Overview

The simplicity principle in perception and cognition

Jacob Feldman*

The simplicity principle, traditionally referred to as Occam's razor, is the idea that simpler explanations of observations should be preferred to more complex ones. In recent decades the principle has been clarified via the incorporation of modern notions of computation and probability, allowing a more precise understanding of how exactly complexity minimization facilitates inference. The simplicity principle has found many applications in modern cognitive science, in contexts as diverse as perception, categorization, reasoning, and neuroscience. In all these areas, the common idea is that the mind seeks the simplest available interpretation of observations— or, more precisely, that it balances a bias toward simplicity with a somewhat opposed constraint to choose models consistent with perceptual or cognitive observations. This brief tutorial surveys some of the uses of the simplicity principle across cognitive science, emphasizing how complexity minimization in a number of forms has been incorporated into probabilistic models of inference. © 2016 Wiley Periodicals, Inc.

How to cite this article:
WIREs Cogn Sci 2016, 7:330–340. doi: 10.1002/wcs.1406

OCCAM'S RAZOR

The principle of *simplicity* or *parsimony*—broadly, the idea that simpler explanations of observations should be preferred to more complex ones—is conventionally attributed to William of Occam, after whom it is traditionally referred to as *Occam's razor.*[a] Since then philosophers of science have adopted a bias toward simpler explanations as a foundational principle of inference, guiding the selection of hypotheses whenever multiple hypothesis are consistent with data—as is nearly always the case.

But *why* should simpler theories be preferred? Practicing scientists have generally assumed it is because they are actually more likely to be correct. But it has never been clear exactly why this should be so. Hume's principle of "Uniformity of Nature" suggests that simpler theories are preferable because they make a good match for a highly regular, lawful world. Conversely, some philosophers have assumed that the bias toward simplicity is an essentially *aesthetic* preference, akin to elegance or beauty that mathematicians prize in theorems, conveying no particular claim to correctness (see Ref 2). Simpler theories were seen as more manageable, more comprehensible, and more testable,[3] but not necessarily more truthful. In the 20th century, some authors (see Refs 4,5) began to argue that simpler theories were, in fact, more likely to be true, but until recently the precise connection between simplicity and truth remained, at best, extremely unclear. To understand the connection, we need at the very least a more precise definition of *simplicity*. Such a definition only arrived in the last few decades.

MATHEMATICAL DEFINITIONS OF SIMPLICITY AND COMPLEXITY

Historically it has generally been assumed that simplicity and complexity were inherently *subjective* notions, impervious to clear, rigorous, or universal definitions. A theory that is simple in one method of expression may appear complex in another, implying that simplicity lies "in the eye of the beholder." A notorious example was the competition between

*Correspondence to: jacob@ruccs.rutgers.edu

Department of Psychology, Center for Cognitive Science, Rutgers University, New Brunswick, NJ, USA

Conflict of interest: The author has declared no conflicts of interest for this article.

Julian Schwinger's elaborate formalization of quantum thermodynamics, and Richard Feynman's apparently simpler account (based on "funny little diagrams")—which, notwithstanding their apparent difference in complexity, were eventually shown by Freeman Dyson to be equivalent.[6] Cases like this appeared to imply that complexity depends on the chosen method of expression, and thus that no quantification of complexity could be universal.

Kolmogorov Complexity

This all changed in the 1960s with the introduction of a principled and convincing mathematical definition of complexity now known as *Kolmogorov complexity* or *algorithmic complexity*. Introduced in slightly different forms by Solomonoff,[7] Kolmogorov,[8] and Chaitin,[9] the main idea is that the complexity of a string of characters reflects the degree of *incompressibility*, as measured by the length of a computer program required to faithfully express the string (see Ref 10). Simple strings are those that can be expressed by brief computer programs, and complex datasets are those that cannot. The idea is usually formalized by considering a string S (a sequence of symbols), and then considering the length (in symbols) of the shortest computer program capable of generating it. For example, the loop (in pseudocode) "for i = 1 to 1000: print '1'" prints a string of 1000 characters, but the program itself contains only 26; this string is highly compressible. By contrast a "typical" random string of 1000 characters (e.g., "39383827262226…") can't be compressed in this way, although it can be expressed by a program that is itself about 1000 characters long (e.g., "Print '9486390348473969683…'," which has 1008 characters). More generally, a string that contains regularities or patterns—of any form that can be expressed in the computer language—can be faithfully reproduced by a short program that takes advantages of these patterns, while a relatively complex or "random" string cannot be similarly compressed. This measure of complexity is inherently capped at approximately the length of the original string, because any string, no matter how irregular, can be reproduced exactly simply by quoting it verbatim as in the example above. In this view, simplicity is essentially *compressibility*.

Critically, Kolmogorov complexity is *universal* in the sense that it does not depend "very much" on the computer language in which the program is written. The caveat "very much" has a very precise meaning here, which derives from the fact that any computer language can be translated into any other computer language. Turing[11] had demonstrated the existence of computers (now referred to as universal Turing machines) that can, in a well-defined sense, carry out any concretely specifiable algorithm. In modern terminology, we can think of them as computer languages that are general enough to express any computable function—including, critically, to "simulate" other computer languages. Assume a string S that can be expressed by computer language L_1 in $K_1(S)$ steps, meaning that its Kolmogorov complexity is at least as low as $K_1(S)$. Assume that some other language L_2 can be expressed in language L_1 in a finite number $|L_2|$ of steps—in modern terms, $|L_2|$ is the length of a compiler for language L_2 written in language L_1. It follows fairly immediately that string S can be expressed in language L_2 in $|L_2| + K_1(S)$ steps, meaning that the complexity of S in language L_2 (i.e., $K_2(S)$) is at least as low as $|L_2| + K_1(S)$—because that is how many steps it takes to translate L_2 into L_1 and then express S in L_2. In other words, the expression of the string in the first language can be translated into the second language, with a general cap on the number of steps required by the process of translation. The "translation component" $|L_2|$ may be very large, if the computer languages are very different, but it is finite—and, critically, it does not depend on the length of the string. This means that as strings get longer and longer, the translation component of their complexity matters less and less, and in this sense, their complexity is asymptotically independent of the programming language. For this reason, the Kolmogorov complexity $K(S)$ of a string S is usually thought of as a *universal* measure of its inherent complexity or randomness.

Note that the actual value of the $K(S)$ is uncomputable.[b] For long strings, it can be approximated by effective string-compression algorithms such as Lempel-Ziv (see Ref 10) implemented in the common utility gzip,[14] meaning that the Kolmogorov complexity of a long string S is approximately the length, in characters, of the gzipped version of S. For shorter strings, such approximations are in principle less reliable, though recent work by Gauvrit et al.[15] has for the first time provided practical techniques for estimating the Kolmogorov complexity of short strings, opening an intriguing research avenue for evaluating the role of complexity in psychological models.

Information-Theoretic Description Length

Another important approach to the quantification of complexity was initiated by Shannon.[16] Shannon showed that in a set of messages m_1, m_2, … which occur with probability p_1, p_2 …, each message conveys information given by $-\log p_i$, which quantifies the degree of "surprise" or unexpectedness entailed by the message. Consequently, if one seeks to convey a set of messages in the fewest symbols possible, one should adopt a coding language in which each message m_i is assigned a code of length approximately $-\log p_i$ symbols. Such a procedure will minimize the expected total code length, that is, which achieves the most compressed expression possible. As a result, the quantity $-\log p_i$ is sometimes referred to as the *Description Length* (DL). The DL of a message is in effect a measure of complexity, because it quantifies how many symbols are required to express m_i in a maximally compressed code. That is, just like Kolmogorov complexity, the DL quantifies how many symbols are required to express a particular message *after maximal compression*. In this way, Shannon showed that complexity is intimately related to probability, a profound insight that pervades the modern understanding of both concepts. Much of modern complexity theory can be thought of as an elaboration of this idea.

Rissanen[17] took the next step by elevating this insight into a fundamental principle of inference, which he called the Minimum Description Length (MDL) principle (see Ref 18 and compare the closely related approach due to[19] called *Minimum Message Length*). In its classic formulation, the MDL principle begins by imagining that we are trying to explain some data X via some set of alternative models Y_i. For any given model Y, the joint probability $p(X \wedge Y)$ that both model and data are true can be written as $p(X|Y)p(Y)$, the product of the probability of the model and the probability of the data conditioned on the model. The DL of this conjunction, that is, its negative log probability, is simply

$$\begin{aligned}-\log p(X \wedge Y) &= -\log p(X|Y)p(Y)\\ &= -\log p(X|Y) + -\log p(Y)\\ &= \mathrm{DL}(X|Y) + \mathrm{DL}(Y).\end{aligned} \tag{1}$$

That is, the complexity (DL) of the model and data is the sum of the complexity of the model, plus the complexity of the data given the model—bearing in mind that, via Shannon's definition, the "complexity" of the data is really its surprisingness given the model. This neat additive formulation captures something very basic about scientific theorizing: that we are trying to simultaneously minimize the complexity of our theories *and* the unexpectedness (surprise) of the data given our theories—that is, that we seek elegant models that also explain the data reasonably well. This perfectly encapsulates Einstein's (perhaps apocryphal) quip that our theories should be "as simple as possible, but no simpler."

Bayesian Inference

The intimate relationship between Occam's razor and rational probabilistic inference was probably first pointed out by Jeffreys,[4] one of the principal developers of the modern conception of Bayesian inference. Jeffreys argued in some detail that the simplest interpretation was indeed the most likely one, and in particular advocated adopting priors that penalize complexity, that is, placing higher priors on simpler models and lower priors on more complex ones. More recently, Edwards[20] has also argued that probability theory inherently favors simpler inductions, though on the basis of the likelihood rather than the prior.

The close connection between Occam's razor and Bayes' rule can be appreciated most directly simply by observing that the hypothesis with the highest posterior is, ipso facto, also the hypothesis with the minimum DL in Shannon's sense. In a Bayesian framework, the posterior belief in hypothesis H after considering data D, notated $p(H|D)$, is proportional to the product of its prior $p(H)$ and its likelihood $p(D|H)$,

$$p(H|D) \propto p(H)p(D|H), \tag{2}$$

(see Refs 21,22 for tutorial introductions). The hypothesis that maximizes this quantity, sometimes called the maximum a posteriori or MAP, is the hypothesis that is the most probable, in that it maximizes the trade-off between prior plausibility and fit to the observed data. Hence in this simple sense, if one assumes an optimal description language in the sense defined by Shannon, the winning hypothesis is both the most probable *and* the simplest.

Even under a broader set of assumptions, Bayesian inference inherently favors simpler hypotheses because of the way it assigns probability,[23] a tendency often referred to as the "Bayes occam" factor. In practice, a Bayesian hypothesis space often consists of one more parameterized families of hypotheses. The more parameters a family has, the smaller the probability volume devoted to each individual hypothesis (i.e., each setting of the parameters), since the total probability assigned to all hypotheses must sum to one. Hence, if one thinks of the number of parameters as a measure of the complexity of the model family, the prior necessarily decreases with complexity. A similar argument applies to the likelihood as well,[24] suggesting that Bayesian inference automatically favors more restrictive (i.e., simpler) hypotheses even without an overt prior bias.

In the modern literature on machine learning and statistical learning, the intimate connection between simplicity and probability is part of what is called the *bias/variance trade-off*, terminology introduced by Geman and coauthors.[25] Very broadly speaking, complex theories are inherently more flexible and thus more capable of fitting training data. But this very flexibility leads them to generalize poorly, because they inevitably fit *noise* in the training data—random fluctuations unlikely to be replicated in future data from the same source (called *overfitting*). Simpler theories tend to generalize better, because they only fit the regularities and not the noise. But theories that are *too* simple miss the regularities as well as the noise (*underfitting*). This leads to a trade-off in which optimal inference requires a balance between simplicity and fit to the training data (see Refs 26,27 for discussion). Unfortunately, there is no way in principle to determine the correct balance, though proponents of Bayesian inference and MDL sometimes argue that these principles do so as well as possible given the information available. The bias/variance trade-off is widely regarded as a central aspect of all probabilistic inference, and is yet another reason why probability and simplicity are intertwined. Almost regardless of the nature of the inference problem, a bias toward simpler theories (in this context sometimes called *regularization)* is required in order to prevent overfitting.

Finally, notice that just as one can create a complexity from a probability by taking a negative logarithm, one can create a probability from a complexity by exponentiation. Given a set of strings S each having Kolmogorov complexity $K(S)$, one can construct a set of probabilities

$$p(S) \propto 2^{-K(S)}, \tag{3}$$

(see Ref 10). Such a distribution assigns higher probability to simpler strings (here thought of as models of data) and lower probability to more complex ones. This construction may appear contrived, but, as Solomonoff[7] observed, it yields a set of probabilities—for example, a prior in the Bayesian sense—that is universal in precisely the same sense that Kolmogorov complexity itself is universal: namely, that is, approximately correct regardless of the details of the coding language. Such a "universal prior" closes the loop connecting probability to complexity.

This connection between simplicity and probability has many nuances not mentioned here, and is not without controversy (see Ref 23 for a more substantial discussion), and is viewed somewhat differently by those from Bayesian and information-theoretic traditions (see Ref 28). However, notwithstanding the many subtleties, it is important to understand that in the modern technical literature and in cognitive science, Bayesian inference and complexity minimization are usually treated as deeply intertwined, if not practically the same thing.

THE SIMPLICITY PRINCIPLE IN PSYCHOLOGY

In psychology and cognitive science, the simplicity principle posits that the mind draws interpretations of the world—mental models or mental representations—that are as simple as possible, or, at least, that are biased toward simplicity.[29,30] The idea takes different forms in different areas of cognition, depending on the nature of the many perceptual and cognitive problems the mind encounters: perceptual interpretations of sense data, memory encodings of experience, causal interpretations of observations, and so forth. In MDL and Bayesian formulations, the principle can be extended to allow a trade-off, inherent in these frameworks, between simplicity and consistency with sense data and experience—meaning that the interpretation drawn by the mind in light of simplicity may not actually be consistent with observation. Nevertheless in many areas of cognition, briefly surveyed in the next few sections, researchers have found that human thought incorporates a bias toward simplicity.

Note that complexity often arises in psychological experiments as a nuisance variable, simply because it can have such a salient effect on performance. Many experiments include simple and complex conditions (often labeled in other ways such as "high-load" and "low-load", etc.), even when complexity *per se* is not the main topic of inquiry. Typically, "complex" conditions simply involve a larger number of items or features, although it should be noted that the sheer number of elements in a construct is not generally a good proxy for complexity, since (as will be seen below) patterns with an equal number of elements can vary widely in regularity or compressibility. This review will not generally include such studies, but will instead focus on studies in which the bias toward simplicity is the main topic of interest.

Perception

The principle of simplicity first arose in perceptual psychology via the Gestalt notion of *Prägnanz*, a broad term meant to encompass "such properties as regularity, symmetry, simplicity, and others" (see Ref 31). The idea is that the mind prefers coherent and plausible interpretations of sensory data, for example, interpreting contours as the boundaries of objects, completing shapes plausibly behind occluders, and so forth. Notwithstanding its somewhat vague definition, Prägnanz is often thought of as a kind of simplicity principle, sometimes under the rubric *minimum principle* (see Refs 32,33). The idea is that more coherent or "Prägnant" interpretations are in some sense simpler than alternatives.[34]

In the 1950s, following the introduction of computers and the dissemination of Shannon's ideas about information, some psychologists began to take up information-theoretic quantifications of complexity and Prägnanz. Attneave's influential paper[35] expressed the idea of simplicity in terms of "economy of perceptual description," and for the first time compared Shannon's formal information measure to human performance. Around the same time, Hochberg and McAlister[36] quantified the complexity of a stimulus in a Kolmogorov-like way (a decade before Kolmogorov, Chaitin, and Solomonoff), adding up the number of steps in the simplest generative procedure required to replicate a stimulus (e.g., the number of segments, turns, corners, and bends required to recreate a given line drawing; see also Ref 37). They demonstrated that subjects shown an ambiguous figure preferred interpretations in inverse proportion to their complexity quantified in this manner.

The attempt to create a general complexity measure for perceptual interpretations reached a greater level of sophistication in the work of Leeuwenberg.[38] Leeuwenberg, along with his followers in the tradition later known as structural information theory (SIT), articulated a coding language based on pattern repetitions, symmetries, and, later, other kinds of regularities.[39,40] Predictions derived from the theory have been used to account for various phenomena of visual completion (e.g. Refs 41,42) as well as motion interpretation.[43] For example, this work provides a concrete account of why, when we see an image of a horse behind a tree, we interpret it as a complete horse behind an occluding tree, rather that as the arbitrary juxtaposition of a front half of a horse, a tree, and a rear half of a horse. In the coding language used by the human visual system, it is simpler to complete the horse than to express this odd juxtaposition of partial horse parts. However, while the contribution of SIT is impressive, it should be noted that complexity in a fixed coding language such as SIT's cannot necessarily be assumed to be universal in the sense of Kolmogorov complexity unless the language has been shown to express *all* visual patterns (including shading, color, texture, etc., which the SIT coding language does not usually include) which to the author's knowledge never been demonstrated.

Regardless of the details of the complexity measure, the simplicity principle in visual perception has often been placed in opposition to the *Likelihood principle* (see Refs 34,44), which is the tendency of the visual system to see the most probable interpretation (see Refs 45–47), which is in turn descended from notions of ecological probability in perception originated by Egon Brunswik.[48,49] However, as discussed above, complexity minimization and probabilistic inference are now recognized to be closely aligned and indeed not always clearly distinguishable from each other. In the perception literature, this connection was first recognized in an influential paper by Chater[50] who argued that in many contexts the simplest visual interpretation is also the most likely to be veridical. The more specific connection between Bayesian inference and complexity minimization has been explored in a number of places since (e.g., see Refs 51,52).

The idea that the visual system chooses the simplest model consistent with visual input was expressed in a particularly memorable way in an influential paper by Adelson and Pentland.[53] They imagined the process of scene model construction via a metaphor in which the scene must be created by a combination of a metalworker, a painter, and a lighting designer, each of whom charges fees for constructive operations such as creating a surface, bending a surface, painting a surface, or adding a light source. The brain's task is to construct a scene consistent with the image data for the least cost—in other words, to construct a scene model with minimum complexity in this particular "coding language." Because the coding language is in principle capable of generating any observable scene (albeit possibly with an enormous number of surfaces and colors, etc.), this dollar cost is a close analog of the Kolmogorov complexity. Simple images are those that can be rendered via inexpensive models, and the simplest (cheapest) model of the image is the most likely hypothesis about what arrangement of surfaces in the real world actually generated it.

The role of simplicity in vision has been particularly prominent in relation to perceptual organization and vision, where it originated. The work of Leeuwenberg and his followers on simplicity in perceptual organization has already been mentioned. In the computational literature, Darrell et al.[54] showed how the visual image can be parsed into coherent objects by choosing the decomposition with minimum DL. Feldman[55] similarly showed how the most plausible grouping interpretation can be chosen via a suitably chosen complexity minimization. Similarly, configurations of dots are clustered in part based on simplicity criteria.[56] Regardless of the specifics of the complexity measures, all these results suggest that the human visual system divides the image into coherent units in part based on the simplicity principle.

Similar principles govern how individual objects are represented. In the Bayesian shape representation framework of Feldman and Singh,[57] individual shapes are parsed into individual parts by choosing the simplest (MDL) skeletal representation consistent with the bounding contour. A closely related complexity measure has been shown to influence the detectability of both open contours in noise[58] as well as closed contours, that is, whole shapes.[59] Finally, a simplicity bias influences how the visual system interprets three-dimensional (3D) structure in line drawings; the system apparently chooses the simplest 3D shape consistent with the configuration of line elements (Ref 60).

Categorization and Concept Learning

The role of simplicity biases was recognized early in the machine learning literature (see Refs 61,62). Algorithmic approaches to learning have grown enormously since then, diverging into a number of frameworks with their own complexity measures. PAC ("probably approximately correct") learning, introduced by Valiant,[63] often uses a complexity measure called VC (Vapnik-Chervonenkis) dimension (see Ref 64), which as with nearly all complexity measures relates to how model complexity needs to be constrained in order to ensure learnability. As mentioned above, statistical learning theory, including the theory of neural networks (e.g. Ref 65) more generally assumes a trade-off between model complexity and fit to training data in order to promote effective generalization.

In the psychological literature on concept learning, the role of simplicity was noticed early.[66–68] Rosch[69] articulated a "principle of cognitive economy" as a motivation for why the mind reflexively organizes the world into coherent categories. However, this idea actually played relatively little role in the models that dominated the categorization literature for the next several decades, exemplar models (e.g., Refs 70,71). Exemplar models assume that categorization is a by-product of the storage of specific examples, which are then used as standards against which to judge the category membership of future examples. Exemplar models do not have an overt simplicity bias, because they do not involve any abstraction process *per se*, although later analysis made it clear that they implicitly regularize to a degree modulated by certain parameter settings.[72,73] Later "hybrid" (prototype plus exemplar) models, such as that of Nosofsky et al.[74] and others that followed, posited that learning proceeds by discovering collections of items that are well described by a simple rule, which can be stored separately from "exceptional" (irregular) items; such a strategy obviously requires an overt simplicity bias. Pothos and Chater[75] showed that unsupervised categorization too can be understood as a process of complexity minimization, using an MDL criterion that maximizes similarity within clusters and minimizes it between them. Similarly, Hahn et al.[76] showed how *similarity*, an essential construct in almost all categorization models, can be understood in terms of the Kolmogorov complexity of the transformation between objects.

The role of complexity in category learning has been studied the most directly in the context of Boolean categories, that is, categories built out of combinations of binary features.[77] Because Boolean categories involve finite combinations of discrete features, it is possible to test them comprehensively, including every distinct logical type.[78] In early work, several studies had suggested that the subjective difficulty of Boolean concepts could be tied to their logical complexity.[66,68] More recently, Feldman[79] undertook a more comprehensive study incorporating a much larger set of concept types. The results show that subjects' ability to learn Boolean concepts declines with their inherent logical complexity, suggesting (yet again) a bias toward simplicity in learning.

In its traditional definition (see Ref 80), Boolean complexity is defined as the length (in variable names, or literals) of the shortest propositional formula equivalent to a given set of examples. For example, the propositional formula $(A \wedge B) \vee (A \wedge B')$ describes two training examples, one with features A and B, the other with features A and not B. (In this notation, A and B are features, A' is the negation of feature A, $\wedge$ means "and," and $\vee$ means "or.") This expression can be reduced to $A \wedge (B \vee B')$ which in turn reduces to A, meaning that the original examples can be fully expressed simply by their common feature A; its Boolean complexity is 1. By contrast, the examples $(A \wedge B) \vee (A' \wedge B')$ cannot be reduced at all (it is "incompressible") so its Boolean complexity is 4. This definition parallels that of Kolmogorov complexity, in that it quantifies the length of the shortest faithful representation of the original formula, and enjoys an analogous kind of universality: Boolean complexity is universal across logical bases (i.e., choice of connectives) up to a multiplicative factor.

The simplicity bias in Boolean concept learning since has been corroborated by several studies, though there are still a number of distinct views about how to properly formulate the complexity measure.[81–86] Mathy et al. have measured complexity in terms of the length of a decision tree,[87] and even shown that response times can be tied to the process of decompression from a simplified representation.[88] Finally, Goodman et al.[89] were able to explain a large swath of concept learning data with a Bayesian model that assigns probability to logical formulae (i.e., models) in proportion to the length of their derivation in a context-free grammar, thus in effect favoring simple models over more complex ones.

Memory

Ideas from complexity theory have also proved useful in the study of memory and mental representation. This connection is somewhat different from others reviewed above, in that it does not overtly involve an attempt to draw inferences from data. In a sense, though the connection is more direct, because the compressibility of a piece of information—that is, its Kolmogorov complexity—relates directly to the amount of storage space required to encode it. Hence, Kolmogorov complexity and other measures of compressibility relate directly to the efficiency of memory encoding.

Mathy and Feldman[90] demonstrated this connection fairly directly by manipulating the compressibility of digit sequences to be held by subjects in verbal short-term memory (STM), for example, including "runs" (chains of rising or falling digits, like 3-4-5-6-7 or 8-7-6) of various lengths. The more regular (and thus compressible) the sequence, the more digits ordinary subjects could retain correctly. The implication is that verbal STM incorporates an active compression or pattern-finding mechanism that allows it to minimize memory resources. Children, too, show an effect of compressibility, meaning that their verbal STM capacity also varies in direct proportion to the compressiblity of the material to be remembered.[91] Intriguingly, while their digit span rises with age, the effect of compressibility is apparently constant, meaning that as children develop they retain a fixed capacity to compress information.

A closely related debate involves the capacity of visual STM, the buffer in which visual information is briefly held. Like verbal STM, visual STM had traditionally been assumed to contain a fixed number (about 3 or 4) of slots without regard to information load. But this "fixed slots" view has been challenged in favor of a "continuous-resource" view in which memory resources are flexibly allocated depending on intrinsic information load. For example, several studies have found that the number of items stored depends on the complexity of each item[92] or the precision with which each is represented,[93] implying that capacity is bound by the total information content rather than by a fixed slot limit. Ma et al.[94] provide a good recent summary, concluding that visual STM capacity depends on a continuous information limit rather than a set of discrete slots. This is a contentious debate with a number of open controversies (see for example Ref 95 for a conflicting view). But broadly speaking, these findings point to an underlying compression system in which information is reflexively represented in the most parsimonious way possible. Although the nature of the underlying neural code is not yet well understood, recent models assume a maximally compressed code that is efficient in the Shannon sense.[96]

Causal Reasoning

Another natural setting for a simplicity bias is in the inference of causal explanations from observations. When a doctor is confronted with a set of symptoms (say, fatigue, fever, and a runny nose), it is simpler and thus more reasonable to diagnose a single cause (the flu) rather than a set of distinct causes (anemia, sepsis, and seasonal allergies). Here again the simplicity bias can be described in Bayesian terms, as the assignment of a higher prior to a single cause than to a set of three distinct causes (which, being independent, would have a prior approximately proportional the third power of that of a single cause, a much lower number). The confidence inspired by a simple explanation of a complex set of facts (*Eureka!*) derives in part from the fact that it is unlikely for a simple theory to fit "by accident".[97] Accordingly, Little and Shiffrin[98] found that subjects favor simple explanations of data (e.g., lower-degree polynomial models) over more complex (higher-degree) explanations. Similarly, studies of children's explanations of causal relations have found that they too favor explanations that minimize the number of distinct causes.[99,100]

Neuroscience

Finally, we briefly mention the role that complexity, and in particular minimization of coding length, plays in theoretical neuroscience. Barlow[101] famously advocated *efficient coding* as a fundamental principle of sensory representation, arguing that the brain encodes sensory signals so as to minimize the redundancy of the raw stimulus array. This idea is essentially equivalent to the notion of compression later connected with Kolmogorov complexity, as it entails an Occam-like compression of the raw sensory signal in order to extract the regularities latent within it.

Barlow's idea has exerted a profound and far-reaching influence on neursocience in the decades since. The idea that neural networks extract regularity from sensory data, sometimes referred to as *dimensionality reduction*, is a central principle of contemporary theoretical neuroscience.[102] Similarly, neural receptive fields are now thought to be designed so as to optimally (that is, with maximal informational efficiency) encode visual stimuli.[103] Another important development along the same lines is in the quantification of information along neural spike trains, which is based on the idea that the sequence of action potential constitutes an optimally efficient encoding of the information conveyed by sensory receptors.[104] All these developments have in common an Occam-like reduction of the complexity of the raw stimulus array in order to further the behavioral goals of the organism.

Finally note that the brain's neural circuitry itself, viewed from the perspective of graph theory, appears to minimize circuit complexity and other computational costs.[105] While this is admittedly speculative, it is possible that (in some not-yet-understood sense) the *mental* bias toward relatively simple interpretations of the world might be related to a *neural* bias toward simplicity in the underlying neural architecture.

CONCLUSION

As the many examples above illustrate, a bias toward simplicity pervades mental function. Examples can be found in perception, learning, categorization, reasoning, and neuroscience. Some of these findings involve Kolmogorov complexity directly; others involve information-theoretic concepts like DL (negative log probability) and information load; and others involve simplicity biases that arise in the context of Bayesian inference—all of which are closely related from a mathematical point of view. Broadly speaking, it may be that, as Hume first suggested, the mind cannot apprehend the world without assuming some form of underlying regularity, an idea sometimes called the "Principle of Natural Modes".[106]

However, notwithstanding the ubiquity of simplicity biases, it actually remains unclear whether Occam's razor is, in fact, a primary driving principle of human inference. As discussed above, simplicity biases are deeply intertwined with—indeed scarcely separable from—information theory and Bayesian probability theory. From the point of view of modern theory, almost any form of rational inference will entail some kind of simplicity bias. Hence rather than being a foundational principle, the human simplicity bias may simply be an epiphenomenon of a more basic goal of mental function, such as veridicality,[107] optimal estimation,[108] or, perhaps most fundamentally, adaptive functionality.[109,110]

NOTES

[a] Occam himself was arguing against the existence of "universals" (i.e., generalizations), maintaining that one should not posit the existence of entities beyond those that can be directly observed; see Ref 1.

[b] Kolmogorov complexity is uncomputable for essentially the same reason there is no "Smallest uninteresting number"—if there were, that would indeed be very interesting. (see Ref 12 on what Bertrand Russell called the Berry paradox.) Similarly, if there were a computable procedure for computing $K(S)$, then some strings of complexity at least $K(S)$ could be reproduced by a program of the form "Print the shortest string with complexity $K(S)$." Such a program would take fewer than $K(S)$ characters to encode—a contradiction. See Schöning and Pruim[13] for a more careful discussion.

ACKNOWLEDGMENTS

Preparation of this article was supported by NIH (NEI) Grant EY021494. I am grateful to Manish Singh and members of the Visual Cognition Lab for many stimulating discussions.

REFERENCES

1. Hannam J. *God's Philosophers: How the Medieval World Laid the Foundations of Modern Science*. London: Icon Books; 2009.
2. Sober E. *Simplicity*. London: Oxford University Press; 1975.
3. Popper KR. *The Logic of Scientific Discovery*. New York: Basic Books; 1934/1959.
4. Jeffreys H. *Theory of Probability*. 3rd ed. Oxford: Clarendon Press; 1939/1961.
5. Quine WVO. On simple theories of a complex world. In: Foster MH, Martin ML, eds. *Probability, Confirmation, and Simplicity: Readings in the Philosophy of Inductive Logic*. New York: Odyssey Press; 1965, 250–252.
6. Krauss LM. *Quantum Man: Richard Feynman's Life in Science*. New York: Atlas & Co.; 2011.
7. Solomonoff R. A formal theory of inductive inference: part II. *Inform Contr* 1964, 7:224–254.
8. Kolmogorov AN. Three approaches to the quantitative definition of information. *Probl Inform Transm* 1965, 1:1–7.
9. Chaitin GJ. On the length of programs for computing finite binary sequences. *J ACM* 1966, 13:547–569.
10. Li M, Vitányi P. *An Introduction to Kolmogorov Complexity and Its Applications*. New York: Springer; 1997.
11. Turing AM. On computable numbers, with an application to the entscheidungs problem. *Proc Lond Math Soc* 1937, 2:230–265.
12. Chaitin GJ. Information-theoretic computational complexity. *IEEE Trans Inform Theor* 1974, IT-20:10–15.
13. Schöning U, Pruim R. *Gems of Theoretical Computer Science*. Berlin: Springer; 1998.
14. Ziv J, Lempel A. A universal algorithm for sequential data compression. *IEEE Trans Inform Theor* 1977, 23:337–343.
15. Gauvrit N, Singmann H, Soler-Toscano F, Zenil H. Algorithmic complexity for psychology: a user-friendly implementation of the coding theorem method. *Behav Res Methods* 2016, 48:314–329.
16. Shannon C. A mathematical theory of communication. *Bell Syst Tech J* 1948, 27:379–423.
17. Rissanen J. Modeling by shortest data description. *Automatica* 1978, 14:465–471.
18. Grünwald PD. A tutorial introduction to the minimum description length principle. In: Grünwald PD, Myung IJ, Pitt M, eds. *Advances in Minimum Description Length: Theory and Applications*. Cambridge, MA: MIT press; 2005.
19. Wallace CS. *Statistical and Inductive Inference by Minimum Message Length*. New York: Springer; 2004.
20. Edwards AWF. *Likelihood*. Cambridge: Cambridge University Press; 1972.
21. Feldman J. Bayesian models of perceptual organization. In: Wagemans J, ed. *Handbook of Perceptual Organization*. Oxford: Oxford University Press; 2014, 1008–1026.
22. Yuille AL, Bülthoff HH. Bayesian decision theory and psychophysics. In: Knill DC, Richards W, eds. *Perception as Bayesian Inference*. Cambridge: Cambridge University Press; 1996, 123–162.
23. MacKay DJC. *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press; 2003.
24. Tenenbaum JB, Griffiths TL. Generalization, similarity, and Bayesian inference. *Behav Brain Sci* 2001, 24:629–640.
25. Geman S, Bienenstock E, Doursat R. Neural networks and the bias/variance dilemma. *Neural Comput* 1992, 4:1–58.
26. Duda RO, Hart PE, Stork DG. *Pattern Classification*. New York: John Wiley & Sons; 2001.
27. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer; 2001.
28. Burnham KP, Anderson DR. *Model Selection and Multi-Model Inference: A Practical Information-Theoretic Approach*. New York: Springer; 2002.
29. Chater N. Simplicity and the mind. *Psychologist* 1997, 10:495–498.
30. Chater N, Vitányi P. Simplicity: a unifying principle in cognitive science. *Trends Cogn Sci* 2003, 7:19–22.
31. Koffka K. *Principles of Gestalt Psychology*. New York: Harcourt; 1935.
32. Boselie F, Wouterlood D. The minimum principle and visual pattern completion. *Psychol Res* 1989, 51:93–101.
33. Kanizsa G. *Organization in Vision: Essays on Gestalt Perception*. New York: Praeger Publishers; 1979.
34. Hatfield G, Epstein W. The status of the minimum principle in the theoretical analysis of visual perception. *Psychol Bull* 1985, 97:155–186.
35. Attneave F. Some informational aspects of visual perception. *Psychol Rev* 1954, 61:183–193.
36. Hochberg J, McAlister E. A quantitative approach to figural "goodness". *J Exp Psychol* 1953, 46:361–364.
37. Hochberg J. *Perception*. Englewood Cliffs, NJ: Prentice-Hall; 1964.
38. Leeuwenberg ELJ. A perceptual coding language for visual and auditory patterns. *Am J Psychol* 1971, 84:307–349.
39. van der Helm PA. Simplicity in perceptual organization. In: Wagemans J, ed. *Handbook of Perceptual Organization*. Oxford: Oxford University Press; 2014.

40. van der Helm PA. *Simplicity in Vision*. Cambridge: Cambridge University Press; 2015.

41. van Lier RJ, Leeuwenberg ELJ, van der Helm PA. Multiple completions primed by occlusion patterns. *Perception* 1995, 24:727–740.

42. van Lier RJ, van der Helm P, Leeuwenberg ELJ. Integrating global and local aspects of visual occlusion. *Perception* 1994, 23:883–903.

43. Restle F. Coding theory of the perception of motion configurations. *Psychol Rev* 1979, 86:1–24.

44. Pomerantz JR, Kubovy M. Theoretical approaches to perceptual organization. In: Boff KR, Kaufman L, Thomas JP, eds. *Handbook of Perception and Human Performance, Volume 2: Cognitive Processes and Performance*. New York: John Wiley & Sons; 1986, 36-1–36-46.

45. Boselie F, Leeuwenberg E. A test of the minimum principle requires a perceptual coding system. *Perception* 1986, 15:331–354.

46. Leeuwenberg ELJ, Boselie F. Against the likelihood principle in visual form perception. *Psychol Rev* 1988, 95:485–491.

47. Moravec L, Beck J. Amodal completion: simplicity is not the explanation. *Bull Psychon Soc* 1986, 24:269–272.

48. Brunswik E. *Perception and the Representative Design of Psychological Experiments*. Berkeley: University of California Press; 1956.

49. Brunswik E, Kamiya J. Ecological cue-validity of 'proximity' and of other Gestalt factors. *Am J Psychol* 1953, 66:20–32.

50. Chater N. Reconciling simplicity and likelihood principles in perceptual organization. *Psychol Rev* 1996, 103:566–581.

51. Feldman J. Bayes and the simplicity principle in perception. *Psychol Rev* 2009, 116:875–887.

52. Vitányi PMB, Li M. Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Trans Inf Theory* 2000, 46:446–464.

53. Adelson EH, Pentland AP. The perception of shading and reflectance. In: Knill C, Richards DW, eds. *Perception as Bayesian Inference*. Cambridge: Cambridge University Press; 1996, 409–423.

54. Darrell, T., Sclaroff, S. & Pentland, A. Segmentation by minimal description. In: *Proceedings Third International Conference on Computer Vision*. Los Alamitos, CA: IEEE Computer Society Press; 1990, 112–116.

55. Feldman J. Regularity-based perceptual grouping. *Comput Intell* 1997, 13:582–623.

56. Gershman SJ, Niv Y. Perceptual estimation obeys Occam's razor. *Front Psychol* 2013, 4:623.

57. Feldman J, Singh M. Bayesian estimation of the shape skeleton. *Proc Natl Acad Sci* 2006, 103:18014–18019.

58. Wilder J, Feldman J, Singh M. Contour complexity and contour detection. *J Vis* 2015a, 15:1–16.

59. Wilder J, Feldman J, Singh M. The role of shape complexity in the detection of closed contours. *Vision Res* 2015b.

60. Li Y, Pizlo Z. Depth cues versus the simplicity principle in 3D shape perception. *Top Cogn Sci* 2011, 3:667–685.

61. Iba, W., Wogulis, J. & Langley, P. Trading off simplicity and coverage in incremental concept learning. In: *Proceedings of the Fifth International Conference on Machine Learning*, Ann Arbor, MI, 1988, 7379, 73–79.

62. Medin DL, Wattenmaker WD, Michalski RS. Constraints and preferences in inductive learning: an experimental study of human and machine performance. *Cognit Sci* 1987, 11:299–339.

63. Valiant L. A theory of the learnable. *Commun ACM* 1984, 27:1134–1142.

64. AbuAbu-Mostafa YS. The Vapnik-Chervonenkis dimension: information versus complexity in learning. *Neural Comput* 1989, 1:312–317.

65. Poggio T, Rifkin R, Mukherjee S, Niyogi P. General conditions for predictivity in learning theory. *Nature* 2004, 428:419–422.

66. Haygood RC. *Rule and Attribute Learning as Aspects of Conceptual Behavior*. University of Utah; 1963.

67. Looney NJ, Haygood RC. Effects of number of relevant dimensions in disjunctive concept learning. *J Exp Psychol* 1968, 78:169–171.

68. Neisser U, Weene P. Hierarchies in concept attainment. *J Exp Psychol* 1962, 64:640–645.

69. Rosch E. Principles of categorization. In: Rosch E, Lloyd B, eds. *Cognition and Categorization*. Hillsdale, NJ: Lawrence Erlbaum; 1978, 27–48.

70. Kruschke J. ALCOVE: An exemplar-based connectionist model of category learning. *Psychol Rev* 1992, 99:22–44.

71. Nosofsky RM. Attention, similarity, and the identification-categorization relationship. *J Exp Psychol Gen* 1986, 115:39–61.

72. Briscoe, E. & Feldman, J. Conceptual complexity and the bias-variance tradeoff. In: *Proceedings of the Conference of the Cognitive Science Society*, Vancouver, Canada, July, 2016, 1038–1043.

73. Jäkel F, Schölkopf B, Wichmann F. Generalization and similarity in exemplar models of categorization: Insights from machine learning. *Psychon Bull Rev* 2008, 15:256–271.

74. Nosofsky RM, Palmeri TJ, McKinley SC. Rule-plus-exception model of classification learning. *Psychol Rev* 1994, 101:53–79.

75. Pothos EM, Chater N. Categorization by simplicity: a minimum description length approach to unsupervised clustering. In: Hahn U, Ramscar M, eds. *Similarity and Categorization*. Oxford University Press: Oxford; 2001, 51–72.

76. Hahn U, Chater N, Richardson LB. Similarity as transformation. *Cognition* 2003, 87:1–32.
77. Feldman J. The simplicity principle in human concept learning. *Curr Dir Psychol Sci* 2003, 12:227–232.
78. Shepard RN, Hovland CL, Jenkins HM. Learning and memorization of classifications. *Psychol Monogr* 1961, 75:1–42.
79. Feldman J. Minimization of Boolean complexity in human concept learning. *Nature* 2000, 407:630–633.
80. Wegener I. *The Complexity of Boolean Functions*. Chichester: John Wiley & Sons; 1987.
81. Aitkin, C. D. & Feldman, J. Subjective complexity of categories defined over three-valued features. In: *Proceedings of the 28th Conference of the Cognitive Science Society*, Vancouver, Canada, July, 2006, 961¢-966.
82. Fass D, Feldman J. Categorization under complexity: a unified MDL account of human learning of regular and irregular categories. In: Becker S, Thrun S, Obermayer K, eds. *Advances in Neural Information Processing 15*. Cambridge, MA: MIT Press; 2002.
83. Feldman J. An algebra of human concept learning. *J Math Psychol* 2006, 50:339–368.
84. Lafond D, Lacouture Y, Mineau G. Complexity minimization in rule-based category learning: revising the catalog of Boolean concepts and evidence for non-minimal rules. *J Math Psychol* 2007, 51:57–74.
85. Mathy F. Assessing conceptual complexity and compressibility using information gain and mutual information. *Tutor Quant Methods Psychol* 2010, 6:16–30.
86. Vigo R. Categorical invariance and structural complexity in human concept learning. *J Math Psychol* 2009, 53:203–221.
87. Mathy F, Bradmetz J. A theory of the graceful complexification of concepts and their learnability. *Curr Psychol Cogn* 2004, 22:41–82.
88. Bradmetz J, Mathy F. Response times seen as decompression times in Boolean concept use. *Psychol Res* 2008, 72:211–234.
89. Goodman ND, Tenenbaum JB, Feldman J, Griffiths TL. A rational analysis of rule-based concept learning. *Cognit Sci* 2008, 32:108–154.
90. Mathy F, Feldman J. What's magic about magic numbers? Chunking and data compression in short-term memory. *Cognition* 2012, 122:346–362.
91. Mathy F, Fartoukh M, Gauvrit N, Guida A. Developmental abilities to form chunks in immediate memory and its non-relationship to span development. *Front Psychol* 2016, 7:201.
92. Luria R, Sessa P, Gotler A, Jolicoeur P, Dell'Acqua R. Visual short-term memory capacity for simple and complex objects. *J Cogn Neurosci* 2010, 22:496–512.
93. Alvarez GA, Cavanagh P. The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychol Sci* 2004, 15:106–111.
94. Ma WJ, Husain M, Bays PM. Changing concepts of working memory. *Nat Neurosci* 2014, 17:347–356.
95. Awh E, Barton B, Vogel EK. Visual working memory represents a fixed number of items regardless of complexity. *Psychol Sci* 2007, 18:622–628.
96. Sims CR, Jacobs RA, Knill DC. An ideal observer analysis of visual working memory. *Psychol Rev* 2012, 119:807–830.
97. Feldman J. How surprising is a simple pattern? Quantifying "Eureka!". *Cognition* 2004, 93:199–224.
98. Little, D. R. & Shiffrin, R. M. Simplicity bias in the estimation of causal functions. In: *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, Amsterdam, Netherlands, July, 2009, 1157–1162.
99. Bonawitz EB, Lombrozo T. Occam's rattle: children's use of simplicity and probability to constrain inference. *Dev Psychol* 2012, 48:1156–1164.
100. Lombrozo T. Simplicity and probability in causal explanation. *Cogn Psychol* 2007, 55:232–257.
101. Barlow HB. Possible principles underlying the transformation of sensory messages. In: Rosenblith WA, ed. *Sensory Communication*. Cambridge, MA: MIT Press; 1961, 217–234.
102. Hinton GE, Salakhutdinov RR. Reducing the dimensionality of data with neural networks. *Science* 2006, 313:504–507.
103. Field DJ. Relations between the statistics of natural images and the response properties of cortical cells. *J Opt Soc Am A* 1987, 4:2379–2394.
104. Rieke F, Warland D, de Ruyter van Steveninck R, Bialek W. *Spikes: Exploring the Neural Code*. Cambridge, MA: MIT Press; 1996.
105. Bullmore E, Sporns O. The economy of brain network organization. *Nat Rev Neurosci* 2012, 13:336–349.
106. Richards WA, Bobick A. Playing twenty questions with nature. In: Pylyshyn Z, ed. *Computational Processes in Human Vision: An Interdisciplinary Perspective*. Norwood, NJ: Ablex Publishing Corporation; 1988, 3–26.
107. Pizlo Z, Sawada T, Li Y, Kropatsch WG, Steinman RM. New approach to the perception of 3D shape based on veridicality, complexity, symmetry and volume. *Vision Res* 2010, 50:1–11.
108. Knill DC, Kersten D, Yuille A. Introduction: a Bayesian formulation of visual perception. In: Knill DC, Richards W, eds. *Perception as Bayesian Inference*. Cambridge: Cambridge University Press; 1996, 123–162.
109. Hoffman DD, Singh M, Prakash C. The interface theory of perception. *Psychon Bull Rev* 2015, 22:1480–1506.
110. Todd PM, Gigerenzer G, the ABC Research Group. *Ecological Rationality: Intelligence in the World*. Oxford: Oxford University Press; 2012.