the observed instance in over 80% of cases. Some of the most successful cognitive models of many cognitive processes (e.g., category learning, motor learning, dynamic decision making) are premised on the idea that people retain all past experiences for later comparison (i.e., exemplar/instance-based models; see discussion in Section 6.12).

4.2 Unnecessarily detailed and complex representations

Excess capacity systems, by definition, allocate excessive representational resources to transform the situations they encounter in various ways. Excess capacity systems expand the information they intake in varied and sometimes convoluted ways, often constructing representations for which many features of situations each contribute a little (Belkin, Ma, et al., 2018; Bartlett et al., 2020; Oneto et al., 2023). While constrained capacity systems discard details of their experiences, excess capacity cognitive systems retain all information about their experiences, constructing representations on the basis of all details, whether relevant or not.

Examples from psychology. Although much research emphasizes compression or abstraction as key principles of human learning (Bates & Jacobs, 2020), other work consistently finds that humans can retain incredibly rich details of their experiences, including details deemed irrelevant

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

or superficial. For example, items that are shown in the same context as previous items (e.g., an image shown in the same location across presentations) are more quickly and accurately recognized, suggesting that humans represent contextual details even when they are irrelevant to performance on the given task (Jacoby, 1983; Jacoby & Brooks, 1984; Kristjánsson & Campana, 2010; Schacter, 1987; Tenpenny, 1995). Further evidence shows that humans often come up with unnecessarily complex rules to make judgments. For instance, in one series of studies on category identification, adult participants were asked to identify whether a presented shape is a "triangle". While being aware of the explicit definition of a "triangle" ("a polygon with three sides having three vertices"), participants relied on unnecessary cues like the relative lengths of the sides of the triangles, to the detriment of their accuracy and speed. Similarly, when asked to identify even numbers, the participants relied on unnecessary cues like whether the number was a multiple of 10 (Lupyan, 2013). Thus, humans sometimes retain a surprising amount of irrelevant information about their experiences, and often employ complex solutions even when simple rules are readily available. Although these studies do not constitute diagnostic evidence for excess capacity learning, such findings challenge the idea that human learning is fully captured by parsimonious compression or abstraction of past experiences, and can inspire ways in which this can be further empirically tested in future work.

4.3 Learned patterns are summaries of past experiences

While excess capacity systems may internally represent situations in unnecessarily detailed and complex ways (see Section 4.2), the predictions of outcomes that arise from these representations can exhibit a surprising degree of structure: As with constrained capacity systems, the learned patterns produced by excess capacity systems often exhibit a surprising degree of smoothness: The similarity of new situations predicts the similarity of the system's predictions and behaviors upon encountering those situations (Curth et al., 2024; Gamba et al., 2023). In other words, these systems make predictions that summarize the outcomes they've experienced in similar situations. As the amount of excess representational resources a system has increases, the patterns the system learns become smoother (Bubeck & Sellke, 2023; Curth et al., 2024; Gamba et al., 2022; H. Liu et al., 2022; Rosca et al., 2020; Somepalli et al., 2022). It has been shown that smooth learned patterns can facilitate successful and robust generalization (Englesson & Azizpour, 2021; Weng et al., 2018).

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

Examples from psychology. Like constrained and excess capacity systems, people have been shown to construct summaries (e.g., an "average") of their experiences (Bowman et al., 2020; Brady et al., 2018; Murphy, 2004; Schlichting et al., 2015; Tompary & Thompson-Schill, 2021). For example, after learning a new category, humans discriminate an average category member—even if it has never been observed—faster and categorize it more accurately than any other new instance of the category (Posner & Keele, 1968, 1970). When asked to memorize a list of words (e.g., "bed", "rest", "awake"), participants construct false memories of thematically associated words (e.g., "sleep"; Roediger & McDermott, 1995). Humans have also been shown to distort their perception of an item to reflect properties of the prototypical member of the category it belongs to (e.g., achromatic bananas are perceived slightly yellow; Hansen et al., 2006; Bates & Jacobs, 2020; Dubova & Goldstone, 2021; Goldstone, 1994; Hansen et al., 2006; Harnad, 2003; Kuhl et al., 1992). As a result, many cognitive models conceptualize (perceptual, categorical, or language) learning as the compression of many and high-dimensional experiences into lower-dimensional patterns or summary representations (Bates & Jacobs, 2020; Landauer & Dumais, 1997; Sims, 2018; J. D. Smith & Minda, 1998).

4.4 More experiences can hurt generalization

Acquiring more experiences provides the cognitive system with more information about the world. Under constrained capacity, more experiences typically lead to better generalization, as the cognitive system can use the additional experiences to refine its understanding. However, the situation is more nuanced for excess capacity systems. While receiving more information still improves their ability to extract patterns from their experiences, the advantage of learning from more experiences can be offset by a critical trade-off: The system uses its excess capacity to account for these additional experiences, diminishing the resources available to construct and effectively adjudicate between multiple solutions that account for these experiences. As a result, the mechanisms that enable robust generalization under excess capacity (see Section 3) may weaken as more experiences are incorporated.

To see this, recall that whether a system is learning with constrained or excess capacity is relative to the amount of experiences it learns from. Consider an excess capacity model with enough parameters to memorize n data points. When fit to n-2 data points, this model has some

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

excess capacity. When fit to *n-1* data points, the same model now has less excess capacity. By using its representational resources to fit the new data points, the model has fewer resources to spare to impose the smoothness constraints that aid generalization.

Therefore, excess capacity learning can lead to a seemingly paradoxical phenomenon where new experiences lead to decreased generalization performance (Dar & Baraniuk, 2022; Nakkiran et al., 2021). Recall from Section 3 that benign overfitting occurs when the system has the excess capacity to both interpolate the data and impose a helpful soft inductive bias. Reductions in this excess capacity relative to experience can degrade generalization performance. The same system may therefore exhibit more or less benign overfitting depending on the amount of excess capacity that remains available after accounting for new experiences.

Examples from psychology. Although evidence on the effect of the amount of experiences on generalization is limited, some evidence from the perceptual learning literature suggests humans may sometimes exhibit learning dynamics consistent with excess capacity systems: Some studies report that an increased amount of perceptual training in humans leads to perceptual deterioration. For example, extensive training in a visual texture discrimination task has been shown to decrease participants' performance. This perceptual deterioration is due to changes in the visual cortex and not to general fatigue (indicated by its persistence when training sessions are separated by hours, identification in a trained location in the visual field, and transfer between the eyes; Censor et al., 2006; Censor & Sagi, 2008, 2009; Mednick et al., 2005; Ofen et al., 2007). Such findings illustrate how increasing experience may sometimes impair generalization. Future work could investigate the occurrence of analogous deterioration effects in other areas of cognitive psychology.

4.5 More experiences lead learned patterns to exhibit less smoothness

As discussed in Section 4.4, acquiring more experiences decreases a system's excess capacity. To the extent that the system's excess capacity creates smooth learned patterns (Section 4.3), more experiences can result in learned patterns that are poorer and less stable summaries of past experiences. In these cases, the system's predictions and behaviors in new situations are more

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

distinct—in other words, as the system acquires more experiences, it can better distinguish between new situations.

What about past experiences? Excess capacity systems identify patterns in their past experiences, and then use their excess representational resources to choose which pattern produces predictions that summarize these experiences best. What types of summaries do excess capacity systems prefer? To the best of the authors' reading, the existing computer science literature pertaining to this question is incomplete or contradictory. Existing work shows that it likely depends on the specific modeling paradigm (Gamba et al., 2023). Summaries can exhibit one of two high-level characteristics: If old experiences are treated as the *rule*, predictions conform around them (Curth et al., 2022; Gamba et al., 2023). In this case, predictions in new situations become better *local* summaries of past experiences, and so the similarity of two situations affects the similarity of the system's predictions: The system's predictions in new situations are summaries of only very similar past experiences. If old experiences are treated as *exceptions* (Gu et al., 2024), the system learns patterns that capture the overall trend in their experiences, and then adds "spikes" to account for these exceptions (Bartlett et al., 2021; Wyner et al., 2017; see Figure 8). In this case, predictions in new situations become better *global* summaries of all past experiences, and so whether a situation was encountered in the past or will be encountered in the future affects the similarity of the system's predictions: The system has retained outcomes from all its past experiences, but predicts a summary of these outcomes (which is relatively invariant to characteristics of the situation) when encountering new situations.

In the latter case, learning on the basis of more experiences (i.e., with less excess capacity) will diminish the system's ability to identify old experiences as exceptions. Figure 8 shows how the amount of experiences can affect the smoothness of the pattern the model learns. These figures show the same degree-one-thousand polynomial fit to 20 and 650 data points, respectively. The model learning from 20 data points (Figure 8a) makes similar predictions on similar unobserved data points, which can result in sharp differences between the predictions for unobserved inputs and previously-observed data points. The model learning from 650 data points (Figure 8b) makes

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

more distinct predictions on unobserved data points, but cannot distinguish between old and new experiences as well as the model in Figure 8a.

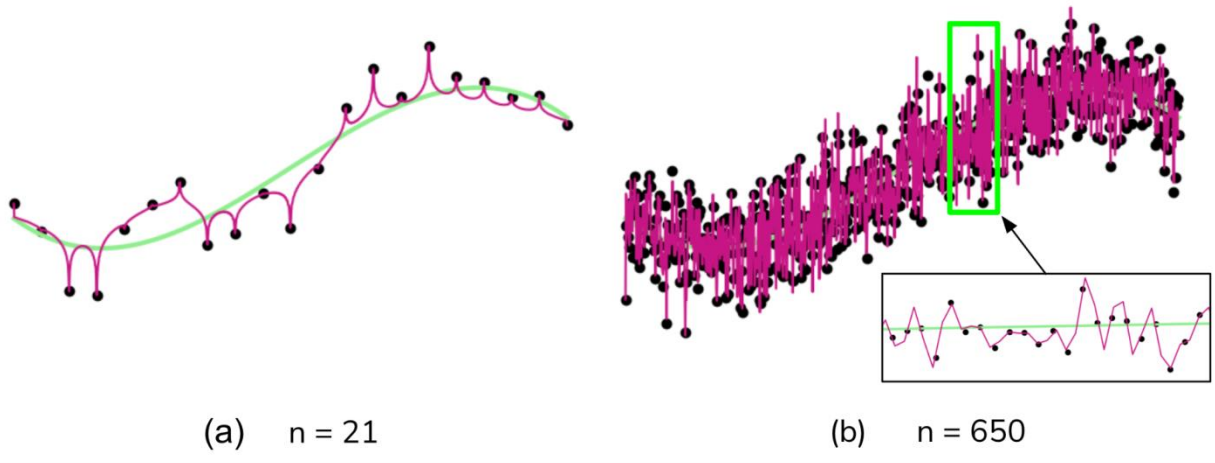


Figure 8. The predictions of a one-thousand-degree polynomial regression model when learning from (a) fewer (21) vs. (b) more (650) data points. The model learns with a higher degree of excess capacity in the setting shown in panel (a) than in the setting shown in panel (b). Values on the *y*-axis are generated as $y = x - 0.5x^3 + \varepsilon$ where $\varepsilon \sim N(0,0.3)$. In panel (a), the model assigns highly distinctive responses to the training data (spikes), while producing relatively smooth predictions for nearby unseen points. In contrast, in panel (b) the model does not capture the training data with distinctive spikes and instead generates more differentiated responses for similar new inputs.

Examples from psychology. As with excess capacity systems, human participants' ability to distinguish similar new experiences from each other has been shown to increase with the number of items they encounter in a given task. For example, studies have shown an increase in participants' ability to accurately perceive differences between new items with training, a phenomenon known as perceptual differentiation (E. J. Gibson, 1969; J. J. Gibson & Gibson, 1955). Some work on human category learning and recognition memory has also pointed to differences in people's ability to distinguish between old and new situations as a function of the number of experiences they encounter. Recall that if the excess capacity system treats old experiences as *exceptions*, as it acquires new experiences, it will begin to treat old and new

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

experiences more similarly (Figure 8). This may lead to recognition errors where participants falsely recognize new items as old.[8] Existing studies report that humans' ability to distinguish old from new experiences decreases when they learn from more experiences. For example, Koutstaal & Schacter (1997) show that adults that are shown more examples from a category (e.g., 18 vs. 6 images from the category "cats") are more prone to falsely recognizing lures from that category as images they have seen before. As another example, Homa et al. (2019) presented a category learning and recognition task to two groups of participants. For one group of participants, the same training instances repeated between the training blocks, whereas for the other group the training instances never repeated, resulting in a much larger set of items experienced over training. When asked to distinguish old from new items after training, the group with fewer experiences was able to successfully identify old items while the group with more experiences was not (Homa et al., 2019; Hu & Nosofsky, 2022). These findings illustrate how increasing experience can simultaneously refine discrimination among new stimuli while increasing confusion between old and new situations, a pattern consistent with the predictions of excess capacity learning.

4.6 Sensitivity to noisy environments

The stability of the patterns learned by constrained capacity systems is relatively unaffected by the degree of noise in their experiences (Bartlett & Kulkarni, 1998; Gamba et al., 2022; Hastie et al., 2022; Nakkiran et al., 2021; Orvieto et al., 2023), whereas the patterns learned by excess capacity systems can be highly affected by noisy experiences (Gamba et al., 2022; Hastie et al., 2022; Nakkiran et al., 2021; Orvieto et al., 2023; B. Wu et al., 2021; Zhang et al., 2021). By definition, excess capacity models do not tolerate any error in their account of past experiences. Noisy experiences—experiences with significant outcome variability for a given situation—directly impact the system's chosen representation, as it modifies its representation to incorporate all deviations from the underlying trend. Consider the constrained capacity (orange line; one-degree polynomial) and excess capacity (magenta line; degree-one-thousand polynomial) models shown in Figure 9: Figure 9a shows the models fit to data points generated

[8]Note that this assumes that recognition is based on similarity to past experiences (e.g., the probability of recalling having seen a "probe" increases with how similar the probe is to items stored in memory). This assumption is common in the literature (Nosofsky, 1988; Schacter et al., 1999).

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

in a low-noise environment, while Figure 9b shows the models fit to data points generated in a high-noise environment. The constrained capacity model's predictions are not affected by the level of noise, while the predictions of the excess capacity model, which modifies its solution to incorporate each data point, is affected by the noise in the data.

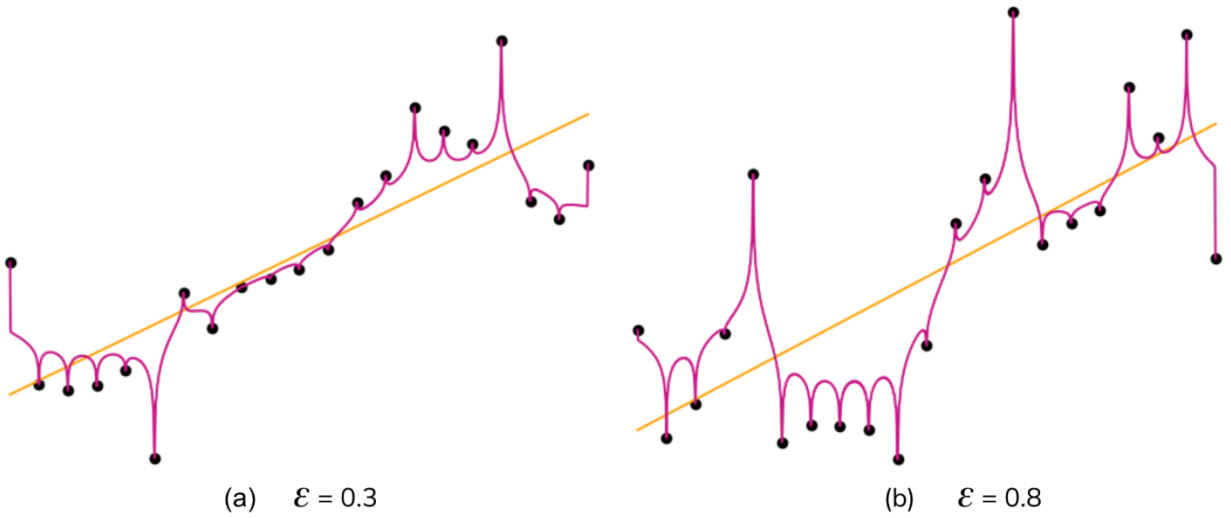


Figure 9. The predictions of a one-degree (orange) and one-thousand-degree (magenta) polynomial regression model when learning in a (a) low-noise vs. (b) high-noise environment. Values on the *y*-axis are generated as $y = x - 0.5x^3 + \varepsilon$ where $\varepsilon \sim N(0,0.3)$ in the low-noise environment and $\varepsilon \sim N(0,0.8)$ in the high-noise environment. The excess capacity model is sensitive to the level of noise (compare the magenta lines between panels a and b), while the constrained capacity model is not (compare the orange lines between panels a and b).

Examples from psychology. Some work has investigated the impact of noise on human learning in the context of visual categorization tasks using a prototype distortion paradigm (a categorization task for which categories are constructed by distorting a corresponding "prototypical" visual stimulus which is unseen by participants). Researchers manipulate the degree of distortion (variability with respect to the prototype, i.e., amount of noise) and examine the effect of this manipulation on participants' ability to categorize new, similar stimuli. Like sufficient and excess capacity systems, participants in these experiments seem to exhibit sensitivity to noise: Increasing the amount of noise in the training stimuli tends to hinder their

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

generalization performance (Bowman & Zeithamova, 2020, 2023; Hu & Nosofsky, 2024). These studies illustrate how manipulations of stimulus variability can be used to empirically test capacity-specific predictions on learner's sensitivity to noise.

5. Broad implications of the capacity framework

In this section, we discuss broad areas and phenomena that could benefit from applying the capacity framework, and the idea of excess capacity learning in particular. Our aim is not to provide an exhaustive overview of these domains, but rather to point to promising directions where the ideas may be productively applied. Ultimately, the value of these applications can only be fully assessed by experts working in these fields.

5.1 Functional relationship between episodic memory and generalization

As discussed in Sections 4.1 and 4.3, excess capacity systems are capable of both memorizing individual experiences and identifying generalizable patterns in these experiences. In cognitive science and neuroscience, these two phenomena are often considered mutually exclusive (see Table 1 for example quotes) and existing cognitive models often treat them as occurring via separate processes (so-called "dual-process architectures"; Ashby et al., 1998; Briscoe & Feldman, 2011; Erickson & Kruschke, 1998; Hahn & Chater, 1998; Juslin et al., 2008; Kumaran et al., 2016; McClelland et al., 1995; Nosofsky et al., 1994). Notable examples of such dual-process accounts are the theories of multiple memory systems (MMS) (Sherry & Schacter, 1987) and complementary learning systems (CLS) (Kumaran et al., 2016; McClelland et al., 1995). According to MMS theory, the human brain must have evolved at least two separate memory systems: the episodic memory system is responsible for memorizing unique experiences and the pattern-learning system is responsible for acquiring generalizable knowledge. CLS theory also posits two learning systems in the human brain: the hippocampus responsible for fast, episodic learning (memorization) and the neocortex responsible for slower, statistical learning (generalization) (this theory was later extended to reflect subprocesses within the hippocampus: e.g., Schapiro et al., 2017).

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

The capacity framework suggests that achieving both memorization and generalization does not require two systems: A single excess capacity system, under the right conditions, can do both.[9] This presents a challenge to the normative rationale for many dual-process accounts, which often operate under the assumption of a computational trade-off between memorization and generalization.

The functional separation between memorizing individual episodes and generalizing from them has historically been motivated by *catastrophic forgetting* exhibited by artificial learning systems (McClelland et al., 1995; Kumaran et al., 2016). Early connectionist models trained sequentially on interfering sets of associations exhibited rapid and near-complete loss of previously learned information when new associations were introduced. To many, this phenomenon suggests the existence of inherent computational limitations that prevent a system that rapidly acquires novel episodes from gradually extracting structure in these episodes. Since humans typically do not show such drastic interference, this phenomenon was interpreted as evidence that human cognition bypasses the supposed computational limitations by learning via multiple systems.

Hewson et al. (2024) present an example use case of the capacity framework by replicating the original demonstrations of catastrophic forgetting in connectionist models, and then showing that increasing the capacity of the connectionist model makes it more capable of retaining what it has learned, despite acquiring interfering associations. More specifically, Hewson et al. (2024) trained simple multi-layer perceptrons on a sequence of data points inspired by tasks designed by McCloskey and Cohen (1989) in which participants were asked to memorize arbitrary word pair associations. McCloskey and Cohen (1989) observed that, unlike many connectionist models, the human participants did not exhibit catastrophic interference: After a word was used in a new association, they were still able to reproduce its previously-learned association. By varying the number of hidden neurons, Hewson et al. (2024) explored neural networks of constrained, sufficient, and excess capacity. The study found that models with constrained and sufficient

[9]Note that this idea has also been discussed in the context of exemplar and Gaussian process regression models of cognition (Gershman & Daw, 2017; Medin & Schaffer, 1978; Nosofsky, 2011; Schulz, 2017) which simultaneously memorize individual experiences and generalize.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

capacity suffer from catastrophic forgetting the most, while learning with excess capacity made the models more robust to catastrophic forgetting in the presence of interfering associations.

While episodic memory and generalization can in principle be achieved by a single learning system, natural brains may nevertheless implement these functions through distinct mechanisms. This is consistent with extensive neural, behavioral, and clinical support for dual-process accounts. Specifically, researchers have documented various tradeoffs between representing individual episodes and generalized structure among them. For example, human and animal behavior can become increasingly governed by "gist"-like representations relative to episodic detail following sleep or over the course of learning (e.g., Tompary & Davachi, 2017; Richards et al., 2014; Zeng et al., 2021; Graves, Antony & Turk-Browne, 2020), or items that allow for better statistical prediction of the future are less likely to be memorized (e.g., Sherman & Turk-Browne, 2020).

The capacity framework developed here provides a new lens through which such findings can be interpreted. While human and animal brains may rely on distinct interacting systems, a capacity-based analysis offers a useful and generative tool for identifying when such trade-offs are expected to arise. In particular, the framework predicts that systems operating without excess capacity will exhibit trade-offs between episodic memory and generalization (see Section 3.1). Alternatively, similar temporal trade-offs may emerge when a single learning system dynamically transitions between capacity regimes in response to task demands. For example, as the number of stimuli increases or task structure becomes more complex, a system may shift from learning with excess or sufficient capacity, characterized by the ability to encode individual episodes, toward constrained capacity, characterized by increasingly "gist"-like representations. In this way, the capacity framework helps clarify which trade-offs are expected in different computational regimes and learning conditions, opening a broader explanatory space beyond one- versus two-system accounts.

5.2 Clinical dissociations

As we discussed in Section 5.1, the frequent human ability to simultaneously memorize and generalize has historically inspired dual-process accounts of learning. However, influential

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

bodies of research have reported *dissociations* in people's abilities to memorize experiences and identify generalizable patterns in them: systematic contexts in which people can either memorize or generalize, but not both. Here, we briefly describe these findings and show how they could potentially be leveraged to diagnose the capacity regimes that characterize learning in specific populations or in specific contexts.

Knowlton et al. (1994, 1996) and Squire & Knowlton (1995) report that when learning new categories of visual stimuli, amnesic patients are worse at recognizing individual training stimuli than neurotypical participants (i.e., have less ability to memorize individual experiences), but are often just as good at identifying the rule that determines the category of new items (i.e., have equal ability to generalize; see an overview of such findings in Ashby & Rosedahl, 2017). Knowlton et al. (1996) report the opposite pattern of behavioral dissociations for Parkinson's disease patients: These patients sometimes exhibit an ability to memorize individual training stimuli on par with neurotypical participants, but perform much worse at generalizing to new items (although later in training these patients often catch up to a generalization performance matching that of neurotypical and amnesic participants). Knowlton et al. (1996) explained these dissociations by invoking a dual memory systems account (see related discussion in Section 5.1: They claimed that amnesic patients' declarative memory system (responsible for memorization) is damaged while their pattern-learning system (responsible for generalization) remains intact, and that patients with Parkinson's disease have an intact declarative memory system and damaged pattern-learning system. Nosofsky & Zaki (1998) reevaluated these phenomena through a one-system model by suggesting that amnesiacs are less accurate in discriminating between similar items in memory.[10]

We suggest that the capacity analysis could be usefully applied to better understand these dissociation cases. Amnesic patients' impaired ability to memorize their experiences and successful ability to generalize from them is consistent with constrained capacity learning. The successful ability of patients with Parkinson's disease to memorize their experiences and

[10]As reported in Squire & Knowlton (1995), Nosofsky & Zaki (1998)'s account was not able to accommodate the case when categorization performance on par with neurotypical patients is accompanied by a complete inability to memorize individual items.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

impaired ability to generalize from them is consistent with sufficient capacity learning. Neurotypical participants' ability to both memorize and generalize is consistent with excess capacity learning. This account can additionally explain the observation that the ability of patients with Parkinson's disease to generalize increases over the course of training: Being exposed to more experiences may cause a system with constant expressivity to transition from sufficient to constrained capacity, which is associated with more successful generalization. Mapping these clinical profiles onto capacity regimes not only systematizes existing findings but also generates new predictions about the representational, temporal, and contextual properties each group should exhibit (Figure 7). These interpretations are merely illustrative, and future empirical work will be required to evaluate the usefulness of applying the capacity framework to clinical phenomena.

5.3 Developmental effects

Human cognitive development is another promising area for applying the capacity framework. A large literature has documented developmental changes in human abilities to memorize and generalize. We conjecture that some of these patterns can be explained by a changing amount of representational resources and experiences across the course of development, which result in transitions between capacity regimes. We note that human development involves asynchronous changes in many processes, such as executive control, attention, motor abilities and others, which, in turn, shape the very experiences learners draw on (Forest et al., 2023; Smith et al., 2018; Thelen & Smith, 1994). Changes in a system's amount of representational resources and experiences should be considered only one component of a more complicated developmental story.

5.3.1 U-shaped development

One broad phenomenon that could be interpreted within the capacity framework is *U-shaped development*: A systematic disruption of competence that has been observed in a variety of domains. U-shaped development is characterized by an initial ability to learn generalizable rules and exceptions (e.g., grammatical rules in the context of language acquisition and their exceptions), followed by abrupt failures due to overapplication of the rules and forgetting of previously-learned exceptions (e.g., using "go-goed" after having correctly used "went"; Kuczaj

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

II, 1977; Bowerman, 1982; Bybee & Slobin, 1982; Marcus et al., 1992; also see Chapter 3 in Braginsky, 2022), and finally a subsequent improvement as the appropriate scope of the rule is learned. The phenomenon has received substantial attention in developmental psychology, with seminal models suggesting that U-shaped development is connected to changes in the ability to retrieve exceptions from memory (the so-called "rule-plus-exceptions" theory; MacWhinney, 1978; Pinker, 2015), decreases in the proportion of exceptions in a child's experience across the course of early development (Rumelhart & McClelland, 1985), or the asynchronous development of multiple processes that together support a given behavior (Gershkoff-Stowe & Thelen, 2004).

Within the capacity framework, the abrupt overgeneralization of rules could indicate that a system has temporarily transitioned from an excess to constrained capacity regime. This could happen due to asynchronous growth in the system's representational resources and the amount of experiences they learn from. This could occur in cases where the amount of experiences the system tries to accommodate grows faster than their expressivity. For example, a child may encounter new words more quickly than their representational expressivity increases. During such periods, the system's capacity may be temporarily overwhelmed by the increasing number and complexity of new experiences. This period of constrained capacity would be characterized by learning patterns—overgeneralization of rules—and failure to recall exceptions. As the rate of new experiences slows down and the system's expressivity continues to increase, the system may transition back to an excess capacity regime, where they can once again apply general rules and recall exceptions to the rules. In the context of language acquisition, this account would predict that, during the temporary constrained capacity stage, overregularization is correlated with vocabulary growth (as reported in Braginsky, 2022). Interestingly, this explanation of U-shaped development predicts a brief stage where the system operates with sufficient capacity, characterized by an ability to recall exceptions and inability to retain the general rule. Although we are not aware of direct evidence of this intermediate stage in the context of human development, it offers a potentially valuable avenue for exploration.

5.3.2 Representational change across development

If human representational resources indeed increase across development, then behavior across the course of development on tasks in which participants are presented with a fixed set of items

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

(e.g., learning from visual or auditory stimuli presented in the lab) should exhibit predictable dissociations, as participants transition from constrained to sufficient to excess capacity for these tasks. In early development, constrained capacity may enable generalization without consistent memory for episodes; later, a brief period of sufficient capacity would enable stronger episodic memory but potentially hinder generalization; in later stages of development, excess capacity would enable both strong episodic memory and generalization. A recent review of behavior across the course of development in statistical learning tasks provides evidence for this trajectory, suggesting that infants first acquire "fuzzy" representations which capture general trends without stable episodic memory; later, episodic memory develops and primarily drives behavior in children; finally, adults succeed at both episodic recall and generalization (Forest et al., 2023). Similarly, studies probing simultaneous episodic recall and generalization often show that the ability to successfully generalize develops earlier, while episodic memory is still fragile. Over time, behavior becomes more reliant on stable episodic memory, eventually resulting in the simultaneous ability to memorize and generalize characteristic of adult behavior and excess capacity learning (Bauer, 2007; Newcombe & Nguyen, 2024; Ngo et al., 2021).

5.4 Changes in representations over the course of training

Empirical results on the effects of the amount of training on the characteristics of learned representations could also present a fruitful playground for theorizing within a capacity framework. For example, as participants are exposed to more examples of category members in category learning tasks, their predictions of the category membership of new items has been shown to shift from reflecting global to local summaries of previously-seen items from each category (i.e., from reflecting a summary of the entire category to a summary of only a few, very similar experiences; Smith & Minda, 1998). To explain this phenomenon, extant models have posited additional components that explicitly correspond to changes in the locality of the system's summary representations: Nosofsky (1987) developed an exemplar model (see Box 1) postulating that the size of the neighborhood from which experiences influence predictions in new situations decreases over the course of training; Griffiths et al. (2007) developed a hierarchical Dirichlet process model which forms increasingly refined category representations over time. Within the capacity framework, this change in the locality of the summaries reflected in the system's learned patterns could be accounted for differently: As discussed in Section 4.5,

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

the smoothness of the patterns learned by excess capacity systems decreases as the system acquires more experiences, resulting in predictions that reflect more local experiences.

6. Responses to criticisms & common questions

6.1 Is the capacity framework falsifiable?

The key goal of the capacity framework is to expand the theoretical repertoire of cognitive science by providing an organizing construct for theorizing about, understanding, and empirically investigating the role of representational resources on learning. As with other theoretical frameworks, its empirical evaluation can proceed only indirectly through specific models and hypotheses derived from it (Lakatos, 1976); see Hewson et al. (2024) for an example of such model testing. Therefore, the framework generates falsifiable predictions only in the context of these specific models and hypotheses: for example, once a cognitive system is hypothesized to operate within a particular capacity regime, the framework predicts the co-occurrence of specific learning properties (see Section 4), which can be operationalized and measured in psychological experiments. In Section 4, we discussed examples of experimental paradigms that can test the properties predicted by each capacity regime.

Although the capacity analysis predicts the co-occurrence of particular learning properties, these properties can also be explained by alternative models (Stanford, 2023). For example, as discussed in Section 5.1, the phenomenon of simultaneous memorization and generalization can also be explained by dual-process accounts. While the capacity framework does not necessarily provide a privileged explanation for observed phenomena, it guides theory development by specifying when and why different learning properties should co-occur and how they are expected to change across learning conditions.

We adopt the view that theoretical frameworks in science are evaluated not by a single decisive empirical test but by their ability to organize existing findings, anticipate unobserved phenomena, and usefully guide model development. In this sense, frameworks differ in their explanatory and predictive utility. We propose the capacity framework as a contribution of this kind (see Section 6.2). Therefore, systematic failure of the predicted relationships described in

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

Section 4 would count as evidence against the framework, undermining its usefulness as a tool for theoretical and empirical progress.

6.2 How can the capacity framework be useful?

The capacity framework is valuable because once a cognitive system is hypothesized or identified to belong to a particular capacity regime—whether constrained, sufficient, or excess—it enables precise predictions about properties of the system's learning process. For example, as described in Section 5, overgeneralization is expected under constrained capacity, while rich representations of rules and retention of exceptions arise under excess capacity. Observing the co-occurrence of several such properties allows researchers to infer the system's capacity regime, and once identified, the framework can predict additional properties (e.g., whether the learning success will be affected by noise). Moreover, the capacity framework provides cognitive scientists with a tool to create new explanations and anticipate properties of learning based on a body of research in another discipline. The framework establishes a connection to an active area of research in computer science, which will enable cognitive scientists to revise the framework to integrate theoretical insights and new regime-specific properties as they are uncovered in computer science.

6.3 Does the capacity framework apply only to certain modeling paradigms (e.g., connectionism)?

One concern might be that the properties discussed in Sections 3 and 4 characterize only models that share a particular structure. For instance, some might wonder whether the ability to both memorize and learn summarizing patterns is unique to large neural networks.

Although the properties of excess capacity learning are an active area of research with many unanswered questions, the available evidence from computer science suggests that the capacity framework and the properties that we outlined for different regimes are not specific to any single modeling paradigm. Rather, they appear to be general consequences of learning in a particular capacity regime. For example, the phenomenon of double descent in generalization error has been demonstrated in a variety of learning algorithms and across modeling frameworks, such as decision trees, regression models, and neural networks (Belkin, 2021; Belkin et al., 2019). The

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

generality of the phenomena indicates that the insights drawn from the capacity framework can be applied broadly to different types of learning systems, making the framework applicable across diverse cognitive areas of study.

6.4 Do excess capacity systems adhere to the principle of parsimony?

The principle of parsimony, or the preference to form simpler representations of experiences, has been extremely influential in cognitive science. The principle that cognitive systems prefer simpler representations of their experiences is often based on the idea that these representations result in better generalization (see Section 3.1). The principle of parsimony has been backed up by extensive evidence in domains that range from human perceptual organization (Pomerantz & Kubovy, 1986) to explanatory preferences (Lombrozo, 2007), and has been proposed as a fundamental cognitive principle (Chater, 1999). Without undermining these contributions, we contextualize excess capacity learning within this literature. We suggest that excess capacity learning prompts us to revise what parsimony is and whether it is necessary for learning from experience.

At an initial glance, it might appear that the paradigm of excess capacity learning undermines the principle of parsimony: Excess capacity models learn and generalize very well (Section 3.2), and might even account for some reported properties of human learning (Section 4), yet by very different means than those that motivate the parsimony principle. Excess capacity systems learn by finding detailed solutions that account for every experience (Section 4.2), and are by definition more flexible, or complex, than the data themselves. These systems can fit any pattern that comes their way.[11] The paradigm of excess capacity learning illustrates that there are many ways to learn successfully, with or without an explicit preference for parsimonious representations.

However, even though excess capacity systems do not have an explicit preference for parsimonious representations, the soft inductive biases that arise in such systems often result in solutions that are consistent with some notions of parsimony (see Section 1.2 and 3.3). This has prompted computer scientists to propose new notions and metrics to capture the form of

[11]Given the amount of experiences does not exhaust the system's excess capacity.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

parsimony that arises in excess capacity learning (Curth et al., 2024; Dwivedi et al., 2023; Dziugaite et al., 2020; Hodgkinson et al., 2023). These discussions enrich our understanding of what parsimony is and whether it is important for successfully learning from experience, both in cognitive science and in science more broadly (see Bunge, 1963; Dubova et al., 2025 for further discussion of the parsimony principle in scientific practice).

6.5 When is overfitting benign?

Excess capacity systems can often both generalize (Section 3) and recover stable patterns in their experiences (Section 4.3). These properties characterize the statistical phenomenon of "benign overfitting" (Bartlett et al., 2020; Belkin, 2021; Belkin et al., 2019; Bubeck & Sellke, 2023; Hastie et al., 2022). However, overparameterization can also lead to unstable predictive patterns and unreliable generalization ("catastrophic overfitting") (Belkin et al., 2019; Dar et al., 2021; Hastie et al., 2022; Mallinar et al., 2022). This leads to the crucial question of *when* learning with excess capacity leads to high generalization ability, and whether these contexts align with those in which humans also successfully generalize (DeLosh et al., 1997; Kane & Broomell, 2020; Zhu et al., 2009).

Some recent work has begun to elucidate the conditions under which one should expect benign overfitting to occur. For example, there is some evidence that the occurrence of benign overfitting depends on the presence of model misspecification (Dar et al., 2021b; Gu et al., 2024; Hastie et al., 2022), the amount of noise (Gu et al., 2024; Hastie et al., 2022; Ju et al., 2023; Nakkiran et al., 2021), and the number of task-relevant features in the experiences (Bartlett et al., 2020). Further exploration of the excess capacity learning regime should entail developing a better understanding of whether the conditions that predict benign vs. catastrophic overfitting can also predict people's ability to generalize in a given context.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

Box 2. Benign overfitting vs. adversarial robustness

The presence of benign vs. catastrophic overfitting is closely related to *adversarial robustness*. Here, adversarial examples are identified by a so-called "adversary" to maximize the likelihood that a model—often a neural network—makes an incorrect prediction (Szegedy et al., 2014). For example, changing only some pixels of an image of a dog can cause a Convolutional Neural Network (CNN) to incorrectly identify the image as an ostrich (Szegedy et al., 2014). There is mixed evidence about whether excess capacity promotes (Chen et al., 2024) or detracts from (Hao & Zhang, 2024) a model's robustness to such adversarial attacks, and the research into the factors that affect the relationship between capacity and adversarial robustness is ongoing (Ribeiro & Schön, 2023).

Importantly, adversarial examples are not *representative* in the same way as situations on which generalization is typically assessed, in the sense that they are selected according to the adversary's criterion rather than to represent the naturally expected distribution of situations. While a simultaneous ability to generalize to representative examples and potential vulnerability to adversarial examples may not at first glance appear to have analogues in the context of human learning, this phenomenon is in some sense both unsurprising and has been discussed in connection to well-known behavioral phenomena. An excess capacity system that tries to account for any degree of imperfection in its experiences will reproduce these imperfections when encountering very similar situations (Belkin, Hsu, et al., 2018; Belkin, 2021). Cognitively, this can be thought of as "over-adjusting" one's predictions and behaviors to reflect unrepresentative ("fluke") outcomes. One line of research is examining whether humans misclassify the same adversarial images as CNNs do, especially under time pressure or when the correct answer is not provided (Elsayed et al., 2018; Veerabadran et al., 2023; Z. Zhou & Firestone, 2019). Separately, there is a debate over whether visual illusions exhibited by humans (e.g., seeing static objects as moving) can be interpreted as analogous to adversarial sensitivity (e.g., Kriegeskorte, 2015; Wichmann & Geirhos, 2023). The degree of adversarial sensitivity exhibited by humans remains relatively underexplored—perhaps due to difficulty defining and measuring it in behavioral studies, or due to the lack of (e.g., gradient) information that would enable efficient search for such examples in humans.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

6.6 Does excess capacity learning lead to good extrapolation (and does it matter)?

Extrapolation, loosely defined as successful generalization beyond the boundaries of past experiences or in substantively new situations, is an important component of ecologically adaptive behavior in cognitive systems. It has been argued that humans not only use their past experiences to generalize within the same environment, but to also successfully navigate new, substantively different environments (Gigerenzer & Brighton, 2009), often doing so from only a few examples (Lake et al., 2017).

An emerging literature has begun to investigate whether and when excess capacity models are able to extrapolate. Most of the results we have discussed so far assume that the data points the model will encounter in the future are qualitatively similar to those it has encountered in the past (e.g., as in Figure 3). This assumption can be relaxed in a variety of ways. For example, the model may extrapolate the patterns it learns to data points that are outside the ranges of what it has seen before. Even for this relatively slight relaxation of the usual setting, the evidence on the effectiveness of excess capacity models is mixed: Some work suggests that excess capacity can facilitate this type of extrapolation (Byrd & Lipton, 2019; Tripuraneni et al., 2021) and other work shows that it hinders it (Hasson et al., 2020). A more challenging setting is where the appropriate pattern to learn itself changes because of changes in the underlying environmental mechanics or because the model is applied to a novel task. Some work has suggested that models with higher capacity perform better in such cases (Hao et al., 2024; Hendrycks et al., 2021; Hendrycks & Dietterich, 2019; Yang et al., 2020). This relative performance depends on several factors, including the similarity between and relative variance of past and future tasks (Dar & Baraniuk, 2022; Lampinen & Ganguli, 2018; Mallinar et al., 2024). In general, excess capacity models can cope with environmental change much better if given the opportunity to relearn in the novel environment, i.e., to "fine-tune" their solutions to the new environment (Dar & Baraniuk, 2022; Huang et al., 2022; Ju et al., 2023).

However, the usefulness of the excess capacity systems for cognitive science may not hinge on their ability to extrapolate: Although there is a seeming consensus that extrapolation is an important human capability, this notion has also been challenged from many perspectives. For example, in many cases, humans seem to project trends they extracted from experiences in past

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

contexts to future contexts in ways that can be incorrect or maladaptive (DeLosh et al., 1997; Li et al., 2022). Some also argue that what appears to be successful extrapolation is actually yet another instance of successful interpolation (in other words, that there's "nothing new under the sun"). For example, Hasson et al. (2020) argued that our real-life experiences are so rich and varied that we are hardly ever encountering substantively new contexts. On the other hand, some work suggests that the ability to extrapolate is required to generalize in even apparently similar situations: From a statistical perspective, it has been shown that a model that generalizes beyond even a few data points is essentially always extrapolating if these data points lie in a sufficiently high dimensional space, i.e., are characterized by many features (Balestriero et al., 2021).

6.7 Do excess capacity systems confabulate?

Confabulation, or the generation of false information, is a phenomenon exhibited by both artificial and natural cognitive systems. For example, Large Language Models (LLMs) sometimes generate entirely fabricated references. Recent work has proposed that confabulation in LLMs may be inevitable (Xu et al., 2024) or even functionally beneficial (Sui et al., 2024). Humans can similarly generate false memories of past events (i.e., be led to believe they experienced events that never occurred; Loftus et al., 2014; Loftus & Pickrell, 1995) or of experimental stimuli (e.g., falsing recalling words or images that are semantically related to studied items; Koutstaal & Schacter, 1997; Roediger & McDermott, 1995).

Confabulation is not unique to excess capacity systems (in particular, LLMs are constrained capacity systems; see Appendix B and Section 6.9). As we discuss in Section 4.6, sufficient and excess capacity systems that commit to memorizing their experiences are bound to make errors when data are noisy. However, the properties of the representation the system ultimately selects (e.g., whether it overfits to or abstracts from noisy experiences) may influence the form that confabulation takes. Understanding whether and how confabulation differs across capacity regimes is an exciting direction for future research.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

6.8 What about biological constraints?

The biological pressure to conserve metabolic resources has been leveraged as an argument against cognitive models that posit high capacity (Laughlin, 2001). In response, we offer three considerations: 1) learning with excess capacity does not necessarily imply a high demand on metabolic resources, and may in fact offer evolutionary advantages in generalization and learning efficiency, 2) the concept of the brain as an extremely high-capacity organ is already widely explored in neuroscience, and 3) excess capacity learning is a paradigm that can be realized in systems with very different biological and physical properties, and so need not reflect the constraints of a specific biological entity, such as a number of neurons or synapses in a brain.

First, excess capacity learning offers advantages that could have contributed to its evolution in biological systems. As discussed in Section 3, excess capacity learning can facilitate generalization, which may enhance the adaptability of cognitive systems across diverse environments (Hasson et al., 2020). Moreover, excess capacity learning is often associated with more efficient convergence on good solutions. Specifically, excess capacity systems are sometimes faster and more likely to converge on globally optimal solutions, instead of getting trapped in local minima (Arora et al., 2018; Bartlett et al., 2021; Simsek et al., 2021). Rather than placing a higher demand on metabolic resources, this ability to learn better solutions more efficiently could result in savings of metabolic resources, suggesting that excess capacity learning could have been selected for during evolution.

Second, the enormous representational capacity of nervous systems (e.g., the human brain) is not a novel idea in neuroscience. Existing evidence suggests that neural circuits may redundantly code and expand information (Spanne & Jörntell, 2015). For example, Olshausen & Field (2004) report that "*in cat V1 there is an approximate 25:1 expansion ratio in terms of the number of axons projecting from layers 2/3 on to higher areas relative to the number of inputs from the lateral geniculate nucleus*". Hasson et al. (2020) offer extensive evolutionary arguments for the plausibility of excess capacity learning, showing that it might be consistent with the amount of representational resources that brains possess.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

Third, excess capacity learning can be realized in different ways and so does not depend on a specific biological implementation, such as the number of neurons or synapses in the human brain. Our framework can be applied to cognitive systems broadly considered, including individual cells (França et al., 2024; Lyon & Cheng, 2023; Shomar et al., 2022), embodied minds, machines, and social and human-machine systems (Dubova et al., 2022). The amount of representational resources multi-agent systems possess is an open question and likely exceeds that of individual natural and artificial learners.

6.9 What capacity regime characterizes modern state-of-the-art neural network models?

Modern state-of-the-art neural networks span a range of capacity regimes. Here, we discuss Convolutional Neural Networks (CNNs) and Large Language Models (LLMs), both of which have been widely used in cognitive science and illustrate different ends of the capacity spectrum (see additional discussion in Appendix B). CNNs have been increasingly used in cognitive neuroscience as models of human and primate object recognition (e.g., Schrimpf et al., 2018). Many CNNs, such as Inception and AlexNet, have been shown to operate in the excess capacity regime. These models can perfectly fit even randomly labeled versions of their training datasets, and still achieve high generalization on structured inputs (Zhang et al., 2016). This ability to simultaneously memorize any dataset of a given size and generalize is characteristic of excess capacity learning.

LLMs have been recently used to model various aspects of human behavior. Despite having a huge number of parameters, based on existing evidence LLMs such as GPT variants are better understood as operating in the constrained capacity regime. These models are trained on massive datasets that appear to exhaust their expressivity. Estimates suggest that these models memorize only a small fraction (0.2%–1.5%) of their training data, with fine-tuning raising that percentage in some cases (Carlini et al., 2022; Nasr et al., 2025). Although evidence pertaining to the capacity regime of state-of-the-art LLMs is limited, based on current findings that production models are unable to fully memorize their training data, we provisionally classify LLMs as falling in the constrained regime. Importantly, language models do not have to be trained on vast amounts of language data, which has resulted in proposals to train them on much smaller, developmentally plausible, datasets (e.g., the BabyLM challenge; Warstadt et al., 2023; Hu et al.,

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

2024). Given the multitude of LLMs and their applications, we acknowledge the possibility that some LLMs may be learning with excess representational capacity.

6.10 What capacity regime characterizes multi-process models?

In Sections 1-4, we have used the capacity framework to characterize individual learning systems, meaning that a single functional component is assumed to do all the work. However, many influential theories in cognitive science posit that cognitive functions are supported by multiple (often, two) interacting systems (Minda et al., 2024). These include the multiple memory systems theory (Sherry & Schacter, 1987), the complementary learning systems framework (McClelland et al., 1995), the rules-plus-exceptions models (MacWhinney, 1978; Nosofsky et al., 1994), and dual-process theories of decision making (Kahneman, 2011; Frank et al., 2009; Sanfey & Chang, 2008), among others. In Section 5.1, we discuss how the capacity framework invites a reassessment of whether positing multiple systems, as opposed to a single excess capacity system, is required to explain behaviors previously thought to be functionally incompatible (such as memorization and generalization).

Here, we reflect on whether and how the capacity framework can be applied in the context of the multi-process architectures. We propose two main directions in which the capacity framework may be informative:

1. **Assessing the capacity of the multi-process system as a whole.** One approach is to treat multi-process architectures as single (composite) cognitive systems. For example, in the CLS model discussed earlier (see 5.1), the hippocampus and neocortex could be jointly considered an excess capacity system which is, thanks to the combined expressivity of the subsystems, capable of both memorizing individual episodes and generalizing across them. While such assessments are possible in principle, it is unclear whether doing so would meaningfully inform predictions about the behavior of multi-process systems. Specifically, the validity of diagnosing the capacity of the multi-process system as a whole (i.e., of assessing whether the properties outlined in Section 4 accurately characterize its behavior) hinges on how the subsystems interact (for example, whether they share representational resources and information or instead compete for these resources). For example, the CLS model involves two systems that cooperate and share

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

information. In contrast, in the COVIS model of category learning (Ashby et al., 1998), explicit and implicit systems process experiences in parallel and compete to determine behavior. In this latter case, it may be more meaningful to separately analyze the capacity of each of the two subsystems, rather than treating the model as a single composite system.

2. **Assessing the capacity of individual processes.** As alluded to with the example of the COVIS model, one can also analyze the capacity regime of each of the subprocesses separately. For example, in the CLS model, the neocortex might be thought of as a constrained capacity process because it abstracts regularities across many experiences but cannot retain all their details. In contrast, the hippocampus resembles a sufficient capacity process, capable of rapidly storing individual experiences in non-overlapping traces. This use of the framework could help modelers anticipate the cognitive properties associated with each subsystem individually and predict what properties might be lost if one of the subsystems was lesioned. Such predictions would still need to be contextualized in details of how the processes interact with each other in the full model architecture.

6.11 What capacity regime characterizes cognitive neuroscience models?

In Sections 1-4, we focused primarily on cognitive models which abstract away from the underlying neural architecture. However, there is a rich modeling tradition in cognitive neuroscience that aims to capture both behavioral phenomena and the details of their underlying neural mechanisms. These models often simulate hundreds of interconnected neurons across multiple brain areas and have been instrumental in advancing our understanding of learning, generalization, memory, cognitive control, and other processes (O'Reilly & Munakata, 2000).

Characterizing the capacity regime of these cognitive neuroscience models is non-trivial and has to be approached with a degree of nuance that is outside the scope of this paper. Many of these models include multiple interconnected processes or regions and specific architectural decisions that make it difficult to know where they fall on the capacity spectrum. The complex interactions between regions corresponding to different learning systems make it unclear how to apply the diagnostic criteria outlined in Section 4. We discussed this issue in Section 6.10. Here, we consider one prominent cognitive neuroscience model, C-HORSE as an example.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

The C-HORSE model (Norman & O'Reilly, 2003; O'Reilly & Rudy, 2001; Ketz et al., 2013; Schapiro et al., 2017) formalizes hippocampal learning in two anatomical pathways. The trisynaptic pathway encodes individual experiences and the monosynaptic pathway constructs overlapping representations of these experiences. The monosynaptic pathway by itself resembles a constrained capacity learning system that compresses the details of its experiences. Moreover, in some applications of the C-HORSE model, the trisynaptic pathway has enough capacity to form a distinct representation of each experience, thus resembling a sufficient or excess capacity system (Zheng et al., 2022; Sučević & Schapiro, 2023). Like the models discussed in Section 6.10, since C-HORSE is characterized by both pathways interacting with each other to support behavior, the model as a whole functionally resembles learning with excess capacity in some applications.

Importantly, this characterization is not guaranteed and depends on specific applications of the model: targeted computational assessments, using the criteria described in Section 1.3 and Appendix A, are required to measure the capacity of individual subprocesses and the full model. This example highlights the potential and open challenges in applying the capacity framework to interpret complex multi-process models, such as those developed in cognitive neuroscience (see Section 6.10).

6.12 What capacity regime characterizes episodic and instance-based models?

Box 1 discussed systems for which the amount of representational resources scales with the number of experiences the system aims to capture. Many episodic and instance-based models, such as exemplar and Gaussian process regression models, fall under this *constant* capacity regime (Figure 2). These models posit that the system maintains the same level of detail when representing past experiences regardless of the number of these experiences.

In the context of categorization, the exemplar model suggests that cognitive systems store each encountered instance of a category in memory as a so-called exemplar (Medin & Schaffer, 1978; Nosofsky, 1988, 2011). When encountering a new item, the system predicts its category by potentially using all these stored exemplars. In the context of reinforcement learning, the episodic reinforcement learning model suggests that cognitive systems store each past episode as

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

a detailed memory trace (Gershman & Daw, 2017). Unlike many other reinforcement learning models, which assume that information from past experiences is discarded, the episodic reinforcement learning model posits that future decisions are influenced by information from all past experiences, proportional to their similarity to the current situation. The amount of representational resources the system uses therefore scales with the amount of experience. Similarly, in the context of function learning and search tasks, Gaussian process regression models posit that cognitive systems model relationships between situations and outcomes by mapping past experiences onto a function space whose complexity grows with the number of experiences (Schulz, 2017; Schulz et al., 2019; Wilson et al., 2015; C. M. Wu et al., 2018)[12]. Overall, such constant capacity models which posit that humans retain all past experiences to guide future behavior have been successful in describing human learning in domains ranging from skill automatization (Logan, 1988), visuomotor learning (Gorman & Goldstone, 2022) to dynamic decision-making (Gonzalez et al., 2003; Gonzalez & Dutt, 2011).

What distinguishes constant capacity models is their assumption that representational resources are not a fixed, invariant characteristic of a cognitive system but instead dynamically expand in response to the number of experiences encountered. Despite this contrast with the models discussed so far, which assume that representational resources are finite, our framework can still characterize many of the properties of constant capacity learning: Each of these models operates within a constant capacity regime—whether constrained, sufficient, or excess (see more in Appendix B). Thus, the properties discussed in Section 4 are also applicable to these systems (except for temporal properties in Sections 4.4 and 4.5, which specifically invoke the dependence of the system's capacity on the amount of its experiences).

7. Excess capacity learning: Limitations and future directions

In this article, we introduce the capacity framework, which conceptualizes learning as a function of the representational resources a cognitive system devotes to capturing past experiences. We distinguish four capacity regimes—constrained, sufficient, excess, and constant—and note that cognitive science has predominantly viewed human learning as a constrained capacity process.

[12]There is a connection between exemplar models and Gaussian process regression; see Ashby & Alfonso-Reese (1995) and Jäkel et al. (2007, 2008, 2009).

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

We highlight excess capacity learning, or learning with more representational resources than are necessary to memorize past experiences, as a new possibility for cognitive theorizing. Whether human or animal minds actually operate with excess representational capacity, how capacity should be measured, and which empirical phenomena provide evidence for each regime remain central open questions that we hope will motivate future research and debate. By clarifying distinctions among capacity regimes, outlining computational and empirical approaches to diagnosing them, and identifying domains across cognitive, psychological, and neural sciences where these ideas may be applied, we hope that the framework supports systematic empirical progress in cognitive sciences.

Alongside these open questions, we also point to current limitations of the framework. Many reported properties of learning are not yet captured through the capacity analysis, particularly those involving temporal dynamics of memorizing, learning, and forgetting. For example, it has been observed that humans and animals may memorize individual experiences first and acquire generalizable patterns later (McClelland et al., 1995; Richards et al., 2014; Tompary & Davachi, 2017; Graves, Antony & Turk-Browne, 2020). Moreover, sleep plays a significant role in supporting successful learning and generalization: It has been argued that sleep is essential for acquiring generalizable patterns, and sleep has been shown to negate the perceptual deterioration discussed in Section 4.4. While the capacity framework considers only the temporal effects of acquisition of new experiences, other recent computer science results could apply to these phenomena. For example, *grokking* is a related phenomenon showing that the generalization performance of a model may improve with training time, even after the model has perfectly accounted for all the training data points (Nanda et al., 2023; Papyan et al., 2020; Power et al., 2022). These computational findings point to promising directions for extending the framework to account for temporal aspects of learning.

Beyond the broad directions for future research outlined in Section 5, we identify three additional directions for progress within the capacity framework. These include, but are not limited to: 1) clarifying and further identifying the learning properties associated with each capacity regime; 2) determining the capacity regimes underlying putatively distinct cognitive processes, such as long-term memory, visual perception, attention, and reinforcement learning;

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

and 3) investigating the realization and feasibility of excess capacity learning in specific biological and artificial systems. We discuss these directions in turn.

First, the empirical blueprint of learning with excess capacity outlined in Section 4 is based on the recent literature in computer science. Many aspects of the mechanisms driving and behaviors exhibited by excess capacity systems remain open research questions. We see further exploration of the empirical properties of learning with excess capacity, including the extent of generality of these properties, as a key step for progress in this paradigm. There remains much work to be done uncovering key properties of excess capacity learning, such as the dynamics of learning and forgetting, properties of the resulting representations, and these systems' ability to extrapolate beyond the boundaries of past experiences.

Second, our proposed taxonomy of the capacity regimes provides cognitive scientists with a theoretical framework useful for investigating and understanding specific cognitive processes, or specific stages of a cognitive process[13]. These processes include reinforcement learning, perceptual learning, and long-term memory. Identification of the capacity regime of a given cognitive process, based on empirical observations of the properties we outline in Section 4, can help scientists make other empirically testable predictions that reflect behavior in that capacity regime.

Third, future research would need to establish the connection between the behavioral signatures of excess capacity learning and the way such learning might be realized in specific biological or artificial systems, such as the human brain, robot, or a multi-agent system. Progress on this front can be accomplished by conducting thorough theoretical examinations of the feasibility of excess capacity learning in specific target contexts (Van Rooij, 2008) and by explicitly constructing biologically- and cognitively-plausible excess capacity models.

Cognitive science benefits from diverse perspectives on cognitive systems (Dale et al., 2022). Conceptualizing cognition as an excess, rather than constrained, capacity process has so far been

[13]Recent work in theoretical neuroscience suggests that compression and expansion might each be at play at different stages of information processing in the brain (Farrell et al., 2022; D. Zhou et al., 2022).

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

largely unexplored and at times even discouraged. Exploring excess capacity learning as a genuine possibility for many domains of cognitive science opens new directions of inquiry. Beyond introducing another mode of learning, it allows us to consider cognition as operating along a continuum of capacity regimes, with the cognitive system flexibly navigating between them. The capacity framework invites a reconceptualization of existing models and lays the groundwork for the development of new formal and empirical approaches to cognition that have yet to be imagined.

Acknowledgements. We thank Paul Bloom, Nick Chater, and three anonymous reviewers for their comments that helped improve this manuscript. We thank Rob Goldstone, Rob Nosofsky, Rich Shiffrin, David J. Halpern, Marisa Carrasco, Steven Sloman, Gary Lupyan, Chris Bates, Dileep George, Sebastian Musslick for their feedback on the ideas and the earlier versions of the manuscript.

Funding statement. The authors report no relevant funding.

Conflicts of Interest statement. The authors report no conflicts of interest.

Use of AI tools declaration. The authors used AI-based tools (OpenAI's ChatGPT) to assist with language editing and grammar during final stages of manuscript preparation. All intellectual content was created and verified by the authors who take full responsibility for the work.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

References

1. Arora, S., Cohen, N., & Hazan, E. (2018). On the optimization of deep networks: Implicit acceleration by overparameterization. *International Conference on Machine Learning*, 244–253. https://proceedings.mlr.press/v80/arora18a.htmlAshby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology*, *39*(2), 216–233.
2. Ashby, F. G., Alfonso-Reese, L. A., & Waldron, E. M. (1998). A neuropsychological theory of multiple systems in category learning. *Psychological Review*, *105*(3), 442.
3. Ashby, F. G., & Rosedahl, L. (2017). A neural interpretation of exemplar theory. *Psychological Review*, *124*(4), 472.
4. Balestriero, R., Pesenti, J., & LeCun, Y. (2021, October 18). *Learning in High Dimension Always Amounts to Extrapolation*. ArXiv.Org. https://arxiv.org/abs/2110.09485v2
5. Bartlett, P. L., & Kulkarni, S. R. (1998). The complexity of model classes, and smoothing noisy data. *Systems & Control Letters*, *34*(3), 133–140.
6. Bartlett, P. L., Long, P. M., Lugosi, G., & Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, *117*(48), 30063–30070. https://doi.org/10.1073/pnas.1907378117
7. Bartlett, P. L., Montanari, A., & Rakhlin, A. (2021). Deep learning: A statistical viewpoint. *Acta Numerica*, *30*, 87–201.
8. Bates, C. J., & Jacobs, R. A. (2020). Efficient data compression in perception and perceptual memory. *Psychological Review*, *127*(5), 891–917. https://doi.org/10.1037/rev0000197
9. Bauer, P. J. (2007). Recall in Infancy: A Neurodevelopmental Account. *Current Directions in Psychological Science*, *16*(3), 142–146. https://doi.org/10.1111/j.1467-8721.2007.00492.x
10. Belkin, M. (2021). Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, *30*, 203–248.
11. Belkin, M., Hsu, D. J., & Mitra, P. (2018). Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper/2018/hash/e22312179bf43e61576081a2f250f845-Abstract.html

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

12. Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849–15854.
13. Belkin, M., Hsu, D., & Xu, J. (2020). Two Models of Double Descent for Weak Features. *SIAM Journal on Mathematics of Data Science*, *2*(4), 1167–1180. https://doi.org/10.1137/20M1336072
14. Belkin, M., Ma, S., & Mandal, S. (2018). To understand deep learning we need to understand kernel learning. *International Conference on Machine Learning*, 541–549. http://proceedings.mlr.press/v80/belkin18a.html
15. Blanc, G., Gupta, N., Valiant, G., & Valiant, P. (2020). Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. *Conference on Learning Theory*, 483–513. http://proceedings.mlr.press/v125/blanc20a.html
16. Bowerman, M. (1982). Reorganizational processes in lexical and syntactic development. In *Language acquisition: The state of the art* (pp. 319–346). Academic Press. https://pure.mpg.de/rest/items/item_468073/component/file_468072/content
17. Bowman, C. R., Iwashita, T., & Zeithamova, D. (2020). Tracking prototype and exemplar representations in the brain across learning. *Elife*, *9*, e59360.
18. Bowman, C. R., & Zeithamova, D. (2020). Training set coherence and set size effects on concept generalization and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *46*(8), 1442.
19. Bowman, C. R., & Zeithamova, D. (2023). Coherent category training enhances generalization in prototype-based categories. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. https://psycnet.apa.org/record/2023-75189-001
20. Brady, T. F., Konkle, T., Alvarez, G. A., & Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, *105*(38), 14325–14329.
21. Brady, T., Schacter, D. L., & Alvarez, G. (2018). *The adaptive nature of false memories is revealed by gist-based distortion of true memories*. https://osf.io/preprints/psyarxiv/zeg95/
22. Braginsky, M. (2022). *Language learning at scale: Data-driven and model-motivated analyses of lexical and morphological development* [PhD Thesis, Massachusetts Institute of Technology]. https://dspace.mit.edu/handle/1721.1/143365

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

23. Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. *Cognition*, *118*(1), 2–16.
24. Bubeck, S., & Sellke, M. (2023). A universal law of robustness via isoperimetry. *Journal of the ACM*, *70*(2), 1–18.
25. Bunge, M. (1963). *The myth of simplicity; problems of scientific philosophy*. Englewood Cliffs, N.J., Prentice-Hall. http://archive.org/details/themythofsimplic0000unse
26. Bybee, J. L., & Slobin, D. I. (1982). Rules and schemas in the development and use of the English past tense. *Language*, *58*(2), 265–289.
27. Byrd, J., & Lipton, Z. (2019). What is the effect of importance weighting in deep learning? *International Conference on Machine Learning*, 872–881. https://proceedings.mlr.press/v97/byrd19a.html
28. Censor, N., Karni, A., & Sagi, D. (2006). A link between perceptual learning, adaptation and sleep. *Vision Research*, *46*(23), 4071–4074. https://doi.org/10.1016/j.visres.2006.07.022
29. Censor, N., & Sagi, D. (2008). Benefits of efficient consolidation: Short training enables long-term resistance to perceptual adaptation induced by intensive testing. *Vision Research*, *48*(7), 970–977.
30. Censor, N., & Sagi, D. (2009). Explaining training induced performance increments and decrements within a unified framework of perceptual learning. *Learning & Perception*, *1*(1), 3–17.
31. Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *The Quarterly Journal of Experimental Psychology: Section A*, *52*(2), 273–302.
32. Chater, N., & Vitányi, P. (2003). Simplicity: A unifying principle in cognitive science? *Trends in Cognitive Sciences*, *7*(1), 19–22. https://doi.org/10.1016/S1364-6613(02)00005-0
33. Chen, Z., Demetrio, L., Gupta, S., Feng, X., Xia, Z., Cinà, A. E., Pintor, M., Oneto, L., Demontis, A., & Biggio, B. (2024). Over-parameterization and Adversarial Robustness in Neural Networks: An Overview and Empirical Analysis. *CoRR*. https://openreview.net/forum?id=eq30pV6xg5
34. Chizat, L., Oyallon, E., & Bach, F. (2019). On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, *32*. https://proceedings.neurips.cc/paper/2019/hash/ae614c557843b1df326cb29c57225459-Abstract.html

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

35. Curth, A., Jeffares, A., & van der Schaar, M. (2024). A u-turn on double descent: Rethinking parameter counting in statistical learning. *Advances in Neural Information Processing Systems*, *36*. https://proceedings.neurips.cc/paper_files/paper/2023/hash/aec5e2847c5ae90f939ab786774856cc-Abstract-Conference.html
36. Dale, R., Warlaumont, A. S., & Johnson, K. L. (2022). The fundamental importance of method to theory. *Nature Reviews Psychology*, 1–12. https://doi.org/10.1038/s44159-022-00120-5
37. Dar, Y., & Baraniuk, R. G. (2022). Double double descent: On generalization errors in transfer learning between linear regression tasks. *SIAM Journal on Mathematics of Data Science*, *4*(4), 1447–1472.
38. Dar, Y., Muthukumar, V., & Baraniuk, R. G. (2021a). A farewell to the bias-variance tradeoff? An overview of the theory of overparameterized machine learning. *ArXiv Preprint ArXiv:2109.02355*.
39. Dar, Y., Muthukumar, V., & Baraniuk, R. G. (2021b). A farewell to the bias-variance tradeoff? An overview of the theory of overparameterized machine learning. *ArXiv Preprint ArXiv:2109.02355*.
40. Dayan, P., & Berridge, K. C. (2014). Model-based and model-free Pavlovian reward learning: Revaluation, revision, and revelation. *Cognitive, Affective, & Behavioral Neuroscience*, *14*(2), 473–492. https://doi.org/10.3758/s13415-014-0277-8
41. DeLosh, E. L., Busemeyer, J. R., & McDaniel, M. A. (1997). Extrapolation: The sine qua non for abstraction in function learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *23*(4), 968.
42. Doroudi, S., & Rastegar, S. A. (2023). The bias–variance tradeoff in cognitive science. *Cognitive Science*, *47*(1), e13241.
43. Du, S., Lee, J., Li, H., Wang, L., & Zhai, X. (2019). Gradient Descent Finds Global Minima of Deep Neural Networks. *Proceedings of the 36th International Conference on Machine Learning*, 1675–1685. https://proceedings.mlr.press/v97/du19c.html
44. Dubova, M., Chandramouli, S., Gigerenzer, G., Grünwald, P., Holmes, W., Lombrozo, T., Marelli, M., Musslick, S., Nicenboim, B., Ross, L. N., Shiffrin, R., White, M., Wagenmakers, E.-J., Bürkner, P.-C., & Sloman, S. J. (2025). Is Ockham's razor losing its

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

edge? New perspectives on the principle of model parsimony. *Proceedings of the National Academy of Sciences*, *122*(5), e2401230121. https://doi.org/10.1073/pnas.2401230121

45. Dubova, M., Galesic, M., & Goldstone, R. L. (2022). Cognitive Science of Augmented Intelligence. *Cognitive Science*, *46*(12), e13229.
46. Dubova, M., & Goldstone, R. L. (2021). The Influences of Category Learning on Perceptual Reconstructions. *Cognitive Science*, *45*(5), e12981.
47. Dwivedi, R., Singh, C., Yu, B., & Wainwright, M. (2023). Revisiting minimum description length complexity in overparameterized models. *Journal of Machine Learning Research*, *24*(268), 1–59.
48. Dziugaite, G. K., Drouin, A., Neal, B., Rajkumar, N., Caballero, E., Wang, L., Mitliagkas, I., & Roy, D. M. (2020). In search of robust measures of generalization. *Advances in Neural Information Processing Systems*, *33*, 11723–11733.
49. Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., & Sohl-Dickstein, J. (2018). Adversarial examples that fool both computer vision and time-limited humans. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper/2018/hash/8562ae5e286544710b2e7ebe9858833b-Abstract.html
50. Erickson, M. A., & Kruschke, J. K. (1998). Rules and exemplars in category learning. *Journal of Experimental Psychology: General*, *127*(2), 107.
51. Farrell, M., Recanatesi, S., Moore, T., Lajoie, G., & Shea-Brown, E. (2022). Gradient-based learning drives robust representations in recurrent neural networks by balancing compression and expansion. *Nature Machine Intelligence*, *4*(6), 564–573.
52. Farrell, M., Recanatesi, S., & Shea-Brown, E. (2023). From lazy to rich to exclusive task representations in neural networks and neural codes. *Current Opinion in Neurobiology*, *83*, 102780.
53. Feng, Y., & Tu, Y. (2021). The inverse variance–flatness relation in stochastic gradient descent is critical for finding flat minima. *Proceedings of the National Academy of Sciences*, *118*(9). https://doi.org/10.1073/pnas.2015617118
54. Forest, T. A., Schlichting, M. L., Duncan, K. D., & Finn, A. S. (2023). Changes in statistical learning across development. *Nature Reviews Psychology*, *2*(4), 205–219.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

55. França, G. S., Baron, M., King, B. R., Bossowski, J. P., Bjornberg, A., Pour, M., Rao, A., Patel, A. S., Misirlioglu, S., & Barkley, D. (2024). Cellular adaptation to cancer therapy along a resistance continuum. *Nature*, *631*(8022), 876–883.
56. Frank, M. J., Cohen, M. X., & Sanfey, A. G. (2009). Multiple Systems in Decision Making: A Neurocomputational Perspective. *Current Directions in Psychological Science*, *18*(2), 73–77. https://doi.org/10.1111/j.1467-8721.2009.01612.x
57. Frankland, S. M., Webb, T., & Cohen, J. D. (2021). *No coincidence, George: Capacity-limits as the Curse of Compositionality*. https://osf.io/preprints/psyarxiv/cjuxb/download/
58. Gamba, M., Englesson, E., Björkman, M., & Azizpour, H. (2022). Deep Double Descent via Smooth Interpolation. *ArXiv Preprint ArXiv:2209.10080*.
59. Geiger, M., Spigler, S., d'Ascoli, S., Sagun, L., Baity-Jesi, M., Biroli, G., & Wyart, M. (2019). The jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, *100*(1), 012115. https://doi.org/10.1103/PhysRevE.100.012115
60. Gershkoff-Stowe, L., & Thelen, E. (2004). U-Shaped Changes in Behavior: A Dynamic Systems Perspective. *Journal of Cognition and Development*, *5*(1), 11–36. https://doi.org/10.1207/s15327647jcd0501_2
61. Gershman, S. J., & Daw, N. D. (2017). Reinforcement learning and episodic memory in humans and animals: An integrative framework. *Annual Review of Psychology*, *68*, 101–128.
62. Gershman, S. J., Markman, A. B., & Otto, A. R. (2014). Retrospective revaluation in sequential decision making: A tale of two systems. *Journal of Experimental Psychology: General*, *143*(1), 182.
63. Gibson, E. J. (1969). *Principles of perceptual learning and development.* https://psycnet.apa.org/record/1969-35014-000
64. Gibson, J. J., & Gibson, E. J. (1955). Perceptual learning: Differentiation or enrichment? *Psychological Review*, *62*(1), 32–41. https://doi.org/10.1037/h0048826
65. Gigerenzer, G. (2008). Why Heuristics Work. *Perspectives on Psychological Science*, *3*(1), 20–29. https://doi.org/10.1111/j.1745-6916.2008.00058.x
66. Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*(1), 107–143.
67. Goldstone, R. L. (1994). Influences of categorization on perceptual discrimination. *Journal of Experimental Psychology: General*, *123*(2), 178–200. https://doi.org/10.1037/0096-3445.123.2.178

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

68. Gonzalez, C., & Dutt, V. (2011). Instance-based learning: Integrating sampling and repeated decisions from experience. *Psychological Review*, *118*(4), 523.
69. Gonzalez, C., Lerch, J. F., & Lebiere, C. (2003). Instance-based learning in dynamic decision making. *Cognitive Science*, *27*(4), 591–635.
70. Gorman, T. E., & Goldstone, R. L. (2022). An instance-based model account of the benefits of varied practice in visuomotor skill. *Cognitive Psychology*, *137*, 101491.
71. Graves, K. N., Antony, J. W., & Turk-Browne, N. B. (2020). Finding the pattern: On-line extraction of spatial structure during virtual navigation. *Psychological Science*, *31*(9), 1183-1190.
72. Griffiths, T. L., Canini, K. R., Sanborn, A. N., & Navarro, D. J. (2007). Unifying rational models of categorization via the hierarchical Dirichlet process. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *29*(29). https://escholarship.org/content/qt5p6505mf/qt5p6505mf.pdf
73. Gu, Y., Zheng, X., & Aste, T. (2024). Unraveling the Enigma of Double Descent: An In-depth Analysis through the Lens of Learned Feature Space. *The Twelfth International Conference on Learning Representations*. https://openreview.net/forum?id=CEkIyshNbC
74. Gunasekar, S., Lee, J. D., Soudry, D., & Srebro, N. (2018). Implicit bias of gradient descent on linear convolutional networks. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper/2018/hash/0e98aeeb54acf612b9eb4e48a269814c-Abstract.html
75. Hahn, U., & Chater, N. (1998). Similarity and rules: Distinct? exhaustive? empirically distinguishable? *Cognition*, *65*(2–3), 197–230.
76. Hampton, J. (1993). *Prototype models of concept representation.* https://psycnet.apa.org/record/1993-97009-003
77. Hansen, T., Olkkonen, M., Walter, S., & Gegenfurtner, K. R. (2006). Memory modulates color appearance. *Nature Neuroscience*, *9*(11), 1367–1368.
78. Hao, Y., Lin, Y., Zou, D., & Zhang, T. (2024). On the Benefits of Over-parameterization for Out-of-Distribution Generalization. *ArXiv Preprint ArXiv:2403.17592*. https://arxiv.org/abs/2403.17592
79. Hao, Y., & Zhang, T. (2024). The Surprising Harmfulness of Benign Overfitting for Adversarial Robustness. *ArXiv E-Prints*, arXiv-2401.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

80. Harnad, S. (2003). Categorical Perception. In *Encyclopedia of Cognitive Science*. Nature Publishing Group: Macmillan. https://eprints.soton.ac.uk/257719/
81. Hasson, U., Nastase, S. A., & Goldstein, A. (2020). Direct fit to nature: An evolutionary perspective on biological and artificial neural networks. *Neuron*, *105*(3), 416–434.
82. Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *The Annals of Statistics*, *50*(2). https://doi.org/10.1214/21-AOS2133
83. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.
84. Hendrycks, D., Basart, S., Mu, N., Kadavath, S., Wang, F., Dorundo, E., Desai, R., Zhu, T., Parajuli, S., Guo, M., Song, D., Steinhardt, J., & Gilmer, J. (2021). *The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization*. 8340–8349. https://openaccess.thecvf.com/content/ICCV2021/html/Hendrycks_The_Many_Faces_of_Robustness_A_Critical_Analysis_of_Out-of-Distribution_ICCV_2021_paper.html
85. Hendrycks, D., & Dietterich, T. (2019). *Benchmarking Neural Network Robustness to Common Corruptions and Perturbations*.
86. Hewson, J., Sloman, S. J., & Dubova, M. (2024). *One system for learning and remembering episodes and rules. Proceedings of the 7th annual conference on Cognitive Computational Neuroscience*. https://2024.ccneuro.org/pdf/316_Paper_authored_ECL_vs__CLS__CCN.pdf
87. Hodgkinson, L., van der Heide, C., Salomone, R., Roosta, F., & Mahoney, M. W. (2023). The Interpolating Information Criterion for Overparameterized Models. *ArXiv E-Prints*, arXiv-2307.
88. Homa, D., Blair, M., McClure, S. M., Medema, J., & Stone, G. (2019). Learning concepts when instances never repeat. *Memory & Cognition*, *47*(3), 395–411. https://doi.org/10.3758/s13421-018-0874-9
89. Hu, M., & Nosofsky, R. M. (2022). Exemplar-model account of categorization and recognition when training instances never repeat. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(12), 1947–1969. https://doi.org/10.1037/xlm0001008
90. Hu, M., & Nosofsky, R. M. (2024). High-variability training does not enhance generalization in the prototype-distortion paradigm. *Memory & Cognition*. https://doi.org/10.3758/s13421-023-01516-1

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

91. Hu, M. Y., Mueller, A., Ross, C., Williams, A., Linzen, T., Zhuang, C., Cotterell, R., Choshen, L., Warstadt, A., & Wilcox, E. G. (2024). Findings of the Second BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. *The 2nd BabyLM Challenge at the 28th Conference on Computational Natural Language Learning*, 1–21. https://aclanthology.org/2024.conll-babylm.1/
92. Huang, Y., Liang, Y., & Huang, L. (2022). Provable generalization of overparameterized meta-learning trained with sgd. *Advances in Neural Information Processing Systems*, *35*, 16563–16576.
93. Jacoby, L. L. (1983). Perceptual enhancement: Persistent effects of an experience. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *9*(1), 21.
94. Jacoby, L. L., & Brooks, L. R. (1984). Nonanalytic cognition: Memory, perception, and concept learning. In *Psychology of learning and motivation* (Vol. 18, pp. 1–47). Elsevier. https://www.sciencedirect.com/science/article/pii/S0079742108603588
95. Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2007). A tutorial on kernel methods for categorization. *Journal of Mathematical Psychology*, *51*(6), 343–358.
96. Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008). Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychonomic Bulletin & Review*, *15*(2), 256–271.
97. Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2009). Does cognitive science need kernels? *Trends in Cognitive Sciences*, *13*(9), 381–388.
98. Jastrzebski, S., Szymczak, M., Fort, S., Arpit, D., Tabor, J., Cho, K., & Geras, K. (2020). The Break-Even Point on Optimization Trajectories of Deep Neural Networks. *International Conference on Learning Representations*. https://openreview.net/forum?id=r1g87C4KwB
99. Ju, P., Lin, S., Squillante, M. S., Liang, Y., & Shroff, N. B. (2023). Generalization Performance of Transfer Learning: Overparameterized and Underparameterized Regimes. *ArXiv E-Prints*, arXiv-2306.
100. Juslin, P., Karlsson, L., & Olsson, H. (2008). Information integration in multiple cue judgment: A division of labor hypothesis. *Cognition*, *106*(1), 259–298.
101. Just, M. A., & Varma, S. (2007). The organization of thinking: What functional brain imaging reveals about the neuroarchitecture of complex cognition. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(3), 153–191.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

102. Kahneman, D. (2011). *Thinking, fast and slow*. macmillan. https://books.google.com/books?hl=en&lr=&id=SHvzzuCnuv8C&oi=fnd&pg=PP2&dq=kahneman+tversky+system+1&ots=NUtgNB5hCH&sig=Sk1qvNseeiX-HQFbuQf0yE_7cus

103. Kane, P. B., & Broomell, S. B. (2020). Applications of the bias–variance decomposition to human forecasting. *Journal of Mathematical Psychology*, *98*, 102417. https://doi.org/10.1016/j.jmp.2020.102417

104. Ketz, N., Morkonda, S. G., & O'Reilly, R. C. (2013). Theta coordinated error-driven learning in the hippocampus. *PLoS Computational Biology*, *9*(6), e1003067.

105. Knowlton, B. J., Mangels, J. A., & Squire, L. R. (1996). A neostriatal habit learning system in humans. *Science*, *273*(5280), 1399–1402.

106. Knowlton, B. J., Squire, L. R., & Gluck, M. A. (1994). Probabilistic classification learning in amnesia. *Learning & Memory*, *1*(2), 106–120.

107. Koffka K. (1935). *Principles Of Gestalt Psychology*. Harcourt, Brace and Company, Inc. http://archive.org/details/in.ernet.dli.2015.7888

108. Kolers, P. A. (1976). Reading a year later. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5), 554.

109. Koutstaal, W., & Schacter, D. L. (1997). Gist-based false recognition of pictures in older and younger adults. *Journal of Memory and Language*, *37*(4), 555–583.

110. Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446.

111. Kristjánsson, Á., & Campana, G. (2010). Where perception meets memory: A review of repetition priming in visual search tasks. *Attention, Perception, & Psychophysics*, *72*(1), 5–18. https://doi.org/10.3758/APP.72.1.5

112. Kuczaj II, S. A. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 589–600.

113. Kuhl, P. K., Williams, K. A., Lacerda, F., Stevens, K. N., & Lindblom, B. (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606–608.

114. 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.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

115. Lakatos, I. (1976). Falsification and the methodology of scientific research programmes. In *Can theories be refuted?* (pp. 205–259). Springer.

116. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*. https://doi.org/10.1017/S0140525X16001837

117. Lampinen, A. K., & Ganguli, S. (2018). An analytic theory of generalization dynamics and transfer learning in deep linear networks. *ArXiv Preprint ArXiv:1809.10374*.

118. Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211.

119. Laughlin, S. B. (2001). Energy as a constraint on the coding and processing of sensory information. *Current Opinion in Neurobiology*, *11*(4), 475–480.

120. Li, Y., Krefeld-Schwalb, A., Wall, D. G., Johnson, E. J., Toubia, O., & Bartels, D. M. (2022). The More You Ask, the Less You Get: When Additional Questions Hurt External Validity. *Journal of Marketing Research*, *59*(5), 963–982. https://doi.org/10.1177/00222437211073581

121. Li, Z., Wang, T., & Arora, S. (2021). What Happens after SGD Reaches Zero Loss?–A Mathematical Framework. *ArXiv E-Prints*, arXiv-2110.

122. Lieder, F., & Griffiths, T. L. (2020). Resource-rational analysis: Understanding human cognition as the optimal use of limited computational resources. *Behavioral and Brain Sciences*, *43*.

123. Liu, C. Y., & Flanigan, J. (2023). Understanding the Role of Optimization in Double Descent. *OPT 2023: Optimization for Machine Learning*. https://openreview.net/forum?id=8vggUWDeLF

124. Loftus, E. F., Coan, J. A., & Pickrell, J. E. (2014). Manufacturing false memories using bits of reality. In *Implicit memory and metacognition* (pp. 195–220). Psychology Press. https://www.taylorfrancis.com/chapters/edit/10.4324/9781315806136-8/manufacturing-false-memories-using-bits-reality-elizabeth-loftus-james-coan-jacqueline-pickrell

125. Loftus, E. F., & Pickrell, J. E. (1995). The Formation of False Memories. *Psychiatric Annals*, *25*(12), 720–725. https://doi.org/10.3928/0048-5713-19951201-07

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

126. Logan, G. D. (1988). Toward an instance theory of automatization. *Psychological Review*, *95*(4), 492.

127. Lombrozo, T. (2007). Simplicity and probability in causal explanation. *Cognitive Psychology*, *55*(3), 232–257. https://doi.org/10.1016/j.cogpsych.2006.09.006

128. Lupyan, G. (2013). The difficulties of executing simple algorithms: Why brains make mistakes computers don't. *Cognition*, *129*(3), 615–636. https://doi.org/10.1016/j.cognition.2013.08.015

129. Lyon, P., & Cheng, K. (2023). Basal cognition: Shifting the center of gravity (again). *Animal Cognition*, *26*(6), 1743–1750. https://doi.org/10.1007/s10071-023-01832-8

130. Mach, E. (1905). *Knowledge and Error: Sketches on the Psychology of Enquiry* (Vol. 3). LEIPZIG Verlag von Johann Ambrosius Barth. https://books.google.com/books?hl=en&lr=&id=tT3qCAAAQBAJ&oi=fnd&pg=PR11&ots=K_rdp7a_f3&sig=_8YsXhTbWs5HH2Cl-HAhaHRPzSw

131. MacWhinney, B. (1978). The acquisition of morphophonology. *Monographs of the Society for Research in Child Development*, 1–123.

132. Mallinar, N. R., Zane, A., Frei, S., & Yu, B. (2024). Minimum-Norm Interpolation Under Covariate Shift. *Forty-First International Conference on Machine Learning*. https://openreview.net/forum?id=Zw7TcnTmHj

133. Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., Xu, F., & Clahsen, H. (1992). Overregularization in language acquisition. *Monographs of the Society for Research in Child Development*, i–178.

134. McClelland, J. L., & Goddard, N. H. (1996). Considerations arising from a complementary learning systems perspective on hippocampus and neocortex. *Hippocampus*, *6*(6), 654–665. https://doi.org/10.1002/(sici)1098-1063(1996)6:6<654::aid-hipo8>3.0.co;2-g

135. 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.

136. Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*(3), 207.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

137. Mednick, S. C., Arman, A. C., & Boynton, G. M. (2005). The time course and specificity of perceptual deterioration. *Proceedings of the National Academy of Sciences*, *102*(10), 3881–3885. https://doi.org/10.1073/pnas.0407866102

138. 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.

139. Minda, J. P., Roark, C. L., Kalra, P., & Cruz, A. (2024). Single and multiple systems in categorization and category learning. *Nature Reviews Psychology*, *3*(8), 536–551.

140. Murphy, G. (2004). *The big book of concepts*. MIT press.

141. Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2021). Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, *2021*(12), 124003.

142. Nanda, N., Chan, L., Liberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. *ArXiv Preprint ArXiv:2301.05217*.

143. Newcombe, N. S., & Nguyen, K. V. (2024). *Developmental Sequences Constrain Models of the Mind*. PsyArXiv. https://files.osf.io/v1/resources/vhuwc/providers/osfstorage/65aed5f997381903b8396d1d?action=download&direct&version=1

144. Ngo, C. T., Benear, S. L., Popal, H., Olson, I. R., & Newcombe, N. S. (2021). Contingency of semantic generalization on episodic specificity varies across development. *Current Biology*, *31*(12), 2690–2697.

145. 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.

146. Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700.

147. Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. *Formal Approaches in Categorization*, 18–39.

148. Nosofsky, R. M., Palmeri, T. J., & McKinley, S. C. (1994). Rule-plus-exception model of classification learning. *Psychological Review*, *101*(1), 53.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

149. Nosofsky, R. M., & Zaki, S. R. (1998). Dissociations between categorization and recognition in amnesic and normal individuals: An exemplar-based interpretation. *Psychological Science*, *9*(4), 247–255.

150. Ofen, N., Moran, A., & Sagi, D. (2007). Effects of trial repetition in texture discrimination. *Vision Research*, *47*(8), 1094–1102. https://doi.org/10.1016/j.visres.2007.01.023

151. Olshausen, B. A., & Field, D. J. (2004). Sparse coding of sensory inputs. *Current Opinion in Neurobiology*, *14*(4), 481–487.

152. Oneto, L., Ridella, S., & Anguita, D. (2023). Do we really need a new theory to understand over-parameterization? *Neurocomputing*, *543*, 126227. https://doi.org/10.1016/j.neucom.2023.126227

153. O'Reilly, R. C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. MIT press. https://books.google.com/books?hl=en&lr=&id=gL5NEAAAQBAJ&oi=fnd&pg=PR19&dq=yuko+munakata&ots=9c8IFHeFI2&sig=tZa0oXm79_gWIJchwu4RMH3QNHk

154. O'Reilly, R. C., & Rudy, J. W. (2001). Conjunctive representations in learning and memory: Principles of cortical and hippocampal function. *Psychological Review*, *108*(2), 311.

155. Orvieto, A., Raj, A., Kersting, H., & Bach, F. (2023). Explicit regularization in overparametrized models via noise injection. *International Conference on Artificial Intelligence and Statistics*, 7265–7287. https://proceedings.mlr.press/v206/orvieto23a.html

156. Papyan, V., Han, X. Y., & Donoho, D. L. (2020). Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, *117*(40), 24652–24663. https://doi.org/10.1073/pnas.2015509117

157. Peterson, L., & Peterson, M. J. (1959). Short-term retention of individual verbal items. *Journal of Experimental Psychology*, *58*(3), 193.

158. Pinker, S. (2015). *Words and rules: The ingredients of language*. Basic Books. https://books.google.com/books?hl=en&lr=&id=2cEuBgAAQBAJ&oi=fnd&pg=PR9&dq=words+and+rules+pinker&ots=y4D-Knd3uw&sig=GQGK9DvYPl64qdBb0WzkzVHYjQM

159. Pomerantz, J. R., & Kubovy, M. (1986). *Theoretical approaches to perceptual organization: Simplicity and likelihood principles.*

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

160. Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, *77*(3p1), 353.

161. Posner, M. I., & Keele, S. W. (1970). Retention of abstract ideas. *Journal of Experimental Psychology*, *83*(2p1), 304.

162. Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets. *ArXiv Preprint ArXiv:2201.02177*.

163. Ribeiro, A. H., & Schön, T. B. (2023). Overparameterized Linear Regression under Adversarial Attacks. *IEEE Transactions on Signal Processing*, *71*, 601–614. https://doi.org/10.1109/TSP.2023.3246228

164. Richards, B. A., Xia, F., Santoro, A., Husse, J., Woodin, M. A., Josselyn, S. A., & Frankland, P. W. (2014). Patterns across multiple memories are identified over time. *Nature neuroscience*, *17*(7), 981-986.

165. Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803.

166. Rumelhart, D. E., & McClelland, J. L. (1985). *On learning the past tenses of English verbs.* https://apps.dtic.mil/sti/html/tr/ADA164233/

167. Russell, S. J., & Subramanian, D. (1994). Provably bounded-optimal agents. *Journal of Artificial Intelligence Research*, *2*, 575–609.

168. Sanfey, A. G., & Chang, L. J. (2008). *Multiple Systems in Decision Making*. *Annals of the New York Academy of Sciences*, *1128*(1), 53–62. https://doi.org/10.1196/annals.1399.007

169. Schacter, D. L. (1987). Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*(3), 501.

170. Schapiro, A. C., Turk-Browne, N. B., Botvinick, M. M., & Norman, K. A. (2017). Complementary learning systems within the hippocampus: A neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160049. https://doi.org/10.1098/rstb.2016.0049

171. Schlichting, M. L., Mumford, J. A., & Preston, A. R. (2015). Learning-related

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

representational changes reveal dissociable integration and separation signatures in the hippocampus and prefrontal cortex. *Nature Communications*, *6*(1), 8151.

172. Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., & Geiger, F. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007.
173. Schulz, E. (2017). *Towards a unifying theory of generalization* [PhD Thesis]. UCL (University College London).
174. Schulz, E., Bhui, R., Love, B. C., Brier, B., Todd, M. T., & Gershman, S. J. (2019). Structured, uncertainty-driven exploration in real-world consumer choice. *Proceedings of the National Academy of Sciences*, *116*(28), 13903–13908.
175. Shepard, R. N. (1967). Recognition memory for words, sentences, and pictures. *Journal of Verbal Learning and Verbal Behavior*, *6*(1), 156–163.
176. Sherry, D. F., & Schacter, D. L. (1987). The evolution of multiple memory systems. *Psychological Review*, *94*(4), 439.
177. Sherman, B. E., & Turk-Browne, N. B. (2020). Statistical prediction of the future impairs episodic encoding of the present. *Proceedings of the National Academy of Sciences*, *117*(37), 22760-22770.
178. Shiffrin, R. M., & Nosofsky, R. M. (1994). *Seven plus or minus two: A commentary on capacity limitations.* https://psycnet.apa.org/journals/rev/101/2/357/
179. Shomar, A., Barak, O., & Brenner, N. (2022). Cancer progression as a learning process. *IScience*, *25*(3), 103924. https://doi.org/10.1016/j.isci.2022.103924
180. Simon, H. A. (1991). Bounded Rationality and Organizational Learning. *Organization Science*, *2*(1), 125–134. https://doi.org/10.1287/orsc.2.1.125
181. Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652–656.
182. Simsek, B., Ged, F., Jacot, A., Spadaro, F., Hongler, C., Gerstner, W., & Brea, J. (2021). Geometry of the loss landscape in overparameterized neural networks: Symmetries and invariances. *International Conference on Machine Learning*, 9722–9732.
183. Smith, J. D., & Minda, J. P. (1998). Prototypes in the mist: The early epochs of category learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *24*(6), 1411.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

184. Smith, L. B., Jayaraman, S., Clerkin, E., & Yu, C. (2018). The Developing Infant Creates a Curriculum for Statistical Learning. *Trends in Cognitive Sciences*, *22*(4), 325–336. https://doi.org/10.1016/j.tics.2018.02.004

185. Somepalli, G., Fowl, L., Bansal, A., Yeh-Chiang, P., Dar, Y., Baraniuk, R., Goldblum, M., & Goldstein, T. (2022). Can neural nets learn the same model twice? Investigating reproducibility and double descent from the decision boundary perspective. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13699–13708.

186. Spanne, A., & Jörntell, H. (2015). Questioning the role of sparse coding in the brain. *Trends in Neurosciences*, *38*(7), 417–427.

187. Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., & Wyart, M. (2019). A jamming transition from under- to over-parametrization affects generalization in deep learning. *Journal of Physics A: Mathematical and Theoretical*, *52*(47), 474001. https://doi.org/10.1088/1751-8121/ab4c8b

188. Stanford, K., (2023). Underdetermination of Scientific Theory. *The Stanford Encyclopedia of Philosophy* (Summer 2023 Edition), Edward N. Zalta & Uri Nodelman (eds.)

189. Squire, L. R., & Knowlton, B. J. (1995). Learning about categories in the absence of memory. *Proceedings of the National Academy of Sciences*, *92*(26), 12470–12474.

190. Sučević, J., & Schapiro, A. C. (2023). A neural network model of hippocampal contributions to category learning. *Elife*, *12*, e77185.

191. Sui, P., Duede, E., Wu, S., & So, R. (2024). Confabulation: The Surprising Value of Large Language Model Hallucinations. *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 14274–14284. https://aclanthology.org/2024.luhme-long.770/

192. Sutton, R. S., & Barto, A. G. (1998). *Reinforcement learning: An introduction* (Vol. 1). MIT press Cambridge. https://www.cambridge.org/core/journals/robotica/article/robot-learning-edited-by-jonathan-h-connell-and-sridhar-mahadevan-kluwer-boston-19931997-xii240-pp-isbn-0792393651-hardback-21800-guilders-12000-8995/737FD21CA908246DF17779E9C20B6DF6

193. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *2nd International Conference on*

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

Learning Representations, ICLR 2014. https://nyuscholars.nyu.edu/en/publications/intriguing-properties-of-neural-networks

194. Tenpenny, P. L. (1995). Abstractionist versus episodic theories of repetition priming and word identification. *Psychonomic Bulletin & Review, 2*(3), 339–363. https://doi.org/10.3758/BF03210972
195. Thelen, E., & Smith, L. B. (1994). *A dynamic systems approach to the development of cognition and action*. MIT press.
196. Todd, P. M., & Gigerenzer, G. (2012). *Ecological rationality: The normative study of heuristics.*
197. Tompary, A., & Davachi, L. (2017). Consolidation promotes the emergence of representational overlap in the hippocampus and medial prefrontal cortex. *Neuron*, *96*(1), 228-241.
198. Tompary, A., & Thompson-Schill, S. L. (2021). Semantic influences on episodic memory distortions. *Journal of Experimental Psychology: General*, *150*(9), 1800.
199. Tripuraneni, N., Adlam, B., & Pennington, J. (2021). Overparameterization improves robustness to covariate shift in high dimensions. *Advances in Neural Information Processing Systems*, *34*, 13883–13897.
200. Van Rooij, I. (2008). The Tractable Cognition Thesis. *Cognitive Science*, *32*(6), 939–984. https://doi.org/10.1080/03640210801897856
201. Veerabadran, V., Goldman, J., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., Shlens, J., Sohl-Dickstein, J., Mozer, M. C., & Elsayed, G. F. (2023). Subtle adversarial image manipulations influence both human and machine perception. *Nature Communications*, *14*(1), 4933. https://doi.org/10.1038/s41467-023-40499-0
202. Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., & Linzen, T. (2023). Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora. *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*. https://cir.nii.ac.jp/crid/1360022501345154560
203. Waugh, N. C., & Norman, D. A. (1965). Primary memory. *Psychological Review*, *72*(2), 89.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

204. Wichmann, F. A., & Geirhos, R. (2023). Are Deep Neural Networks Adequate Behavioral Models of Human Visual Perception? *Annual Review of Vision Science*, *9*(1), 501–524. https://doi.org/10.1146/annurev-vision-120522-031739

205. Wilson, A. G. (2025). Deep Learning is Not So Mysterious or Different. *ArXiv E-Prints*, arXiv-2503.

206. Wilson, A. G., Dann, C., Lucas, C., & Xing, E. P. (2015). The human kernel. *Advances in Neural Information Processing Systems*, *28*.

207. Wu, B., Chen, J., Cai, D., He, X., & Gu, Q. (2021). Do Wider Neural Networks Really Help Adversarial Robustness? *Advances in Neural Information Processing Systems*, *34*, 7054–7067. https://proceedings.neurips.cc/paper/2021/hash/3937230de3c8041e4da6ac3246a888e8-Abstract.html

208. Wu, C. M., Schulz, E., Speekenbrink, M., Nelson, J. D., & Meder, B. (2018). Generalization guides human exploration in vast decision spaces. *Nature Human Behaviour*, *2*(12), 915–924.

209. Wu, L., & Ma, C. (2018). How sgd selects the global minima in over-parameterized learning: A dynamical stability perspective. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper/2018/hash/6651526b6fb8f29a00507de6a49ce30f-Abstract.html

210. Wu, L., Wang, M., & Su, W. (2022). The alignment property of SGD noise and how it helps select flat minima: A stability analysis. *Advances in Neural Information Processing Systems*, *35*, 4680–4693.

211. Wyner, A. J., Olson, M., Bleich, J., & Mease, D. (2017). Explaining the success of adaboost and random forests as interpolating classifiers. *Journal of Machine Learning Research*, *18*(48), 1–33.

212. Xu, Z., Jain, S., & Kankanhalli, M. S. (2024). Hallucination is Inevitable: An Innate Limitation of Large Language Models. *CoRR*. https://openreview.net/forum?id=Fyh9EpsmSE

213. Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365.

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press

214. Yang, Z., Yu, Y., You, C., Steinhardt, J., & Ma, Y. (2020). Rethinking bias-variance trade-off for generalization of neural networks. *International Conference on Machine Learning*, 10767–10777.

215. Zeng, T., Tompary, A., Schapiro, A. C., & Thompson-Schill, S. L. (2021). Tracking the relation between gist and item memory over the course of long-term memory consolidation. *Elife*, *10*, e65588.

216. Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2021). Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, *64*(3), 107–115. https://doi.org/10.1145/3446776

217. Zheng, Y., Liu, X. L., Nishiyama, S., Ranganath, C., & O'Reilly, R. C. (2022). Correcting the hebbian mistake: Toward a fully error-driven hippocampus. *PLoS Computational Biology*, *18*(10), e1010589.

218. Zhou, D., Kim, J. Z., Pines, A. R., Sydnor, V. J., Roalf, D. R., Detre, J. A., Gur, R. C., Gur, R. E., Satterthwaite, T. D., & Bassett, D. S. (2022). Compression supports low-dimensional representations of behavior across neural circuits. *BioRxiv*, 2022–11.

219. Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications*, *10*(1), 1334.

220. Zhu, J., Gibson, B., & Rogers, T. T. (2009). Human rademacher complexity. *Advances in Neural Information Processing Systems*, *22*.

221. Zhu, Z., Wu, J., Yu, B., Wu, L., & Ma, J. (2019). The Anisotropic Noise in Stochastic Gradient Descent: Its Behavior of Escaping from Sharp Minima and Regularization Effects. *International Conference on Machine Learning*, 7654–7663. http://proceedings.mlr.press/v97/zhu19e

https://doi.org/10.1017/S0140525X2610510X Published online by Cambridge University Press