A. Mathematical details: Measures of model expressivity and capacity

We here provide a non-exhaustive list of measures from the computer science literature of the complexity of a model class. We refer to measures that don't depend on a given set of training data points—i.e., measures that reflect the number of patterns that can be expressed by a model class—as *measures of expressivity*. We refer to measures that reflect the complexity of or variance in the versions of a model fit to a particular set of training data points as *measures of capacity*.

A.1 Measures of expressivity

Number of parameters. Usually, a model with many parameters (components of the specification of the model whose specific values are influenced by training data) can express more patterns than a model with few parameters. For example, a degree-twenty polynomial regression model (Figure 3b) can exhibit a wide variety of shapes, while a degree-one polynomial regression model (Figure 3a) can only exhibit straight lines. Typically, each possible value the additional parameters can take is associated with a unique expression of the model (although see additional discussion below on the nuances of computing the *effective* number of parameters). In many cases, simply counting the number of a model's parameters can provide information about its expressivity (Hastie & Tibshirani, 2009; Belkin et al., 2019; Dubova et al., 2025).

VC dimension. The Vapnik-Chervonenkis (VC) dimension of a classifier is the maximum number of randomly-assigned data points the classifier could provide a perfect fit to (V. N. Vapnik & Chervonenkis, 1971). For example, the VC dimension of a binary linear classifier of data points in a two-dimensional space is 3: Given any three data points in a two-dimensional space, there is always a line that can separate the inputs that share a given label from the inputs that share the other label. This is not the case for four data points. For example, consider the case where the points [(0,1), (0,-1)] have the label "+" and the points [(1,0), (-1,0)] have the label "-"; no linear classifier can correctly separate the inputs labeled "+" from the inputs labeled "-".

Effective model complexity. The notion of effective model complexity (EMC) was proposed in the double descent literature to extend the basic idea of the VC dimension—a guarantee of the number of arbitrary data points a model can express—to an entire training procedure, i.e., to depend, in addition to the model class, on aspects of training like the optimization algorithm used. The EMC of a training procedure is the maximum number of randomly-labeled inputs[1] on which a training procedure results in essentially zero

[1]This is in contrast to the VC dimension, which considers the maximum number of inputs on which a classifier can achieve zero training error *regardless* of the labels. The EMC instead considers the average error across inputs and labelings drawn from a pre-specified distribution.

training error[2] (Nakkiran et al., 2021). Using the EMC as their measure of expressivity, Nakirran et al. (2021) demonstrate systematic presence of the double descent behavior described in Section 3.2.

A.2 Measures of capacity

Overparametrization ratio. The overparametrization ratio refers to the ratio of the number of parameters (see above) to the number of data points (Hastie et al., 2022. This measure "normalizes" the model's expressivity (as measured by the parameter count) by the number of data points that constrain the solutions it selects.

Rademacher complexity. The empirical error of a model measures the size of the errors it continues to commit on its training data even after being fit to those data. The generalization error of a model measures the size of the errors it commits on an entire population of data, most of which it hasn't encountered during training. Usually, the empirical error is an optimistic estimate of the generalization error, in the sense that the model will tend to commit smaller errors on data it has encountered during training than on data it hasn't. The degree of this optimism—the difference between the empirical and generalization errors—is known as the generalization gap.

The Rademacher complexity of a model is an estimate of the generalization gap achieved by the model, based on data points from a particular training sample (Shalev-Shwartz & Ben-David, 2014). Although this is superficially unrelated to the VC dimension (which in effect measures the size of the set of data points at which a model exhausts its excess capacity), the two measures are in fact closely connected. Informally, both can be connected to the number of distinct expressions (possibly learned patterns) a model has for a given number of data points (its "growth function"; Shalev-Shwartz & Ben-David, 2014; these connections are established by results known as the Massart lemma and Sauer's lemma).

Information criteria. The connection between the generalization gap and the flexibility of the model's expressions underlies a body of literature developing information criteria that capture both a model's empirical error and degree of freedom. As mentioned above, each additional free parameter included in a model generally increases the model's degree of freedom. This motivates the development of information criteria that use the number of a model's parameters to compute its degree of freedom, such as the Bayesian information criterion (Schwarz, 1978) and Akaike information criterion (Akaike, 1973).

Effective number of parameters. In nonlinear models, the number of degrees of freedom is generally less than the "parameter count" of the model: Each additional parameter interacts with the others in sometimes counterintuitive ways, and these interactions constrain the space of patterns the model can

[2]More specifically, the EMC captures the number of training data points under which the training error does not exceed some user-specified error tolerance ε.

exhibit. The effective number of parameters refers to the number of functionally independent parameters in a model. There are a variety of approaches to estimating a model's effective number of parameters (Moody, 1991; Spiegelhalter et al., 2002), which generally capture the size of the "influence" of the training data points on the learned patterns.[3] These have been leveraged in the construction of additional information criteria (Spiegelhalter et al., 2002; Van Der Linde, 2012).

Effective number of parameters of a smoother. A smoother is any model whose prediction at a given test input is a weighted average of the outcomes it's observed. Typically, smoothers assign higher weights to inputs that are more similar to the test input; smoothers that consider only very similar inputs produce less stable predictions in the sense that their predictions are highly dependent on the values observed at these specific inputs. Smoothers are particularly relevant in the context of theories of overparameterized machine learning: Overparameterized models exhibit properties of, and in many cases can be reduced to, smoothers (Belkin et al., 2018; Curth et al., 2024; Jacot et al., 2018; see also discussion in Section 4.3). The effective number of parameters of a smoother is a function of the norm of the vector of weights used to aggregate observed outcomes into a given prediction (Curth et al., 2024; Hastie & Tibshirani, 1986), reflecting the connection between how the smoother distributes these weights across data points and the stability of the predictions it produces. Like measures of the effective number of parameters more generally, these measures are typically a function of the training data and designed to directly capture the stability of outcomes predicted by a model trained on those data (Curth et al., 2024; Hastie et al., 2009). Intriguingly, some work has suggested that this measure is fundamental to understanding the generalization of overparameterized models, in the sense that measuring a model's expressivity as the effective number of parameters of the implied smoother eliminates the apparent double descent of generalization error (Curth et al., 2024).[4]

Description length. An alternative paradigm construes the fitted model as a representation of its training data, and the capacity of the model as the size of that representation. For example, the complexity with which the implied models in Figure 1a represent the same image can be measured in terms of the number of pixels, or number of bits, of the resulting representation. Measures of description length (Rissanen, 1978) and Kolmogorov complexity measure the size of the model's representation of the data (using

[3]The methods in the cited papers generalize the standard influence matrix in the context of linear models.

[4]The concept of a weight vector also appears in definitions of support vector machines (SVMs): The decision boundary of an SVM is a hyperplane in the space of possible inputs, and can be represented by a weight vector that linearly combines each of the inputs' dimensions. In certain settings, the VC dimension of an SVM is also related to the norm of its weight vectors (Schölkopf & Smola, 2002; V. Vapnik, 2006), and the expressivity of the SVM can be controlled by bounding the weight vector's norm. This result has been leveraged to establish connections between VC theory and the double descent phenomenon (Cherkassky & Lee, 2024).

concepts such as the size of the computer program that would be required to reconstruct its data representation). Such measures are the basis of principles like that of minimum description length, the principle of preferring models with shorter associated description length (Grünwald & Roos, 2019; Rissanen, 1978; Shalev-Shwartz & Ben-David, 2014).

B. Identifying the capacity of extant cognitive models

Here we describe the specific models and evidence that we used to identify the capacity regime of the extant cognitive models in Figure 2.

B.1 Constrained capacity

Distributional Semantics Models (DSMs). Human semantic learning and memory have been successfully modeled by a large class of DSMs as the process of abstracting word co-occurrence statistics onto a semantic space (Jones et al., 2015; Kumar, 2021; Landauer & Dumais, 1997). According to DSMs, each word is represented as a vector in the corresponding semantic space. These models learn vector representations for each word by trying to predict the word's neighbors in the contexts in which it occurred. DSMs have successfully captured semantic relationships between words, such as identifying synonyms based on the distance between words' vectors in the latent semantic space. There are numerous methods for learning distributional semantic representations, with Latent Semantic Analysis (Landauer & Dumais, 1997) and Word2Vec (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013) among the most recognized ones. The diverse approaches all rely on the key principle that the meaning of a word can be captured in a space of lower dimension than the number of specific contexts in which the word occurred. DSMs are essentially *dimensionality reduction* methods trained on vast linguistic corpora; therefore, they never fully memorize their data (Landauer & Dumais, 1997; Mikolov, Chen, et al., 2013). Jones (2019) compares DSMs to prototype models of categorization, saying that "virtually all DSMs are prototype models in that they create a single abstract representation of a word's meaning" aggregated across all contexts. Since DSMs never memorize their data (internally representing only lower-dimensional statistics of those data), we place them within the constrained capacity regime. Below, among the "Instance-based Models", we present a single exception to this: an instance-based DSM model which learns with sufficient (constant) capacity (Jamieson et al., 2018).

Compression & Autoencoder Models. Processes including perceptual learning and memory have been conceptualized as an optimal compression process over a channel with limited information-processing capacity. The system's limited cognitive resources form a bottleneck, preventing it from fully reconstructing its past experiences. The system learns to optimally allocate these resources to capture as many useful regularities about the environment as possible. Rate distortion theory provides optimal bounds on the amount of information that can be captured despite a bottleneck of a given information-processing capacity, and has been extensively used to model information retention in cognitive systems (Bates & Jacobs, 2020; Imel & Zaslavsky, 2024; Sims, 2016, 2018). Neural compression models, such as

neural autoencoders, have been used as approximate instantiations of the optimal compression processes described by rate distortion theory, and have been used to model human responses to real visual stimuli. Autoencoders learn to represent ("encode") their experiences by trying to reconstruct ("decode") them through a "bottleneck", a network layer of insufficient width to reproduce all the details of the network's inputs (for example, a 100x100 pixel image might be represented as the activation of just a few nodes). Neural autoencoder models have been especially successful as models of perceptual memory (Bates & Jacobs, 2020). Like the DSMs discussed above, autoencoders constitute a dimensionality reduction technique applied as a model of human cognition. Since the fundamental assumption of a limited capacity bottleneck is the central motivation for rate distortion theory and autoencoder modeling, we place these models in the constrained capacity regime.

Prototype Models. Prototype models have been used to capture the process of learning and using categories, such as "birds" and "healthy food items". The key idea is that each member of a category is represented economically, using one representation per all the items in the class (Hampton, 1993; Rosch, 1988). For example, each member of the category of "birds" is represented by the same set of average properties (e.g., has wings, can fly, is 10 cm tall). Prototype models posit that, when categorizing a new creature, systems assess the new creature's similarity to the prototypes of all previously-encountered categories. The key idea underlying the prototype model is that a variety of examples of a category are summarized by the system's representation of a single, prototypical example. Details of the examples are explicitly omitted in the prototype representation, which is why we assigned the prototype models to the constrained capacity regime.

Decision Heuristics. There is a huge body of work on the simple and fast heuristics that humans and animals use to make decisions in an uncertain world with scarce data. Many of these heuristics involve agents ignoring most available situational information to make a decision (G. Gigerenzer & Brighton, 2009; G. E. Gigerenzer et al., 2011; G. Gigerenzer & Goldstein, 1996). For example, systems operating according to the "take-the-best" heuristic base their decisions on a single informative feature of the situations they encounter. As an example of application of the take-the-best heuristic, a peahen chooses a mate with the highest number of eyespots on the tail, instead of assessing and integrating all the features to find the mate with the highest expected value (Petrie & Halliday, 1994). Other heuristics include recognition (choosing an option that is recognized) and tallying (counting the number of positive cues to make a decision).

Application of most decision heuristics assumes a pre-specified set of cues, or representations. In contrast, our focus is on the ways in which cognitive systems construct such representations in the first place. While not directly capturing the learning process that is the key target of our paper, the use of heuristics is often motivated by the postulation of cognitive limitations, such as memory and attention constraints, that would make it impossible for cognitive systems to remember the details of all past

experiences (Simon, 1990). Moreover, the ecological success of such heuristics (as compared to systems that base their decisions on extensive details of their past experiences) has been used as evidence for constraints on representational resources (viz., the "adaptivity" of forgetting; Gigerenzer & Todd, 1999; Schooler & Hertwig, 2005). Since heuristics *rely on* and are often used to *justify* simplified representations of situations, we broadly map them onto the constrained capacity regime.

Large Language Models (LLMs). Detecting the amount of training data production LLMs can memorize is an active area of research (due to privacy considerations, LLMs do not usually output exact sequences from their training data). Given that many modern LLMs and their training data are not publicly available, this research instead relies on publicly available models (e.g., GPT-Neo) and often indirect assessments of memorization. For example, one approach is to uniformly sample data points from datasets known to be used to train LLMs (e.g., Pile 825GB, a publicly-available dataset comprised of text collected from various sources), and then prompt the model with the sampled data points. If subsequent data points are reproduced by the model, the model is considered to have memorized the training input. Because of the inclusion of safeguards that prevent the models from outputting the training data verbatim, these indirect methods can provide only approximate lower bounds on the degree to which LLMs memorize. Current estimates based on a variety of LLMs of different sizes suggest that these models memorize somewhere between 0.2% and 1.5% of their training data. Fine-tuning the models can raise the percentage of memorized data, sometimes up to 16% (Carlini et al., 2022; Nasr et al., 2025). Given the lack of evidence for memorization of the full training dataset in the LLM studies so far, we broadly classify existing production LLMs as constrained capacity (Hoffmann et al., 2022; Sardana et al., 2024).

B.2 Excess capacity

*Convolutional Neural Networks (CNNs)***.** CNNs have recently served as a successful model of visual information processing in the brain. We used evidence from Zhang et al. (2016), who assessed the ability of the popular convolutional architectures (e.g., Inception, Alexnet) to fit the CIFAR10 and ImageNet ILSVRC 2012 datasets with randomly reshuffled labels (outputs) or pixels (inputs). The authors found that the CNNs were able to perfectly memorize even these sets of perturbed data points; when trained on the original datasets, these networks fully memorized those but still achieved high generalization scores. Perfect memorization and high generalization suggest these models belong to the excess capacity regime (as also concluded by Zhang et al., 2016). More recent, pretrained versions of these CNNs have been used in cognitive neuroscience as cognitive models of object recognition (Schrimpf et al., 2018).

B.3. Constant capacity

Exemplar Models. In the context of categorization, exemplar models suggest that cognitive systems store each encountered instance of a category in memory as a so-called exemplar, and represent new experiences as a weighted average of potentially all stored exemplars (Medin & Schaffer, 1978; Nosofsky, 1988, 2011; see also Box 1 and Section 6.12). In exemplar models, the representational resources the

system dedicates to storage of a given set of *n* exemplars are mathematically represented as the coordinates of *n* points in a multidimensional space, where each dimension represents a feature in terms of which the exemplars vary. When encountering a new experience, the system's representation is influenced by all exemplars stored in memory, and so the potential complexity of the representation scales with the number of stored exemplars. Exemplar models exhibit both constant capacity (the amount of representational resources they employ increases as they acquire more experiences) and sufficient capacity (given a set of *n* exemplars, they typically construct representations using the coordinates of exactly *n* points in the multidimensional feature space).

Similarly to exemplar models, multiple-trace models of memory also exhibit constant sufficient capacity: These models work by adding new so-called traces for each experience. Like exemplar models, these models posit that human cognition is unlimited in terms of the amount of information that can be encoded (Hintzman, 1984).

Gaussian Process Models (GPs). Closely related to exemplar models are Gaussian Process (GP) models (Ashby & Alfonso-Reese, 1995; Jäkel et al., 2007, 2008, 2009; see also Box 1 and Section 6.12). In the context of function learning and search tasks, GP models of human cognition posit that people map past experiences into an internal function space whose complexity is determined by the amount of their experiences: As with exemplar models, outcomes in new situations are predicted to covary with old experiences in structured, systematic ways (Schulz, 2017; Schulz et al., 2019; Wilson et al., 2015; C. M. Wu et al., 2018). GPs have been used to model human generalization across tasks (Schulz, 2017), notably including reinforcement learning (Gershman & Daw, 2017; Schulz et al., 2019; Wu et al., 2018). Artificial neural networks have been shown to be analogous to the Gaussian Processes in the infinite width (Jacot et al., 2018; Lee et al., 2017). Like exemplar models, GP models and the cognitive systems they represent exhibit constant capacity: They employ more representational resources as they acquire more experiences.

ALCOVE (Attention Learning Covering Map) (Kruschke, 1992) is a constant capacity model originally proposed as a connectionist alternative to the exemplar model of category learning (see above). ALCOVE has a hidden "exemplar" layer which maps each individual experience to a separate unit. As new exemplars (experiences) are encountered, ALCOVE expands to incorporate new hidden units to encode each new exemplar. In other words, the model's expressivity scales linearly with the number of training examples. When the network has to categorize a new exemplar, it integrates activations across the exemplar layer using the principles of the exemplar model. Because of the fixed exemplar-to-unit mapping (the representational resource of memory grows in strict proportion to training exposure), ALCOVE exhibits constant capacity regime.

References

1. Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, 267–281. https://hero.epa.gov/hero/index.cfm/reference/details/reference_id/591
2. 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
3. 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
4. Carlini, N., Ippolito, D., Jagielski, M., Lee, K., Tramer, F., & Zhang, C. (2022). Quantifying memorization across neural language models. *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=TatRHT_1cK
5. Cherkassky, V., & Lee, E. H. (2024). To understand double descent, we need to understand VC theory. *Neural Networks*, *169*, 242–256.
6. 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
7. 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.
8. Gigerenzer, G., & Brighton, H. (2009). Homo Heuristicus: Why Biased Minds Make Better Inferences. *Topics in Cognitive Science*, *1*(1), 107–143. https://doi.org/10.1111/j.1756-8765.2008.01006.x
9. Gigerenzer, G. E., Hertwig, R. E., & Pachur, T. E. (2011). *Heuristics: The foundations of adaptive behavior*. Oxford University Press. https://psycnet.apa.org/record/2011-10624-000
10. Gigerenzer, G., & Goldstein, D. G. (1996). Reasoning the fast and frugal way: Models of bounded rationality. *Psychological Review*, *103*(4), 650.
11. Gigerenzer, G., & Todd, P. M. (1999). *Simple heuristics that make us smart*. Oxford University Press, USA.
12. Grünwald, P., & Roos, T. (2019). Minimum description length revisited. *International Journal of Mathematics for Industry*, *11*(01), 1930001. https://doi.org/10.1142/S2661335219300018
13. Hampton, J. (1993). *Prototype models of concept representation*. https://psycnet.apa.org/record/1993-97009-003
14. 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

15. Hastie, T., & Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, *1*(3), 297–310. https://doi.org/10.1214/ss/1177013604
16. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Vol. 2). Springer.
17. Hintzman, D. L. (1984). MINERVA 2: A simulation model of human memory. *Behavior Research Methods, Instruments, & Computers*, *16*(2), 96–101.
18. Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., de Las Casas, D., Hendricks, L. A., Welbl, J., & Clark, A. (2022). Training compute-optimal large language models. *Proceedings of the 36th International Conference on Neural Information Processing Systems*, 30016–30030. https://dl.acm.org/doi/abs/10.5555/3600270.3602446
19. Imel, N., & Zaslavsky, N. (2024). Optimal compression in human concept learning. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *46*. https://escholarship.org/uc/item/7pc1g61d
20. Jacot, A., Gabriel, F., & Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, *31*.
21. Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior*, *1*, 119–136.
22. Jones, M. N. (2019). When does abstraction occur in semantic memory: Insights from distributional models. *Language, Cognition and Neuroscience*, *34*(10), 1338–1346. https://doi.org/10.1080/23273798.2018.1431679
23. Jones, M. N., Willits, J., & Dennis, S. (2015). 11 Models of Semantic Memory. *The Oxford Handbook of Computational and Mathematical Psychology*, 232.
24. KRUSCHKE, J. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review*, *99*(1), 22–44.
25. Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, *28*(1), 40–80. https://doi.org/10.3758/s13423-020-01792-x
26. 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.
27. Lee, J., Bahri, Y., Novak, R., Schoenholz, S. S., Pennington, J., & Sohl-Dickstein, J. (2017). Deep Neural Networks as Gaussian Processes. *ArXiv E-Prints*, arXiv-1711.
28. Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, January 16). *Efficient Estimation of Word Representations in Vector Space*. ArXiv.Org. https://arxiv.org/abs/1301.3781v3
29. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems*, *26*. https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html

30. Moody, J. (1991). The effective number of parameters: An analysis of generalization and regularization in nonlinear learning systems. *Advances in Neural Information Processing Systems*, *4*. https://proceedings.neurips.cc/paper/1991/hash/d64a340bcb633f536d56e51874281454-Abstract.html
31. 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.
32. Nasr, M., Rando, J., Carlini, N., Hayase, J., Jagielski, M., Cooper, A. F., Ippolito, D., Choquette-Choo, C. A., Tramer, F., & Lee, K. (2025). *SCALABLE EXTRACTION OF TRAINING DATA FROM ALIGNED, PRODUCTION LANGUAGE MODELS*.
33. Petrie, M., & Halliday, T. (1994). Experimental and natural changes in the peacock's (Pavo cristatus) train can affect mating success. *Behavioral Ecology and Sociobiology*, *35*, 213–217.
34. Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(5), 465–471.
35. Rosch, E. (1988). Principles of Categorization. In A. Collins & E. E. Smith (Eds.), *Readings in Cognitive Science, a Perspective From Psychology and Artificial Intelligence* (pp. 312–322). Morgan Kaufmann Publishers.
36. Sardana, N., Portes, J., Doubov, S., & Frankle, J. (2024). Beyond Chinchilla-optimal: Accounting for inference in language model scaling laws. *Proceedings of the 41st International Conference on Machine Learning*, 43445–43460. https://dl.acm.org/doi/abs/10.5555/3692070.3693840
37. Schölkopf, B., & Smola, A. J. (2002). *Learning with kernels: Support vector machines, regularization, optimization, and beyond*. MIT press. https://books.google.com/books?hl=en&lr=&id=y8ORL3DWt4sC&oi=fnd&pg=PR13&dq=Bernhard+Sch%C3%B6lkopf,+,+Alexander+J.+Smola+(2001).+Learning+with+Kernels+:+Support+Vector+Machines,+Regularization,+Optimization,+and+Beyond.+MIT+Press.&ots=bNxZbAQ3BC&sig=Kw53bZ4BoGuALpTOSxc0nngsG-s
38. Schooler, L. J., & Hertwig, R. (2005). How forgetting aids heuristic inference. *Psychological Review*, *112*(3), 610.
39. 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.
40. Schulz, E. (2017). *Towards a unifying theory of generalization* [PhD Thesis]. UCL (University College London).
41. 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.
42. Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, *6*(2), 461–464.

43. Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press. https://books.google.com/books?hl=en&lr=&id=Hf6QAwAAQBAJ&oi=fnd&pg=PR15&dq=Shai+Shalev-Shwartz+and+Shai+Ben-David+(2014).+Understanding+Machine+Learning:+From+Theory+to+Algorithms.+Cambridge+University+Press.&ots=2JuoSjlOK2&sig=Tir2kaX3gUsIOwuBwB6ZBUegEXY
44. Simon, H. A. (1990). Bounded Rationality. In J. Eatwell, M. Milgate, & P. Newman (Eds.), *Utility and Probability* (pp. 15–18). Palgrave Macmillan UK. https://doi.org/10.1007/978-1-349-20568-4_5
45. Sims, C. R. (2016). Rate–distortion theory and human perception. *Cognition*, *152*, 181–198.
46. Sims, C. R. (2018). Efficient coding explains the universal law of generalization in human perception. *Science*, *360*(6389), 652–656.
47. Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & Van Der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *64*(4), 583–639. https://doi.org/10.1111/1467-9868.00353
48. Van Der Linde, A. (2012). A Bayesian view of model complexity. *Statistica Neerlandica*, *66*(3), 253–271. https://doi.org/10.1111/j.1467-9574.2011.00518.x
49. Vapnik, V. (2006). *Estimation of dependences based on empirical data*. Springer Science & Business Media. https://books.google.com/books?hl=en&lr=&id=N_-5VRWai84C&oi=fnd&pg=PA8&dq=Vladimir+Vapnik+(2006).+Estimation+of+Dependences+Based+on+Empirical+Data:+Second+Edition.+Springer.&ots=RhFEzpkFNY&sig=Jp9QDGm3bJCOL_PqSLTf5CwDG9o
50. Vapnik, V. N., & Chervonenkis, A. Y. (1971). On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory of Probability and Its Applications*, *16*(2), 264.
51. 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.
52. Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). *Understanding Deep Learning Requires Rethinking Generalization*. *arXiv preprint arXiv:1611.03530*. https://pluskid.org/files/slides/ICLR2017-Poster.pdf