The generalized universal law of generalization

Nick Chater[a],[*],[1] and Paul M.B. Vitányi[b],[2]

[a] *Department of Psychology, University of Warwick, Institute for Applied Cognitive Science, Coventry, CV4 7AL, UK*
[b] *Centrum voor Wiskunde en Informatica, Kruislaan 413, 1098 SJ Amsterdam, The Netherlands*

Abstract

It has been argued by Shepard that there is a robust psychological law that relates the distance between a pair of items in psychological space and the probability that they will be perceived as similar. Specifically, this probability is a negative exponential function of the distance between the pair of items. In experimental contexts, distance is typically defined in terms of a multidimensional space—but this assumption seems unlikely to hold for complex stimuli. We show that, nonetheless, the Universal Law of Generalization can be derived in the more complex setting of arbitrary stimuli, using a much more universal measure of distance. This universal distance is defined as the length of the shortest program that transforms the representations of the two items of interest into one another: The algorithmic information distance. It is universal in the sense that it minorizes every computable distance: It is the smallest computable distance. We show that the Universal Law of Generalization holds with probability going to one—provided the probabilities concerned are computable. We also give a mathematically more appealing form of the Universal Law.

1. Introduction

Shepard (1987) has put forward a 'Universal Law of Generalization' as one of the few general psychological results governing human cognition. The law states that the probability of perceiving similarity or analogy between two items, a and b, is a negative exponential function of the distance $d(a, b)$ between them in an internal psychological space. A large body of empirical data, from a variety of psychological domains, has been collected in support of the Universal Law, and theoretical derivations have been provided to support it, in specific mathematical settings. Shepard emphasizes that the structure of the mental spaces in which items are

represented may have a variety of different forms, depending on the nature of the items that are being represented. Such spaces can range from spaces with Euclidean or Minkowski metrics, to tree structures (Pruzansky, Tversky, & Carroll, 1982). In practice, though, both empirical evidence and theoretical justifications (Shepard, 1987; Tenenbaum & Griffiths, 2001) generally focus on cases where stimuli can be embedded in a Euclidean space. This focus on traditional metric spaces raises two interesting issues, one empirical and one theoretical.

The empirical issue concerns whether the Universal Law applies where confusability data is best modelled by non-standard metrics. Evidence on this question appears to be sparse, perhaps because scaling techniques that embed items in Euclidean spaces are particularly well-developed and widely used. One piece of evidence that the law may extend to other metrics is given in (Cunningham & Shepard, 1974). Confusability data for Morse Code signals collected by (Rothkopf, 1957) was analysed by a very general scaling method, which makes only the metric assumptions. This data shows qualitatively the same pattern as in conventional non-metric multidimensional scaling analysis, consistent with the Universal Law. Below, we shall follow Shepard in assuming that the Universal Law does hold good, using

---

[*] Corresponding author. Fax: +24-7652-4225.

E-mail addresses: nick.chater@warwick.ac.uk (N. Chater), paulv@cwi.nl (P.M.B. Vitányi).

whatever metric is most appropriate for representing the data.

This assumption then raises the theoretical issue of how the Universal Law is to be explained. Current theoretical proposals focus, quite appropriately, on the Euclidean case (Shepard, 1987; Tenenbaum & Griffiths, 2001). But if the Universal Law applies to a range of different metrics, then likely as not for each metric, a separate explanation of why the Universal Law holds in that metric appears to be required. And if it turns out that, indeed, the law is universal, it is not clear how this fact can be explained, aside from as a remarkable co-incidence. Thus, if we are to provide a unified explanation of the Universal Law of generalization which applies to stimuli which come from different underlying spaces, it seems that a more general analysis will be required. The present article attempts to provide such an analysis.

Specifically, we note that there exists a universal cognitive metric, the 'information metric' (Bennett, Gács, Li, Vitányi, & Zurek, 1998), that accounts for all possible similarities that can be perceived. It assigns nearly as small a distance between two objects as any cognitive distance will do. Thus, while the positive and negative of the same picture are far away from each other in terms of Euclidean distance, they are at almost zero distance in terms of universal distance because interchanging the black and white pixels transforms one picture into the other. The universal cognitive metric, also called 'information distance', is a mathematical notion derived from mathematical logic, computer science, information theory, and the theory of randomness. It is an 'ideal' notion in the sense that it ignores the limitations on processing capacity, or the evolutionarily acquired goal-driven restrictions of the cognitive system. Nonetheless, we show the following concrete generalization of the Universal Law (of generalization): if we randomly pick items a and b, where we allow the most complex objects, then with overwhelming probability the Universal Law of Generalization holds with the internal psychological space, where the metric of that space is a specific, very general, metric, the 'information' metric.

1.1. Level of abstraction of the analysis

It is reasonable to ask whether an 'ideal' notion such as information distance, which idealizes away from processing details of the cognitive system, is too abstract to be of use in explaining specific psychological data. We suggest the following viewpoint: (i) The new treatment marks the ultimate limits of the validity of the Universal Law of Generalization; (ii) the Universal Law arises from very general features of the cognitive system, and hence is appropriately explained at an abstract level; and (iii) the theoretical results, although requiring unlimited

computing power and computing time in worst-case scenarios, are relevant in practical situations where the problem instances are often simple and, moreover, where only approximate and non-optimal answers are required.

Let us expand on this a little. In (Feldman, 2000) the question of the subjective difficulty of learning a Boolean concept (that is, a formula in elementary mathematical logic) is analysed, with the purpose of giving a more fundamental underpinning of classification of Boolean concepts by their ease of learning by humans, presented in (Shepard, Hovland, & Jenkins, 1961). The empirical result described is that the subjective difficulty experienced by humans appeared to be directly proportional to the length of the shortest equivalent logical formula (Feldman, 2000). These results suggest that the cognitive system compresses the Boolean concept to its smallest equivalent Boolean formula, and learns that at a rate that is related to its length. Yet, according to basic results from computer science, compression of Boolean concepts to its smallest equivalent Boolean formula by any general method in human or machine, is out of the question. The problem is NP-hard (Garey & Johnson, 1979), which means that to solve even moderate size instances of the problem will require more than the life time of the universe for all contemporary and future computing machines (and humans) alike—at least, according to universally accepted assumptions in mathematics. Yet in the experimental situation of (Feldman, 2000) this problem was apparently resolved both explicitly by the experimenters and implicitly by the cognitive system of the subjects. How can that be? The answer, as in many such practical cases, is that the actual problem instances were extremely simple. In real life, even for mathematically hard (NP-hard) problems, the actual instances one meets are simple or regular enough to be amenable to fast cognitive solution, either optimally or approximately. By the same token, even though our analysis seems to rely on unlimited capabilities by the cognitive system, in the real-life situations, which are mostly regular, the cognitive system can get by with pedestrian and plausible capabilities and fully satisfy the results of the kind of abstract analysis presented here.

The level of abstraction that we use may nonetheless raise concerns among psychologists and cognitive scientists. For example, in using the very notion of algorithmic, or Kolmogorov, complexity, introduced below, we will be considering a quantity that is uncomputable. Assuming that the cognitive system is limited to the computable, this implies that Kolmogorov complexities are not, in general, calculated by people (or for that matter by computers). Nonetheless, it turns out that the general theoretical framework can be 'scaled-down' to provide computationally concrete and useful computational and statistical methods (Gao, Li, &

Vitányi, 2000; Rissanen, 1986, 1989; Wallace & Boulton, 1968; Wallace & Freeman, 1987), and that some of these have been used as the basis for models in cognitive science (Brent & Cartwright, 1996; Goldsmith, 2001; Pothos & Chater, 2002).

This situation seems to us a common one in science. If we are interested in gaining insight into some complex system, it is common to attempt to formulate a radical (and knowingly unrealistic) idealization, that one hopes capture the minimal assumptions needed to make theoretical progress. In filling out more detailed and concrete models that will ultimately be required to apply the account to specific circumstances, one hopes that, at least reasonably often, the general findings from the simplified general case will hold good. Almost invariably, while the level of formal rigor in the highly abstract analysis may be reasonably high, there will be little in the way of rigorous formal justification that the approach will still work well, even when its sweeping assumptions are replaced with a more detailed concrete model. For example, in physics, highly complex processes are routinely modelled with deliberately simplified assumptions—e.g., the Ising model, models of laser function, or almost any model in the 'complexity theory' in physics (Bak, 1997; Cartwright, 1983). Equally, in economics, human behaviour is routinely modelled as an aggregate of decisions by agents following the prescriptions of decision theory, which is computationally intractable, requires an unrealistic amount of information to be available to the reasoner, and which is known to be a poor model of the psychology of human decision processes (Akerlof & Yellen, 1985; Friedman, 1953; Kreps, 1990; Simon, 1959). Nonetheless, by making such assumptions, these disciplines, and science in general, have been able to make non-trivial progress. We hope that the high level of abstraction involved in the analysis given may at least potentially support such progress—but we grant that the degree to which these results can ultimately be scaled-down to build theories that make detailed psychological predictions about similarity and confusability (rather than merely capturing high level generalizations, as here) remains a project for further research.

1.2. The Universal Law of Generalization

Although intended to have broader application, the Universal Law of Generalization is primarily associated with a specific experimental paradigm—the identification paradigm. In this paradigm, humans or animals are repeatedly presented with stimuli concerning a (typically small) number of items. We denote items as $a, b, \ldots$, the representations of the corresponding perceptual stimuli as $S_a, S_b, \ldots$, and the representation of the corresponding responses as $R_a, R_b, \ldots$. So, for example, suppose that the experimental paradigm requires identifying English phonemes. Then the representations $a, b, \ldots$, stand for the representation of the individual phonemes of English. The representations $S_a, S_b, \ldots$ stand for representations of the specific perceptual stimuli associated with these phonemes-e.g., the acoustic and/or phonetic representations of the particular instantiations of those phonemes that are used in the experiment. Finally, the representations $R_a, R_b, \ldots$ encode the responses (which might be vocal or manual, depending on the experimental set-up) corresponding to each type of phoneme. We assume that in every specific situation there is an appropriate stimulus–response space.

In the identification paradigm, experimental participants are required to associate a specific, and distinct, response with each item—a response that can be viewed as 'identifying' the item concerned. The stimulus S_a is associated with item a and is supposed to evoke response R_a. With some probability, stimulus S_a can evoke a response R_b with $b \neq a$. This means that item b is 'confused' with item a, although the use of this term is purely descriptive. We leave open, for now, the question of whether these responses arise from confusion of perception, or memory, or through deliberate generalization from one item to another.

The matrix of $\Pr(R_a|S_b)$ values is known as the *confusability matrix*. In these terms, the Universal Law can be written as

$$\Pr(R_a|S_b) \text{ is proportional to } e^{-d(a,b)},$$

although we shall see below that the precise formulation is somewhat more complex. The law is not straightforward to test, because psychological distance $d(\cdot, \cdot)$ can only be inferred by indirect means. Moreover, even for the simplest sets of stimuli, such as pure tones differing in frequency, the nature and even existence of the corresponding internal psychological space, in terms of which distance can be defined, is highly controversial. Shepard has, nonetheless, provided an impressive case for the universal law.

1.3. The empirical case for the Universal Law

Shepard has shown that the technique of non-metric multidimensional scaling, of which he is a pioneer, can be used to derive an underlying metric psychological space from the confusability data itself. Specifically, the confusability data are used to derive a rank ordering of the distances between items on the basis of the relations between corresponding stimuli and responses (imposing certain assumptions, for example, to ensure that the 'distance' between two points is symmetrical, so that for all a, b, we have $d(a,b) = d(b,a)$. This rank ordering is fed into a non-metric multidimensional scaling procedure, which aims to find a way of embedding the items in a low dimensional Euclidean space. The goal is that the rank ordering of distances between the points should

correlate as well as possible with the rank ordering of confusabilities between items. The underlying rationale for this procedure is that the embedding of the items in a low-dimensional Euclidean space can be viewed as a model of the underlying psychological space used by the experimental participants. The probability with which two items are confused will be determined by the distance between them in this psychological space—the closer together they are, the more likely they are to be confused with each other. Given that we have a model of the putative psychological space, and hence can measure the distance $d(a, b)$ between items in that space, we can therefore assess whether the rate at which the probability of confusion decays is a negative exponential of the distance between that pair of items in psychological space, as stated by the Universal Law.

Shepard has amassed a large and diverse body of empirical data that, when analysed in this way, are consistent with the Universal Law. The diverse set of data that conforms to the law includes confusions between linguistic phonemes (Miller & Nicely, 1955), sizes of circles (McGuire, 1961), and spectral hues, in both people (Ekman, 1954) and pigeons (Guttman & Kalish, 1956), and spatial generalization by honeybees (Cheng, 2000). This evidence builds an impressive case for the Universal Law. There are, though, a number of points on which the case might be challenged.

The first challenge concerns the 'universality' of the Universal Law. We have already noted that, with few exceptions, there is little direct evidence that the Universal Law holds for stimuli whose confusability matrices cannot be readily embedded into a Euclidean space. As we have said, we set this issue aside throughout this paper, and assume that the Universal Law does apply in general. But, perhaps more worrying, is that there appear to be large numbers of data sets (Nosofsky, 1985, 1988a, 1988b) from identification paradigms which provide prima facie counterexamples to the Universal Law. Specifically, in these cases, confusability appears to be a Gaussian, rather than a negative exponential, function of psychological distance:

$\Pr(R_a|S_b)$ is proportional to $e^{-d(a,b)^2}$.

Indeed, the Gaussian generalization function is so successful empirically that it is central to a widely used class of exemplar models of categorization (Nosofsky, 1985; Medin & Schaffer, 1978).

The empirical picture is complex, but Shepard (1987) notes that one plausible reconciliation of the Universal Law with apparent examples of Gaussian confusability is that the Gaussian confusability originates from problems of perceptually distinguishing the stimuli (indeed, stochastic measurement errors almost always have a Gaussian distribution), whereas the Universal Law applies when perceptual discrimination is not the limiting factor in performance. From the present

perspective, then, the Universal Law applies where the determining factor in confusability might be confusion of representations in memory; or, from Shepard's interpretation of the data, the Universal Law applies where the critical variable may be judgments concerning what we calls 'consequential regions.' Elaborating on this viewpoint, Ennis (1989) provides a useful mathematical analysis of how perceptual noise might interact with non-perceptual confusability in accordance with the Universal Law.

A second possible challenge concerns the difficulties of curve-fitting. Comparing different classes of model for fit with a set of data is a controversial and subtle matter and fits are frequently surprisingly inconclusive, even when very large sets of data are available (Myung, Forster, & Browne, 2000b). Moreover, Myung, Pitt, and colleagues (Myung, Balasubramanian, & Pitt, 2000a; Myung & Pitt, 1997; Pitt, Myung & Zhang, 2002) have recently argued that comparisons between models are frequently systematically biased because one class of models is less restrictive than the other with respect to the class of data sets that it can model. This can lead to the counterintuitive consequence that, using standard statistical methods, one may be likely to conclude that the data were generated by model class A rather than B, irrespective of whether it was generated by model class A or B. Fortunately, however, the exponential fares relatively well from the point of view of this kind of analysis, at least in relation to the natural comparison with the power law which, in this context, would hold that for some positive constant c,

$\Pr(R_a|S_b)$ is proportional to $d(a, b)^{-c}$.

One would expect to be able to distinguish the two possibilities for large sample sizes. As far as we are aware, though, recent model comparison techniques such as those suggested by Myung and Pitt have not been applied to confusability data.

A third possible concern, which we have touched on above, is that pinning down the structure of internal psychological spaces is a notoriously difficult matter, and one that can be tackled from a range of theoretical perspectives, differing from that which Shepard adopts. Indeed, the problem of mapping magnitudes, such as sound pressure, onto a one-dimensional internal sensory scale (perceived loudness) has occupied the attention of psychophysicists for a century and a half without apparent resolution. Most famously, Fechner (1966/ 1860) Fechner (1966) argued for a logarithmic relationship between physical intensity and internal magnitude, whereas Stevens (1961) argued for a power-law relationship. Not all theorists will be confident in relying on non-metric multidimensional scaling of confusability matrices as the solution to all these difficulties (see Falmagne, 1986, for a review of the complexities of this area). Moreover, some theorists have even

doubted the coherence of internal scales of any form (Laming, 1997).

A fourth kind of concern, and one which the present paper seeks to address, is that it is not clear whether or how distance metrics can be applied at all to representations that the cognitive system may use for many kinds of complex stimuli. It is typically assumed that the cognitive representation formed of a visually presented object, a sentence or a story, will involve *structured* representations (e.g., Biedermam, 1987; Fodor, Bever, & Garrett, 1974; Fodor & Pylyshyn, 1988; Marr, 1982; Minsky, 1977; Schank & Abelson, 1977; Ullman, 1996). Structured representations can describe an object not just as a set of features, or as a set of numerical values along various dimensions, but in terms of parts and their interrelations, and properties that attach to those parts. Thus, in describing a bird, it is important to specify not just the presence of a beak, eyes, claws, and feathers, but the way in which they are spatially and functionally related to each other. Equally, it is important to be able to specify that the beak is yellow, the claws orange and the feathers white—to tie attributes to specific parts of an object. Although many distance metrics between structured representations can be envisaged (and we shall describe just such a metric below), it may appear that each metric will have to be specifically tailored to the particular stimuli concerned, and moreover, that there may be no non-arbitrary way of deciding on an appropriate metric for particular classes of structured representation. Thus, the principles for determining the similarity between the appearance of different kinds of bird, different Shakespearian speeches, or different court cases, might appear necessarily beyond a single style of analysis. And without an agreed metric for determining distance between items, we cannot even apply the Universal Law, let alone empirically confirm it.

This line of argument raises the possibility that the Universal Law may be restricted in scope to stimuli which are sufficiently simple to have a simple multi-dimensional representation—perhaps those that have no psychologically salient part-whole structure. We shall argue, however, that the Universal Law may nonetheless be applicable quite generally, since all these aspects are taken into account by the algorithmic information theory approach. This leads to a more generalized form of the Universal Law, as well as to a mathematically more appealing and less arbitrary form. We will suggest below that the universal cognitive distance that we describe can serve as an appropriate metric for comparing representations of all kinds, including structured representations.

The fifth, and final, concern that we consider is more fundamental. This is that the Universal Law presupposes that the confusability between items can be properly captured by a metric. A metric must have three properties (we state these precisely below): distances must be symmetrical (the distance from Warwick to Amsterdam is the same as the distance from Amsterdam to Warwick); distances must obey the triangle inequality—the distance from Warwick to Amsterdam via Paris must be no greater than the direct distance between Warwick and Amsterdam; and distances must obey the identity axiom, which implies that the distance between Warwick and Warwick is 0.

Do these properties hold in relation to confusability data generated in psychological experiments? Each has been challenged as unjustified, from a psychological point of view (Tversky, 1977)—the strongest challenges have concerned symmetry and the triangle inequality. Let us consider first the case of symmetry. The problem for this axiom is that, in the psychological data, there seem to be genuine and systematic asymmetries across many ways of measuring confusability. For example, complex things tend to be confused with simple things; but simple things are less often confused with complex things. Thus, people misremember complex shapes and simple, wobbly street plans in terms of straight lines and right angles, unusual colors in terms of 'focal colours' (e.g., 'mauve' becomes 'bright red') (Bartlett, 1932). For now, we shall leave this systematic asymmetry in confusability aside, although we note that it may be explicable within the mathematical framework that we describe below. Roughly, we shall suggest that an asymmetry may arise because it is, in a precise sense, easier to delete rather than to add, information in memory. Note that, while of considerable theoretical interest, these concerns over symmetry are not immediately relevant in the context of the empirical data relevant to the Universal Law, because symmetry is enforced by the nature of the data analysis, as we shall see below. Hence, we set concerns about symmetry aside henceforth.

The triangle inequality can also be undermined by examples of the following kind. Before the fall of communism, people typically judged Russia and Cuba to be highly similar counties; and moreover people typically judged Cuba and Jamaica to be highly similar. Hence it appears that Russia and Cuba should lie nearby in psychological space, as should Cuba and Jamaica. But Russia and Jamaica are typically judged to be highly dis-similar—and hence are presumed to be very distant in psychological space. So, in psychological terms, it may appear that the direct 'distance' between Russia and Jamaica, via Cuba, is shorter than the distance between Russia and Jamaica direct—and hence that the triangle inequality is violated.

As Peter van der Helm (personal communication) has pointed out to us, there are elegant perceptual analogs of this kind of case, which help clarify the origin of this kind of phenomenon, illustrated in Fig. 1.
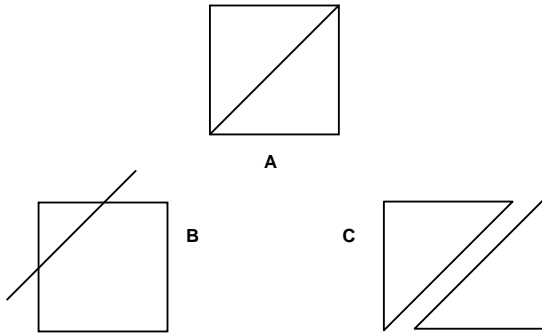
Fig. 1. Cognitive example apparently violating triangle inequality.

We see that A and B appear highly similar. A can be transformed into B, or vice versa, simply by a translation of the diagonal line. Equally, A and C appear highly similar. A can be transformed into C, or vice versa, simply by a translational movement of one of the triangles, relative to the other. But B and C appear to be quite dissimilar. It appears that the route from B to C via A may be shorter, in psychological terms, that the direct route from B to C, in contravention of the triangle inequality.

A possible response to this kind of case is that the case against the triangle inequality is still unproven, because it is not clear how the intuitive sense of similarity maps on quantitatively to psychological distance. If the relationship is highly non-linear, then the apparent counterexamples might break down, when expressed in terms of psychological distance, rather than in terms of intuitive similarity. But this response, while defensible, seems somewhat unsatisfactory, because there appears to be a systematic cognitive force at work, which stands in need an explanation. Specifically, where a stimulus can be represented in more than one way, the choice of 'foil' stimuli can bias the perceptual system towards one interpretation or the other. This bias then modifies judgments of similarity (and hence of psychological distance). Thus, when Cuba is presented in the context of Russia, the cognitive system focuses on political attributes; by contrast, when Cuba is presented in the context of Jamaica, the cognitive system focuses on geographical and cultural attributes (more generally, factors that are 'aligned' between two items tend to receive more attention than factors that are not aligned (Goldstone, 1990; Goldstone, Medin & Gentner, 1991; Hahn, Chater, & Henley, 1996). But this suggests that the apparent violation of the triangle inequality can be explained merely by assuming that the representation of the ambiguous item is different, depending on the comparison stimulus. Then the apparent violation of the triangle inequality vanishes. To continue the geographical analogy, it is simply not puzzling that Warwick is near Cambridge and Boston is near Cambridge, but that Warwick and Boston are very distant,

once one realizes that there are two distinct places, Cambridge, Massachusetts, and Cambridge, England.

This line of explanation relies, of course, on the assumption that similarity (and presumably confusability data) is defined over *interpreted* images, rather than, for example, images represented as raw arrays of pixels. Only from this point of view does it makes sense to say that the same visual input can be assigned to very different representations, which may have very different properties when entered into a similarity comparison, or in relation to confusability. A satisfactory cognitive theory of confusability, judged similarity, or related notions would, of course, be required to specify a particular level and style of representation as a starting point for an analysis. The fact that similarity is defined not over 'raw' perceptual inputs but rather over representations is, of course, common to any theory of similarity. One advantage of the present approach, as we have noted, is that the same formal machinery can apply over representations of any kind (rather than being limited to, say, spatial locations or bundles of features). This means that we can provide a formal analysis that has the potential to generalize over aspects of similarity and confusability involving different kinds of representations; and we can proceed without being committed to specific, and potentially controversial, representational assumptions.[3,4]

1.4. Theoretical perspectives on the Universal Law

We have so far described the Universal Law as capturing regularities in the confusability between stimuli, in a particular experimental set-up, the identification paradigm. The term 'confusability' is standard in this area, but masks a crucial theoretical issue: are cases where the response does not identify the stimulus appropriately viewed as 'mere' error, as the term 'confusability' suggests; or should they, rather, be viewed as deliberate acts of generalization?

To clarify the difference, imagine that you are presented with an identification paradigm, where the stimuli are colored blobs, and the responses are color words from a language that you do not know. Suppose that a pale red shade has previously been labelled BLIB;

---

[3] In particular, we do not have to make any specific assumptions about the 'space' of stimuli and responses in our analysis below—or even whether the representational format of these stimuli and responses is usefully thought of as contained within a space at all.

[4] Note, also, that allowing that information distance operates on interpreted images, rather than raw images, does not in any way imply that algorithmic information theory may not have an important role to play as a theoretical framework for understanding how raw images are mapped to interpretations. In particular, it may be that, in accordance with the 'simplicity principle' in perception, discussed below, that the interpretation that provides the shortest encoding of the 'raw' data is preferred (Chater, 1996), or some variant of this idea (van der Helm, 2000).

and a bright, focal red has previously been labelled GILP. You are then presented with the pale red shade again, and respond GILP. The 'confusion' interpretation of this response is that you simply made a mistake—you should have said BLIB, but mis-remembered (or mis-perceived) the relevant stimuli or responses (or the association between them). But the other 'generalization' interpretation is that your perceptual and memory processes are entirely intact—it is simply that you have generalized from focal red being GILP to other red colors also being GILP. On this interpretation, you might very well be able to report that the pale red shade can also be called BLIB, and even that this is how it is labelled in the training that you have received; but you nonetheless chose to use a different, generalized, response that you also believe to be appropriate.

Clearly, the confusion and generalization interpretations are not mutually exclusive: making a deliberate attempt to generalize is quite compatible with the additional influence of memory or perceptual lapses. A key question is which factors are most influential.

Shepard (1987) suggests that confusability data may often arise principally from deliberate generalization. Thus, for Shepard, a critical question arises concerning the degree to which such generalization is *justified*. He provides an elegant mathematical derivation that shows that, given specific assumptions, items that are represented as nearby in psychological space can be expected to have similar properties. Moreover, Shepard is able to derive the conclusion that the probability that such generalization is correct decays exponentially with distance in psychological space, just as exhibited in empirically collected confusability data. It is from this basis of viewing confusability as arising from deliberate generalization that Shepard describes his law as the Universal Law of Generalization, rather than, for example, the Universal Law of Confusability. This type of analysis has been followed up in more recent work (Myung & Shepard, 1996; Tenenbaum & Griffiths, 2001).

We suggest, however, that the emphasis on generalization may not always be appropriate. One reason is that, in most experimental paradigms, human participants are instructed to follow identification instructions. Hence, to indulge in deliberate generalization would seem to directly flout what appear to be clear experimental instructions. Moreover, the experimental tasks from which confusability data are collected typically generate substantial numbers of errors, even if participants do attempt to follow the instructions correctly. As the personal experience of anyone unlucky enough to have spent many hours performing identification judgments will testify, such tasks are typically quite challenging, and cannot be performed perfectly. This suggests that plain error is certainly one factor in

confusion responses; and perhaps in some settings it is the dominant factor.

Let us suppose, then, for a moment, that errors and generalization both play a role in generating confusion responses, in the kinds of experimental paradigms that have been widely used to test the Universal Law (the relative importance of the two factors might be quite uneven). Suppose that some participants can be persuaded to follow the task instructions reasonably closely, and not engage in unsolicited generalization—then we should obtain a relatively pure data concerning genuine errorful confusability, and effects of generalization should be eliminated. Does the Universal Law hold for such data sets? We suggest that the answer must be 'yes' because so many data sets for which the law has been tested seem likely to approximate such 'pure' conditions reasonably closely.

What would happen, instead, if participants were instructed not to feel obliged to make the response that identified the stimulus during the training phase of the experiment, but to engage in deliberate acts of generalization, whenever they felt appropriate? This would be the crucial empirical test for the Universal Law, as a law of *generalization*. Experimental conditions of this kind have not, to our knowledge, been tried. But there is circumstantial evidence that suggests that the resulting behaviour might not be highly lawful. The reason to suspect this is that, across many experimental paradigms, where performance is mediated by deliberate decision, performance tends to be highly variable in tasks that appear quite ill-defined from the point of view of participants. This may be, perhaps because people deploy all manner of strategies and background knowledge in a flexible and unpredictable way. For example, Stewart and Chater (2002) investigated generalization to novel stimuli intermediate between two categories that differ in variability. The effect of the variability of the categories differed greatly between participants—some participants classified intermediate stimuli into the more similar, less variable category, others classified the intermediate stimuli into the less similar, more variable category. Further, altering the variability of the training categories had large effects on individual participants' generalization. When the difference in variability between the two categories was increased, some people increased generalization to the more variable category, and some increased generalization to the less variable category. Existing exemplar (e.g., Nosofsky, 1988a) and parametric/distributional (e.g., Ashby & Townsend, 1986) models of generalization in categorization cannot predict the large variation between participants. This individual variation in performance suggests that there may be no single law governing human generalization, and therefore that performance may not fit easily into a lawful theoretical analysis, although it is too early to draw firm conclusions on this issue. If this is correct, we

might expect that, empirically, the Universal Law may apply more accurately to confusions resulting from genuine error than from confusions resulting from deliberate generalization.

In view of these considerations, here we shall not treat confusion data as representing generalization; and hence we shall not attempt to provide any justification for such generalization. Instead, we shall adopt a neutral framework, which makes quite minimal assumptions concerning the relationship between representations of stimuli and representations of the corresponding responses. We stress, though, that while we suspect that confusions may be the main factor in the identification paradigm data that has primarily been used to support the Universal Law, we expect that these confusions are typically due to lapses in memory, rather than to the difficulties in perceptual discriminability. Indeed, as we noted above, for stimuli where perceptual discrimination appears to limit performance, generalization typically follows Guassian, rather than exponential, decay.

2. Mathematical preliminaries

Shepard's (1987) article raises the question whether psychological science has any hope of formulating a law that is comparable in scope and possibly accuracy to Newton's universal law of gravitation. The Universal Law of Generalization for psychological science is a tentative candidate. In the *Principia* (Newton, 1687), Newton gives a few rules governing scientific activity. The first rule is 'We are to admit no more causes of natural things than such as are both true and sufficient to explain the appearances. To this purpose the philosophers say that Nature does nothing in vain, and more is in vain when less will serve; for Nature is pleased with simplicity, and affects not the pomp of superfluous causes.' Here, we generalize the 'Universal Law of Generalization' by essentially using Newton's maxim.

We have noted that the empirical analysis of internal psychological spaces from experimental data has proved extremely contentious. Here, we take a complementary approach and derive the Universal Law from first principles using the novel notion of the information contents of individual objects. That is, we motivate a measure of distance between representations of objects on a priori grounds, drawing on recent advances in Algorithmic Information Theory (a mathematical theory based on Kolmogorov complexity, Li & Vitányi, 1997). It turns out that there is a very natural, and general, measure of the 'distance' between representations, of whatever form: the information distance. Using this very general measure, the Universal Law of Generalization still holds, subject to quite minimal

restrictions on the process by which the experimental participant maps stimuli onto responses in the identification paradigm.

The presentation of this section has three parts. First, we provide some general background and also describe some basic results in Kolmogorov complexity theory. Second, we introduce and motivate the notion of 'information distance,' which we shall use as a fundamental measure of psychological distance. Third, we consider the nature of the probabilitistic process by which the participant maps stimuli to responses, which generates the confusion matrix in the identification paradigm—we shall need to make only very weak assumptions about this probabilistic process. In the next section, we show that, given these notions, the Universal Law holds: confusability is a negative exponential function of distance between representations.

2.1. Algorithmic information theory

This subsection gives general background concerning algorithmic information theory (also known as Kolmogorov complexity theory). We begin by providing a very brief intuitive sketch, and relating algorithmic information theory to similar ideas that have a long history in perceptual psychology. We then run through the formal machinery required to outline the elements of the theory that we shall draw on below.

We begin, then, with the core idea: algorithmic information theory provides a measure of the complexity of an individual (formal) object. Roughly, the complexity an object is given by the length of the shortest (effective) description or computer program that generates that object—the precise meaning of this is described below. This has led to a rich mathematical theory of simplicity, that has been used, in particular, as a foundation for inductive inference (Solomonoff, 1964, 1978). According to this approach, inductive inference involves finding the shortest description of, or program that generates, the available data. This is a formal version of an idea with a long intellectual history, from the Greeks, through William of Ockham, to, as we have seen, Newton (Li & Vitányi, 1997).

From the point of view of the psychological and cognitive science literatures, a particularly interesting advocate of this viewpoint was the physicist, philosopher, and perceptual theorist Mach (1959/1886) Mach (1959). According to Mach, the goal of science and the goal of perception is the same: providing an economical (i.e., brief) description of sensory data. Mach's views on the philosophy of science can be seen as feeding into the intellectual project of attempting to build a formal inductive logic for scientific inference, which requires a formal notion of simplicity (Carnap, 1952; Kemeny, 1953). This project was a key source of impetus for the development of algorithmic information theory

(Solomonoff, 1964). But Mach's views on perception can also be viewed as feeding into a rich tradition in perceptual theory, running through Gestalt psychology (e.g., Koffka, 1962/1935) to the present day. This tradition argues that perception is governed by a *simplicity principle*: this is the hypothesis that the perceptual system chooses between the multiple possible interpretations of a perceptual stimulus by choosing the *simplest*.

This long tradition of perceptual research has had a parallel evolution to the formal work on algorithmic information theory that we use in this article. Whereas the focus in algorithmic information theory has been on establishing and applying a very general measure of simplicity/complexity, the focus in perceptual research has been focussed on devising particular coding schemes for specific kinds of experimental stimuli, to make the simplicity principle concrete, and to allow it to be subject to direct empirical test. Specifically, there have been a great number of proposals for encodings of patterned sequences, where perceptual complexity is then defined in terms of those encodings (see for example the survey in Simon, 1972). This general approach has been applied in a variety of contexts, from the organization of simple sequences (Leeuwenberg, 1969, 1971; Restle, 1970; Simon, 1972; Simon & Kotovsky, 1963; Vitz & Todd, 1969), to judgments of 'figural goodness' (Hochberg & McAlister, 1953), the analysis of (Johansson, 1950) experiments on the perception of motion configurations (Restle, 1979), and figural completion (Buffart, Leeuwenberg, & Restle, 1981). It has also been advanced as a general framework for understanding perceptual organization (Attneave & Frost, 1969; Leeuwenberg, 1971; Leeuwenberg & Boselie, 1988), and even cognitive processes more generally (Chater, 1997, 1999). Although, as we have noted, this strand of perceptual research has developed independently of the mathematical methods of algorithmic information theory that we shall describe and employ below, recently, some initial formal connections have been developed between the two approaches (Chater, 1996; van der Helm, 2000).

Now let us turn to outlining the formal core of algorithmic theory. We have said that algorithmic information theory aims to measure the complexity of individual formal objects. This immediately raises the question of what kind of formal objects we are concerned with. We take the viewpoint that the set of objects we are interested in may be finite or infinite but must be countable, just like the natural numbers. We also assume that each such object can be described by using, for example, English. That means we can describe every object by a finite string in some fixed finite alphabet. By encoding the different letters of that alphabet in bits (0's and 1's) we reduce every description or representation of the object to a finite binary string. A
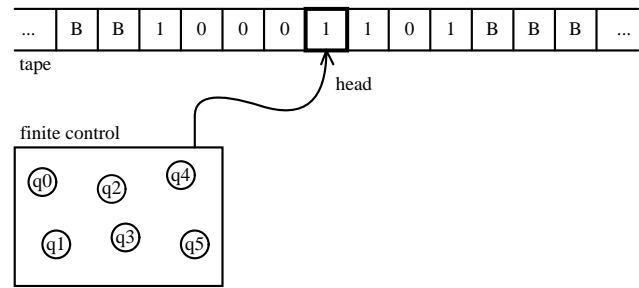

Fig. 2. Turing machine.

similar argument presumably holds for the physical manner by which an object is represented in an agent's cognitive system. In this way we reduce the representation of all objects that are relevant in this discussion to finite binary strings. In the unlikely case that there are relevant objects that cannot be so represented, we simply agree that they are not subject of this discussion.

From the abstract mathematical perspective used in this article, we need not consider the details of the various perceptual codes that have been proposed, and mentioned above: we note merely that all these codes are special types of computable codes, which means that all of them can be decoded by appropriate machines or programs. Mathematically, one says that every such code can be decoded by an appropriate *Turing machine*: a convenient model introduced by A.M. Turing in a celebrated paper (Turing, 1936) to formally capture the intuitive notion of 'computation' in its greatest generality.

How does a 'Turing machine' work? Fig. 2 (taken from Li & Vitányi, 1997) shows a Turing machine, which consists of a finite program, called a *finite control*, capable of manipulating a linear list of *cells*, called a *tape*, at a particular location, called the *head*. The device can write and delete cells at the current location of the head; and it can shift the head one step at a time along the tape. The device follows a *list of rules*, which determine from the current state of the finite control and the symbol contained in the cell that is currently being scanned. The rules determine the next operation of the finite control (e.g., shifting the head, or writing or deleting a symbol), given only the current state of the control and the symbol on the cell of the tape that is currently being scanned. The possible rules are restricted so that the behaviour of a Turing machine is deterministic: a particular computational state of the machine can only continue in a specified way. One state of the control can be designated as the 'halt' state, from which no further operations can be performed. If a Turing machine halts, the symbols on the tape encode the result of its computation; if it does not halt, and continues computing indefinitely, the computation is viewed as having no determinate result.

It has turned out that all different mathematical proposals to formulate a more general notion of computability turned out to be equivalent to the Turing machine. The so-called *Church-Turing thesis* states that the Turing machine captures the most universal and general notion of effective computability, and is the formal equivalent of our intuitive notion of what is calculable. The Turing machine model of computation is universally used in formal arguments. There is no need to go into details here, as they can be found in any textbook on computable functions and effective processes, for example, Odifreddi (1989) or Section 1.7 in Li and Vitányi (1997). What is important here is that there is a general code that subsumes all computable codes mentioned above. This is the code decodable by a so-called 'universal' Turing machine. In effect, such a machine works with a code book that enumerates all computable codes. By prefixing an encoded item with the index in the enumeration of the particular code that has been used, the universal Turing machine can decode that item. Clearly, this universal encoding need not be longer than the shortest two-part code consisting of the index of a particular code used plus the length of the resulting encoding. Stating that the universal code can be decoded by a universal Turing machine (which is assumed to have an infinite tape) is equivalent to that it is a program in a universal programming language like C++ or Java (running on a machine with no unlimited memory). For example, we can now formulate the length of the shortest code for Tolstoy's *War and Peace*. These considerations led the Russian mathematician Kolmogorov (1965) to propose a general theory of the information content in individual objects. In this paper we keep the discussion informal; an introduction, epistimology and rigorous treatment of the theory is given in Li and Vitányi (1997).

The *Kolmogorov complexity* $K(x)$ of a finite object x, is defined as the length of the shortest binary computer program that produces x as an output.[5] Thus, objects such as a string of one billion '1's, or a binary code for a digitized picture of an untextured rectangle, or the first billion digits of $\pi = 3.1415...$ are reasonably simple, because there are short programs that can generate these objects. On the other hand, a typical binary sequence generated by tossing a coin is complex—the sequence is its own shortest program, because there is no hidden

structure in such a sequence that can be used to find a shorter code. The Kolmogorov complexity is an absolute measure of the amount of information in an individual object. It has been applied to resolve a long-standing debate on the proper definition of individual random sequences and an objective formulation of the notion of 'simplicity' in the inductive principle known as 'Occam's Razor.' Kolmogorov complexity contrasts with standard (probabilistic) information theory (Cover & Thomas, 1991), which is only concerned with the average information of a random source. This algorithmic notion of information should be contrasted with Shannon's statistical notion of information (Shannon, 1948; Cover & Thomas, 1991), from standard (probabilistic) information theory, which deals with the average number of bits required to communicate a message from a random source.

From a psychological point of view, Kolmogorov complexity may seem unsatisfactory, because it takes no account of whether an extant, in principle computable, regularity will in fact be detected by the limited computing power, or evolutionary bias of attention, of the cognitive system. For example, an array of pixels that corresponded to a binary encoding of π would have a low Kolmogorov complexity, because π can be generated by a short program; and a random array of pixel values would have a high Kolmogorov complexity, because, in virtue of its randomness, there would no short program that could generate it. But, from a perceptual point of view, these arrays of pixels would seem indistinguishable to most human observers. we shall see below, however, that the approach is still applicable, even when we are concerned with agents with limited abilities to find regularities.

The definition of Kolmogorov complexity may appear to be rather specific. But this appearance is misleading. For example, the restriction to a binary coding alphabet can easily be dispensed with—switching to an alphabet with n letters amounts merely to rescaling all Kolmogorov complexities by a multiplicative constant,[6] but has no other impact. The binary alphabet is used by convention, to provide a fixed measuring standard. More interestingly, it might appear that the length of the shortest program that generates a specific code must inevitably be relative to the choice of programming language. But a central result of Kolmogorov complexity theory, the Invariance Theorem (Li & Vitányi, 1997), states that the shortest description of any object is invariant (up to a constant) between different universal languages. Therefore, it does not matter whether the universal language chosen is C++, Java or Prolog—the length of the shortest description for each object will be approximately the same. Let us introduce the notation

---

[5] Strictly, it is important that the program is a prefix program—that is, that no initial segment of the binary string comprising the program itself defines a valid program; and, equally, that no non-trivial continuation of the binary string comprising the program defines a valid program. The restriction to prefixes ensures that, for example, given a binary string that corresponds to a concatenation of programs, there is no ambiguity concerning how the string should be divided into discrete programs. Although tangential to the discussion here, the use of prefix complexity is of considerable technical importance (Li & Vitányi, 1997; Vitányi & Li, 2000).

[6] Specifically, this constant is $\log n$. All logarithms in this paper are binary logarithms unless otherwise noted.

$K_{C++}(x)$ to denote the length of the shortest C++ program which generates object x; and $K_{Java}(x)$ to denote the length of the shortest Java program. The Invariance Theorem implies that $K_{C++}(x)$ and $K_{Java}(x)$ will only differ by some constant, c, (which may be positive or negative) for *all* objects x, including, of course, all possible perceptual stimuli. Formally, there exists a constant c such that for all objects x:

$$|K_{C++}(x) - K_{Java}(x)| \leqslant c.$$

Thus, in specifying the complexity of an object, it is therefore possible to abstract away from the particular language under consideration. Thus the complexity of an object, x, can be denoted simply as $K(x)$—referring to *the* Kolmogorov complexity of that object.

Why is Kolmogorov complexity language invariant? To see this intuitively, note that any universal language can be used to encode any other universal programming language. This follows from the preceding discussion because a programming language is just a particular kind of computable mapping, and any *universal* programming language can encode any computable mapping. For example, consider two universal computer languages which we call 'C++' and 'Java.' Starting with C++, we can write a program, known in computer science as a compiler, which translates any program written in Java into C++. Suppose that this program has length has length c_1. Suppose that we know $K_{Java}(x)$, the length of the shortest program which generates an object x in Java. What is $K_{C++}(x)$, the shortest program in C++ which encodes x? Notice that one way of encoding x in C++ works as follows—the first part of the program translates from Java into C++ (of length c_1), and the second part of the program, which is an input to the first, is simply the shortest Java program generating the object. The length of this program is the sum of the lengths of its two components: $K_{Java}(x) + c_1$. This is a C++ program which generates x, if by a rather roundabout means. Therefore $K_{C++}(x)$, the shortest possible C++ program must be no longer than this: $K_{C++}(x) \leqslant K_{Java}(x) + c_1$. An exactly symmetric argument based on translating between languages in the opposite direction establishes that: $K_{Java}(x) \leqslant K_{C++}(x) + c_2$. Putting these results together, we see that $K_{Java}(x)$ and $K_{C++}(x)$ are the same up to a constant, for all possible objects x. This is the Invariance Theorem, that Kolmogorov complexity is language invariant.

The implication of the Invariance Theorem is that the functions $K(\cdot)$ (and $K(\cdot|\cdot)$, that we introduce below), though defined in terms of a particular programming language, are language-independent up to an additive constant and acquire an asymptotically universal and absolute character through the Church–Turing thesis, i.e., from the ability of universal machines to simulate one another and execute any effective process. The

Kolmogorov complexity of a string can be viewed as an absolute and objective quantification of the amount of information in it, giving a rigorous, formal and highly general notion corresponding to our intuitive notion of shortest effective description length. This may be called *Kolmogorov's thesis*. This leads to a theory of *absolute* information *contents* of *individual* objects in contrast to classical information theory which deals with *average* information *to communicate* objects produced by a *random source*. Since the former theory is much more precise (there are no issues of quantities being defined only up to a constant, depending on the programming language chosen), it is perhaps surprising that analogs of many central theorems in classical information theory nonetheless hold for Kolmogorov complexity, although in a somewhat weaker form.

We have mentioned that shortest code length is invariant for *universal* programming languages. How restrictive is this? The constraint that a system of computation is universal turns out to be surprisingly weak—all manner of computation systems, from a simple automaton with under 100 states supplied with unlimited binary tape on which it can read and write, to numerous word processing packages, spreadsheet and statistical packages, turn out to define universal programming languages. Therefore, it seems that universality is likely to be obeyed by a computational system as elaborate as that involved in cognition.

The basic notion of Kolmogorov complexity has been elaborated into a rich mathematical theory, with a wide range of applications in mathematics and computer science. It has also been applied in a range of contexts in psychology, from perceptual organization (see Chater, 1996, or van der Helm, 2000, for different uses of the theory), to psychological judgements of randomness (Falk & Konold, 1997), to providing the basis for a theory of similarity (Chater & Hahn, 1997; Hahn, Chater, & Richardson, 2003). Indeed, the idea that cognition seeks the simplest explanation for the available data, inspired by results in Kolmogorov complexity, has even been suggested as a fundamental principle of human cognition (Chater, 1997, 1999).

2.2. Information distance

In this section, we introduce the notion of information distance, that will be central to the argument that we develop below. Informally, information distance may be viewed as the most general and most neutral measure of the cognitive distance between two items. It is maximally general, because it applies to items which can have any representational format whatever—roughly, information distance is determined by the length of programs that are required to transform the representation of one item into the representation of the other. This is clearly well-defined, whether the items

concerned are bundles of features, specifications of locations in a multidimensional space, syntactic trees, structured representations of an image, or whatever. On the other hand, most cognitive distance measures discussed in the psychological literature have much narrower scope—they are applicable only to a particular representational format. Information distance is also maximally neutral, as a measure of cognitive distance, in the following sense—if, according to any 'reasonable' cognitive distance (and we will see below that the specific assumptions concerning reasonableness are quite modest), two items are judged to be close, then these items will also be close according to information distance. That is, information distance captures similarities that are recognized by any reasonable cognitive distance—and hence is maintains a maximal level of neutrality between specific cognitive distances. In view of the fact that, from a psychological point of view, we frequently do not have a good analysis either of the structure of the underlying representations used by the cognitive system, or reason to favour a specific distance measure between such representations, information distance is an attractive notion. It allows us to make progress in analysing cognitive distance and related notions, without having to take firm stands on these contentious issues. Moreover, as we have already discussed, the generality of the arguments that we can frame using notions of this kind allows the possibility of providing a general explanation for phenomena such as the universal law, across many types of stimulus materials, rather than having to frame a more specific explanation for each case.

Kolmogorov complexity is defined for a single object, x. But an immediate generalization, conditional Kolmogorov complexity, $K(y|x)$ provides a measure of the degree to which an object y differs from another object x. $K(y|x)$ is defined as the length of the shortest program (in a universal programming language, as before) that takes x as input, and produces y as output. The intuitive idea is that if items are distant from each other, then it should require a complex program to turn one into the other. At this point one may wonder about how large $K(x)$ is? As stated before, we require that the set of programs be prefix-free, no program being a proper prefix of another one. This restricts the set of binary strings that are available as proper programs, and, therefore, has the consequence that for some strings x the shortest programs that describe them are somewhat longer than the literal description of x which has *length* $l(x)$, the number of bits in x. How bad can this get? It turns out that $K(x)$ can be as large as $l(x) + K(l(x))$. If we consider x as the integer described by the binary numeral '$1x$', then $l(x) \approx \log x$. Repeated substitution of the expression for the largest value of $K(x)$ shows that it can rise to about $l^*(x) = \log x + \log \log x + \log \log \log x + \cdots +$ (all positive terms ending with a small additive constant). It is not difficult to see that
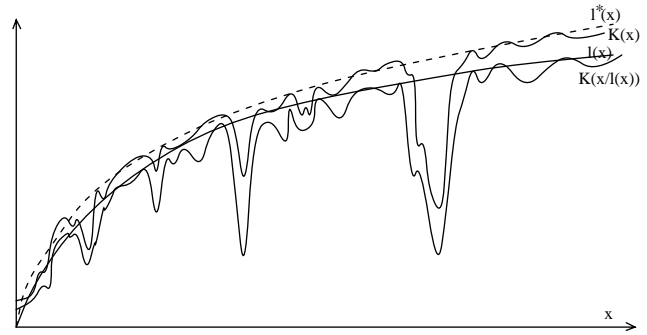

Fig. 3. The graph of the function $K(x)$ and $K(x|l(x))$—here we consider x to be the integer index of the string x in the length-increasing lexicographical order.

$K(x)$ never exceeds $l^*(x) < + 2 \log \log x$ by more than a couple of bits. On the other hand, the shortest program to compute x when it has $l(x)$ as input can be shown to be not larger than about $l(x)$, and therefore $K(x|l(x))$ is upper bounded by about $\log x$. In Fig. 3 (taken from Li & Vitányi, 1997) we have given an 'artist's impression' of the graphs of $K(x)$ and $K(x|l(x))$, viewed as integer functions. The graphs closely 'hug' the graph of $\log x$, for the random or irregular x's, but sometimes have deep indentations. Those indentations correspond to x's having lots of regularity. Note that $K(x)$ rises to infinity with x, both its upper bound which slightly exceeds $\log x$, and its lower bound that rises very, very slowly. In fact, while the bottoms of the dips eventually rise to infinity (because at some x_0 all programs of a given length $\leqslant k$ are in use for some $x \leqslant x_0$ and consequently $K(x) > k$ for all $x > x_0$) this increase is slower than that of any unbounded computable function. In contrast, the graph of $K(x|l(x))$ has dips down to a fixed constant level forever. For example, there is a fixed constant k such that, for every n with $x_n = 00\ldots0$ (a string of n zeros), we have $K(x_n|n) \leqslant k$.

We are now ready to turn to the notion of information distance itself. It is useful to recall the mathematical formulations of the notions of 'distance' and 'metric.' A *distance* function D with nonnegative real values, defined on the Cartesian product $X \times X$ of a set X, is called a *metric* on X if, for every $x, y, z \in X$:

- $D(x, y) = 0$ iff $x = y$ (the identity axiom);
- $D(x, y) + D(y, z) \geqslant D(x, z)$ (the triangle inequality);
- $D(x, y) = D(y, x)$ (the symmetry axiom).

(Recall that we informally discussed the import of these axioms in the context of psychological theories of similarity above). A set X provided with a metric is called a *metric space*. For example, every set X has the trivial *discrete metric* $D(x, y) = 0$ if $x = y$ and $D(x, y) = 1$ otherwise. All information distances in this paper are defined on the set $X = \{0, 1\}^*$ (that is, the set of all finite strings composed of 0's and 1's) and satisfy the metric

conditions up to an additive constant or logarithmic term while the identity axiom can be obtained by normalizing.

The conditional complexity function $K(y|x)$ trivially obeys identity, because no program at all is required to transform an item into itself.[7] Conditional complexity also obeys the triangle inequality: $K(x|z) \leqslant K(x|y) + K(y|z)$. This follows immediately from considering the concatenation of a program mapping z into y (with minimum length $K(y|z)$), and a program mapping y into x (with minimum length $K(x|y)$). Using this concatenation, it is clearly possible to map z to x using a program of length no more than the sum of these individual programs: $K(x|y) + K(y|z)$. This sum must therefore be at least as great as the length of the shortest program mapping from z to x, that is $K(x|z)$, where $K(x|z)$ is typically smaller, by there being shorter programs which perform this mapping without going through the intermediate stage of generating y. Thus, the triangle inequality holds for $K(\cdot|\cdot)$.

But, as it stands, $K(y|x)$ is not appropriate as a distance measure, because it is asymmetric. Consider the null string ε. $K(\varepsilon|x)$ is small, for every x, because to map the input x onto the null string simply involves deleting x, which is a simple operation. Conversely, $K(x|\varepsilon) = K(x)$, can have any value whatever, depending on the complexity of x. More generally, the length of program required to turn a complex object into a simpler object is shorter than the length of program required to turn the simple object back to the complex object. Interestingly, this systematic asymmetry appears to map, qualitatively, onto the empirically observed asymmetry in confusability, that we mentioned above: that a complex object is more likely to be confused with a similar, but simpler, object, than the reverse. This observation may be useful in applying Kolmogorov complexity to understanding confusability data, but we will not pursue it further here.

This is because, in the context of the present article, we need a measure that is symmetric, because, as we have already noted above, symmetry is enforced on the data against which the Universal Law is tested. Symmetry can be restored by, for example, taking the sum of the complexities in both directions: $K(x|y) + K(y|x)$, or alternatively, the maximum of both complexities $\max\{K(x|y), K(y|x)\}$. It is easy to verify that the resulting measures, known as *sum distance* and *max distance*, respectively, qualify as distance metrics (Bennett et al., 1998; Li & Vitányi, 1997). For example, the sum distance and the max distance between x and the

null string ε are given by $K(x|\varepsilon) + K(\varepsilon|x) = K(x) = \max\{K(x|\varepsilon), K(\varepsilon|x)\}$.

Max and sum-distances are close but not necessarily equal. Denoting sum and max distance respectively by D_{sum} and D_{max}, it is easy to verify that, for every x, y:

$$D_{max}(x, y) \leqslant D_{sum}(x, y)) \leqslant 2D_{max}(x, y). \qquad (1)$$

For the present purpose of putting the Universal Law on a formal mathematical footing, it is important to consider the epistimological motivation of these distances. The *information distance* is defined in Bennett et al. (1998) as the length of a shortest binary program that computes x from y as well as computing y from x. Being the shortest such program, it should take advantage of any redundancy between the information required to go from x to y and the information required to go from y to x. The program functions in a catalytic capacity in the sense that it is required to transform the input into the output, but itself remains present and unchanged throughout the computation. Note that while a program of length $K(x|y) + K(y|x)$ by definition can compute from y to x (a subprogram of length $K(x|y)$) and from x to y (a subprogram of length $K(y|x)$), it is by no means clear (and happens to be false) that such a program is necessarily the shortest that performs both the mapping from x to y and the mapping from y to x. A $(K(x|y) + K(y|x))$-length program is not minimal if the information required to compute y from x can be made to overlap with that required to compute x from y.

In some simple cases, *complete* overlap can be achieved, so that the same minimal program suffices to compute x from y as to compute y from x. We first need an additional notion. A binary string x of n bits is called *incompressible* if $K(x) \geqslant n$. A simple argument suffices to show that the overwhelming majority of strings is incompressible (Li & Vitányi, 1997). We continue the main argument. For example if x and y are independent incompressible binary strings of the same length n (up to additive constants we have $K(x|y), K(y|x) \geqslant n$), then their bitwise exclusive-or $x \oplus y$ serves as a minimal program for both computations. (If $x = 01011$ and $y = 10001$, then $z = x \oplus y = 11010$. Since $z \oplus y = x$ and $z \oplus x = y$ we can use z as a program both to compute from y to x and to compute from x to y.)

Similarly, if $x = uv$ and $y = vw$ where u, v, and w are independent incompressible strings of the same length, then $u \oplus w$ along with a way to distinguish x from y is a minimal program to compute either string from the other. Now suppose that more information is required for one of these computations than for the other, say,

$$K(y|x) > K(x|y).$$

Then the minimal programs cannot be made identical because they must be of different sizes. In some cases it is easy to see that the overlap can still be made complete,

---

[7] Note that, throughout, due to language invariance, Kolmogorov complexities are only specified up to an additive constant. So, in a particular language, $K(x|x)$ could be non-zero—if, for example, some instructions are required to implement the 'null' operation (this is typically true of real programming languages, in which the null string is not treated as a valid program). But the length of this program will, by language invariance, be bounded by a constant, for all possible x.

in the sense that the larger program (for y given x) can be made to contain all the information in the shorter program, as well as some additional information. This is so when x and y are independent incompressible strings of unequal length, for example u and vw above. Then $u \oplus v$ serves as a minimal program for u from vw, and $(u \oplus v)w$ serves as one for vw from u.

A principal result of Bennett et al. (1998) shows that, up to an additive logarithmic error term, the information required to translate between two strings can be represented in this maximally overlapping way in *every case*. That is, the minimal program to translate back and forth between x, y has length not larger than $\max\{K(x|y), K(y|x)\}$. It is straightforward that the minimum length program to do this back and forth translation cannot be shorter, since by the definition of Kolmogorov complexity, the translation in direction x to y requires a program of length at least $K(y|x)$ and the translation in the direction of y to x requires a program of length at least $K(x|y)$. Therefore, the length of the shortest binary program that translates back and forth between two items is called the *information distance* between the two items, and it is equal to $D_{\max}(x, y)$—to be precise, up to an additive logarithmic term which we ignore in this discussion.

Max-distance has a particularly attractive universal quality: it is, in a sense, the *minimal* distance measure, in a broad class of distance measures that might be termed 'computable,' as we now see.

We say that a function from a discrete domain to the reals (for example a distance metric) is *semicomputable from above* if it can be approximated from above by some computable process. This is a very weak condition. For example, it is weaker than the assumption that a function is computable. It requires merely that there is some computable process that outputs a sequence of successive approximations to the function value, that are successively decreasing, and which converge to be as close as desired to the distance metric, given sufficient computation.[8] If we assume the Church–Turing thesis, which implies that human cognition can encompass only computable processes, then it seems that this assumption follows automatically.

To make sense of the notion of a 'minimal' distance measure, we need some normalization condition, to fix the 'scale' of the distances. Without such a condition, we could simply divide the values given by a distance metric by an arbitrarily large constant c to get a more 'minimal' distance metric.

For a cognitive similarity metric the metric requirements do not suffice: a distance measure like $D(x, y) = 1$

for all $x \neq y$ must be excluded. For each x and d, we want only finitely many elements y at a distance d from x. Exactly how fast we want the distances of the strings y from x to go to ∞ is not important: it is only a matter of scaling. In analogy with Hamming distance in the space of binary sequences, it seems natural to require that there should not be more than 2^d strings y at a distance d from x. This would be a different requirement for each d. With prefix complexity, it turns out to be more convenient to replace this double series of requirements (a different one for each x and d) with a single requirement for each x:

$$\sum_{y:y \neq x} 2^{-D(x,y)} < 1.$$

We call this the *normalization property* since a certain sum is required to be bounded by 1.

We consider only distances that are computable in some broad sense. This condition will not be seen as unduly restrictive. As a matter of fact, only upper-semicomputability of $D(x, y)$ will be required. This is reasonable: as we have more and more time to process x and y we may discover more and more similarities among them, and thus may revise our upper bound on their distance. The upper-semicomputability means exactly that $D(x, y)$ is the limit of a computable sequence of such upper bounds.

Definition 1. An *admissible distance* $D(x, y)$ is a total nonnegative function on the pairs x, y of binary strings that is 0 if and only if $x = y$, is symmetric, satisfies the triangle inequality, is upper-semicomputable and normalized, that is, it is an upper-semicomputable, normalized, metric. An admissible distance $D(x, y)$ is *universal* if, for every admissible distance $D'(x, y)$, we have $D(x, y) \leqslant D'(x, y) + c_D$, where c_D may depend on D but not on x or y.

In Bennett et al. (1998), a remarkable theorem shows that $D_{\max}$ is a universal (that is, optimal in the sense of being minimal) admissible distance. Formally, every admissible distance metric D has an associated constant c such that

$$D_{\max}(x, y) \leqslant D(x, y) + c,$$

for every x and y.

As already discussed above, the universal distance $D_{\max}$ happens also to have a 'physical' interpretation as the approximate length of the smallest binary program that transforms x into y and vice versa. That is, for all the infinitely many x, y, and hence the infinite number of distances between them, the $D_{\max}$ distance is never more than a finite additive constant term greater than the corresponding D-distance with respect to any admissible distance metric D, where the additive

---

[8] It does not require, for example, that it is possible to actually output the 'correct' distance values—or, indeed, to announce the degree of approximation that has been achieved after a given amount of computation.

constant may depend on D, but is independent of x and y.[9]

Intuitively, the significance of this is that the universal admissible distance minorizes *all* admissible distances: if two pictures are d-close under some admissible distance, then they are a fortiori d-close up to a fixed additive constant under this universal admissible distance. That is, the latter discovers all effective feature similarities or cognitive similarities between two objects: it is the universal cognitive similarity metric. The remarkable thing about information distance measures such as D_{max} is that, with respect to the class of computable distance measures (subject to the normalization condition described above), they are minimal. That is, if any computable measure treats two items as near, then information distance measures will also treat the items as 'reasonably' near.

The typical distance measures considered in psychology, artificial intelligence or mathematics are *not* universal. This is because they favour some regularities among the items that consider, but entirely ignore other regularities—and some of these regularities may be the basis of computable (and hence allowable) distance measures. Let us look at some examples. Identify digitized black-and-white pictures with binary strings. There are many distances defined for binary strings. For example, the Hamming distance and the Euclidean distance. The Hamming distance between two n-bit vectors is the number of positions containing different bits; the Euclidean distance between two n-bit vectors is the square root of the Hamming distance. Such distances are sometimes appropriate. For instance, if we take a binary picture, and change a few bits on that picture, then the changed and unchanged pictures have small Hamming or Euclidean distance, and they do look similar. However, this is not always the case. The positive and negative prints of a photo have the largest possible Hamming and binary Euclidean distance, yet they look similar to us. Also, if we shift a picture one bit to the right, again the Hamming distance may increase by a lot, but the two pictures remain similar. As another example, a metric of similarity based on comparing overlap of features (Tversky, 1977) will treat items that have precisely opposite patterns of features as very distant. But, of course, with respect to the D_{max} measure such items are very close since the program saying 'take the opposite of every feature' suffices to change one item into the other. Hence, such a feature-based metric is not a minimal distance. Similarly, if items are represented as real-valued vectors, and the Euclidean distance metric is used, then items corresponding to vectors $\mathbf{v}$ and $2\mathbf{v}$ will have distance equal to the Euclidean length of $\mathbf{v}$, while D_{max} is small.

We believe that, in the present context, the minimality of information distance is a substantial virtue, because minimal distance measures make the least commitment to the specific similarity metric used by the cognitive system. This is because, if two items are similar according to any computable (strictly, upper-semicomputable) metric, then they are similar according to information distance; and we have assumed that the cognitive system is restricted to computable metrics. Thus, we stress that we adopt information distance as our measure not because we assume that the cognitive system uses information distance: clearly, it does not, because many computable regularities between stimuli are not apparent to the perceptual system, and indeed, the regularities that are readily apparent to the perceptual system appear to be quite limited (van der Helm & Leeuwenberg, 1996; Palmer, 1982; Wagemans, 1995). Instead, we adopt information distance because, as we have just described, it captures any relations of nearness that may actually be used by the cognitive system. That is, if any computable (strictly, upper-semicomputable) metric used by the cognitive system treats two objects as close together, then these objects will also be close together according to the information distance metric. Note, though, that the converse does not hold. Information distance may, however, put two objects close together, while the cognitive system, which cannot take all computable similarities into account, will put them at a distance. In sum, then, information distance can be viewed as an 'amalgam' of all computable distance metrics (and hence, as taking account of any psychologically plausible distance metrics); it therefore seems an ideal starting point for our analysis, because it appears to require the minimum in the way of specific psychological assumptions.

We have considered some technical reasons why measures based on information distance are attractive general distance measures. These measures gain some additional psychological interest because of their relation to the recently proposed Representational Distortion theory of psychological similarity (Hahn, Chater, & Richardson, 2001, 2003). According to Representation Distortion, the psychological similarity of two items depends on the complexity of the transformation required to 'distort' the representation of one of the items into a representation of the other item. The notion of complexity is then assumed to be related to the notion of conditional Kolmogorov complexity, as described here. According to this viewpoint, the flexibility of measures like information distance is appropriate because it reflects the flexibility of the cognitive system—to choose arbitrary ways of interrelating, aligning and connecting representations, rather than being constrained to use a fixed similarity measure. This account of similarity, although early in its development, has received some empirical support. For example,

---

[9] By (1, $D_{sum}(x, y) \leqslant 2D(x, y) + 2c$, because the two measures D_{max} and D_{sum} are within a factor of 2 of each other.

Hahn et al. (2001, 2003) constructed various stimuli, which could be transformed into each other via sequences of elementary operations. One set of stimuli, for example, were different configurations of children's 'lego' building bricks. The number of elementary transformations required to turn one stimulus into another was used as a crude measure of the complexity of the process of distorting the representation of one stimulus to the representation of the other stimulus. They found that (dis)similarity judgments for pairs of stimuli were strongly correlated with this measure of transformational complexity, in line with the predictions of the Representational Distortion account. To choose a very different example, an experiment by Reznikova and Ryabko (1986) (also reported in Li & Vitányi, 1997) on the information transmission rate and message-compressing capabilities of ants showed evidence that ants could communicate simple alternations of left- and right-turns in a maze, like LLLLLL or LRLRLR, faster than more random alternations. This seems to indicate that the ants compress the information before transmitting it. Similarly, as mentioned above, the subjective difficulty of learning a Boolean concept (that is, a formula in elementary mathematical logic) by humans appears to be directly proportional to the length of the shortest equivalent logical formula, Feldman, 2000), giving a more theoretical underpinning of a well-known stimulus classification of the order of empirical difficulty of learning Boolean concepts (Shepard, Hovland & Jenkins, 1961).

2.3. How items are confused

We assume a very general, and weak, model of similarity—based on information distance. We next need a general model of the how items are confused with each other. Fortunately, only a very weak assumption is required. First, we assume that there is a discrete set of items a, stimuli S_a, and responses R_a. Moreover, these are associated with one another in the sense that there is a fixed program that on input x computes y where x and y are choosen from among S_a, R_a. Secondly, for each stimulus S_a, the probability distribution $\Pr(R_b|S_a)$ over the different responses, S_b, is itself semicomputable from below. That is, it can be approximated from below by a computable process that produces a monotonically increasing series of approximations to $\Pr(R_b|S_a)$ which approach arbitrarily if closely, given sufficient computing time. This is a weaker condition, of course, than the condition that the probability distribution can be actually be computed exactly by some computable process—which is equivalent to the distribution being both semicomputable from above and from below. Recall that the celebrated Church–Turing thesis, see for example Odifreddi (1989), states that everything which is intuitively computable can be computed formally by a Turing machine, or, equivalently, a standard computer supplied with a large enough memory. Assuming the Church–Turing thesis implies that processes executed by the cognitive system are computable functions. In particular, therefore, this condition will include any computational account of the process by which stimuli S_a are mapped onto responses R_b.

3. The generalized Universal Law of Generalization

We have now outlined the notion of information distance. Moreover, we have made the modest assumption that the process by which the cognitive system confuses one item with another must be a mixture of computable and random factors. We are now in a position to show how the generalized 'algorithmic' version of the Universal Law can be derived from first principles.

In intuitive terms, the logic of the argument is as follows. We assume only that the confusion probabilities are semicomputable. That is, we need make no specific further assumptions about the particular cognitive algorithms or random processes that lead one item to be mis-identified as another. Standard results in Kolmogorov complexity theory can then be applied, showing that the negative log-probability of mis-labelling stimulus a with response b, $-\log \Pr(R_b|S_a)$, is close to the conditional Kolmogorov complexity from a to b, $K(R_b|S_a)$. We can then apply this key result to link together Shepard's specific measure of the confusability between a pair of items (which is defined in terms of probabilities) and information distance (which is defined in terms of Kolmogorov complexity). This allows us to derive Shepard's universal law, where the distance measure between items is information distance.

3.1. Optimal codes and entropy

For technical reasons, we recall some notions from information theory (Cover & Thomas, 1991). Suppose we have a random source emitting letters from the alphabet with certain frequencies. Our task is to encode messages consisting of many letters in binary in such a way that, on average, the length of the encoded message is as short as possible. It is evident that by assigning the few shortest binary sequences to the most common letters and the longer sequences to the rare ones, the expected length of a message is less than if we assigned equal length codes to all letters. Thus, the Morse code in telegraphy is adapted to the frequency of letter-occurrences in English. It assigns short sequences of dots and dashes to more frequently occurring letters: 'a' is encoded as '. -' and 't' is encoded as '-'. Long sequences of dots and dashes are assigned to less frequently occurring letters such as 'z' which is encoded

as '-- ..'. A *prefix code* has the property that no code word starts with another code word as proper initial segment (prefix). This property makes it possible to parse an encoded message into the sequence of code words from which it is composed in only one way: We can unambiguously retrieve the encoded message. Note that the Morse code is not a prefix-code. A prefix code for the letters a, b, ..., z is, for example, to encode 'a' by '.-', the letter 'b' by '.-', and so on. This example is not very efficient; it is essentially a tally code. It is easy to design more efficient prefix-codes. Nonetheless, since prefixes are excluded, it is clear that prefix-codes cannot be as concise as general codes. But prefix-codes have a very general and central property that makes them more practical than other codes: for every code that is uniquely decodable there is a prefix-code that has precisely the same lengths of code words. Thus, when we want unambiguous codes then we can as well restrict ourselves to prefix-codes: they are uniquely decodable and have the additional advantage that we can parse them in one pass going left-to-right. Moreover, it is well-known that there is a tight connection between prefix codes, probabilities, and notions of optimal codes: Call the letters to be encoded by the name 'source words'. Consider an ensemble of source words with source word x having probability $P(x)$. Assign code words with code word length $l_P(x)$ to source word x. The so-called Noiseless Coding Theorem of Shannon states that among all prefix codes the minimal average code word length, the average taken with respect to the distribution P, satisfies

$$H(P) \leqslant \sum_x P(x) l_P(x) \leqslant H(P) + 1,$$

where $H(P) = -\sum_x P(x) \log P(x)$ is called the *entropy* of P. This minimum is reached by the so-called Shannon–Fano code (the details of which do not matter here) where we assign a code word of length $-\lceil \log P(x) \rceil$ to source word x. Intuitively, this code is optimally 'adapted' to the probability distribution P of the source words.

3.2. The Kolmogorov code

A trivial application of this result, generalized to conditional probability, is that using a code that is well-adapted to probability distribution $\Pr(\cdot|S_a)$, the Shannon–Fano code length of R_b, given S_a, is $-\log \Pr(R_b|S_a)$. This allows us to quantify the code length of a particular way of mapping S_a onto R_b. First, specify the probability distribution $\Pr$—this can be done using a computer program of length $K(\Pr)$ (the existence of such a computer program is guaranteed by the condition that $\Pr$ is computable). Then specify R_b, given S_a using the probability distribution $\Pr$, which takes length $-\log \Pr(R_b|S_a)$ using the Shannon–Fano code.

Thus, the total length of this way of mapping from S_a to R_b is: $K(\Pr) - \log \Pr(R_b|S_a)$. Obviously, every computable code that maps S_a to R_b must be at least as long as the shortest computable code which does this, the length of which is, by definition, $K(R_b|S_a)$. Thus, we can infer that

$$K(R_b|S_a) \leqslant K(\Pr) - \log \Pr(R_b|S_a),$$

which, when rearranged, provides an *upper bound* on $\Pr(R_b|S_a)$:

$$\Pr(R_b|S_a) \leqslant 2^{K(\Pr) - K(R_b|S_a)}.$$

In the following it is convenient to use a special notation for (in)equality up to an additive constant. From now on, we will denote by $\overset{+}{<}$ an inequality to within an additive constant, and by $\overset{+}{=}$ the situation when both $\overset{+}{<}$ and $\overset{+}{>}$ hold.

We derive a *lower bound*: Suppose we sample from a distribution $\Pr$, and encode the outcomes using an optimally adapted code, as described above. We can then write down the expected code length as

$$E_{\Pr}(-\log \Pr(\cdot)) \overset{+}{=} \sum_x \Pr(x)(-\log \Pr(x))$$

$$\overset{+}{=} -\sum_x \Pr(x) \log \Pr(x).$$

Here $E_{\Pr} f(\cdot) = -\sum_x \Pr(x) f(x)$ is called the *expectation* of $f(x)$ with respect to $\Pr$. With $f(x) = -\log \Pr(x)$ this is the above expression for the *entropy* of $\Pr$. Now suppose that we consider, instead, the expected value of the Kolmogorov complexity of x—the shortest code length for x, in a universal programming language. In general, of course, this will be at least as great as the entropy—because the entropy reflects the shortest expected code length for x, using a code which is optimally adapted to $\Pr$. So this means that

$$E_{\Pr}(-\log \Pr(x)) \leqslant E_{\Pr} K(x).$$

Nonetheless, though, there will typically be individual values of x for which Kolmogorov complexity is significantly less than the code length ($\approx -\log \Pr(x)$) optimized to $\Pr$. For example, suppose that $\Pr$ is an extremely simple distribution over binary strings, such that 0 and 1 values both have a probability of 0.5, and are independent—as if, for example, the string were generated by a series of fair coin flips. Consider the string that consists of a million consecutive 1s. According to $\Pr$, the probability of this string is $2^{-1,000,000}$, and the code length according to the code optimally adapted to $\Pr$ is $-\log 2^{-1,000,000} = 1,000,000$. Indeed, this same code length will be assigned for every binary string of 1,000,000 characters generated by $\Pr$, because according to $\Pr$ all such strings have the same probability of occurring. However, the Kolmogorov complexity of this particular string will, of course, be considerably less

than 1,000,000 bits—because a short computer program can print a million 1's and then halt.

The reason that this particular string generated by Pr has a smaller Kolmogorov complexity that is associated with the optimal code for Pr, is that the string has some additional structure, that is unexplained by Pr. The existence of this additional structure (such as being a sequence of repeated items, or alternating items, or encoding $\pi = 3.14\ldots$ in binary, or whatever it may be) can therefore be used to provide an unexpectedly short code for the string. Intuitively, though, it seems that strings generated by Pr with such additional useful structure must be rare—it would seem likely that the overwhelming majority of strings generated by Pr will merely be typical of the distribution, and hence will not contain any useful 'unexpected' structure. The Kolmogorov complexity of these items will, therefore, be at least as great as the code length according to the code optimally adapted to Pr. This intuition is indeed correct. It can be shown that the probability that an item, x, drawn from Pr, is such that

$$-\log \Pr(x) \leqslant K(x)$$

goes to 1 as the length of x grows unboundedly (Li & Vitányi, 1997, 2001). That is, almost all probability is concentrated on items x satisfying this inequality—and, if the probability is not dramatically skewed this implies that the overwhelming majority of x's do so. Items for which this inequality holds are known as $\Pr(\cdot)$-random, indicating that they do not have sufficient 'unexpected' structure to support an shorter coding than would be expected from Pr.[10]

A straightforward generalization of this result to conditional probability, and its application in the present context yields the result that, for the R_b that are $\Pr(.|S_a)$-random (and the probability of sampling

---

[10] Here, we touch on the more general idea that the randomness of a string may be assessed by considering its Kolmogorov complexity. This idea has been developed into a deep mathematical theory of 'algorithmic' randomness. The common meaning of a 'random object' is an outcome of a random source. Such outcomes have expected properties but particular outcomes may or may not possess these expected properties. In contrast, we use the notion of randomness of individual objects. This elusive notion's long history goes back to the initial attempts by von Mises (1919), to formulate the principles of application of the calculus of probabilities to real-world phenomena. Classical probability theory cannot even express the notion of 'randomness of individual objects.' Following almost half a century of unsuccessful attempts, the theory of Kolmogorov complexity (Kolmogorov, 1965), and Martin-Löf tests for randomness (Martin-Löf, 1966), finally succeeded in formally expressing the novel notion of individual randomness in a correct manner, see Li and Vitányi (1997). Every individually random object possesses individually all effectively testable properties that are only expected for outcomes of the random source concerned. It will satisfy *all* effective tests for randomness—known and unknown alike. Details are beyond the scope of this treatment, but see the discussions in Martin-Löf (1966) and Li and Vitányi (1997).

such an item from $\Pr(\cdot|S_a)$ will be almost 1), then

$$-\log \Pr(R_b|S_a) \leqslant K(R_b|S_a).$$

This equation can be rearranged to give a lower bound on $\Pr(R_b|S_a)$:

$$2^{-K(R_b|S_a)} \leqslant \Pr(R_b|S_a).$$

Putting the upper and lower bounds together, we can conclude that, for $\Pr(\cdot|S_a)$-random items:

$$2^{-K(R_b|S_a)} \leqslant \Pr(R_b|S_a) \leqslant 2^{K(Pr)-K(R_b|S_a)}.$$

This result implies that, for almost all items (the $\Pr(\cdot|S_a)$-random items), $\Pr(R_b|S_a)$ is close to $2^{-K(R_b|S_a)}$, to within a multiplicative factor, $2^{K(Pr)}$. Since $K(Pr)$ is constant, independent of the items a and b we can simplify the formulas, using the earlier introduced notation '$\stackrel{\pm}{=}$', to

$$\log \Pr(R_b|S_a) \stackrel{\pm}{=} -K(R_b|S_a), \tag{2}$$

for almost all items b (the $\Pr(\cdot|S_a)$-random items), with respect to every item a. That is, (2) holds for almost all pairs of items a, b, with $\Pr(\cdot|a)$-probability going to 1 for b increasing with every fixed a.

3.3. Formal derivation of the law

Now we are in a position to directly relate Shepard's Universal Law to information distance. Shepard uses a specific measure, $G(a, b)$, as a measure of what he terms the 'generalization' between items a and b. Here S_a is the stimulus related to item a with the correct corresponding response R_a. Possibly, the stimulus S_a elicits another response R_b ($b \neq a$). The probability of this happening is $\Pr(R_b|S_a)$.

$$G(a, b) = \left[\frac{\Pr(R_a|S_b)Pr(R_b|S_a)}{\Pr(R_a|S_a)Pr(R_b|S_b)} \right]^{\frac{1}{2}} \tag{3}$$

To express $G(a, b)$ in terms of Kolmogorov complexity, observe the following. We have assumed at the outset that there is a simple fixed program, of length say C bits, that maps S_x to R_x for all x's. This means that $K(R_a|S_a)$ and $K(R_b|S_b)$ are upper bounded by a fixed constant C independent of variable items a and b. Moreover, $K(R_a|S_a)$ and $K(R_b|S_b)$ are strictly positive, as a consequence of the definition of Kolmogorov complexity (the Universal Turing Machine must have some program to do the transformation). Therefore, the denominator in (3) can be replaced by a positive constant independent of a and b. Taking this into account, and substituting (2) into (3) we obtain that, for almost all a, b (the almost all $\Pr(\cdot|S_a)$-random items b with respect to every item a, in the above sense of concentration of Pr-probability,

$$\log G(a, b) \stackrel{\pm}{=} \frac{1}{2} [-K(R_a|S_b) - K(R_b|S_a)]. \tag{4}$$

We have also assumed at the outset that there are fixed length programs that compute S_a from a, R_a from a, S_a from R_a, and so on, for every item a. Therefore, $K(R_b|S_a) \overset{+}{=} K(b|a)$ and $K(R_a|S_b) \overset{+}{=} K(a|b)$. Earlier, we defined the 'sum'-information distance $D_{\text{sum}}(a,b)$ between a and b as the sum $K(b|a) + K(b|a)$ of the conditional complexities between the two items. Therefore, $D_{\text{sum}}(a,b) = K(b|a) + K(a|b) \overset{+}{=} K(R_b|S_a) + K(R_a|S_b)$, which can be substituted into (4) to give

$$\log G(a,b) \overset{+}{=} -\frac{1}{2} D_{\text{sum}}(a,b)$$

or equivalently, shifting to base e,

$$\ln G(a,b) \overset{+}{=} -\frac{\ln 2}{2} D_{\text{sum}}(a,b), \tag{5}$$

for almost all a and b (in the above sense of concentration of Pr-probability).

This means that $G(a,b)$ is a negative exponential function of information distance D_{sum}, which is Shepard's Universal Law. This is a surprising result. It indicates that $G(a,b)$, a measure of the confusability between the items a and b, has a specific functional relationship with a general measure of distance, subject only to the mild assumption that the probability distribution determining confusability is computable.

Two points concerning this result are worth noting. The first is that it might appear that the result is somewhat *too* precise. Shepard's Universal Law allows two free parameters, A and B:

$$G(a,b) = Ae^{-B \cdot D(a,b)}$$

whereas (5) has no apparent free parameters. But this disparity is deceptive: The $\overset{+}{=}$-symbol hides the parameter A, because it gives equality—but only up to an additive constant term (which translates into an multiplicative constant factor since (5) gives the logarithmic version of the relation). Moreover, the units for D_{sum} are arbitrary, because they depend on the choice of a binary alphabet for measuring Kolmogorov complexity. Shifting to an alphabet with a different number of elements (which can be viewed as having any real value), or to a different computable correspondence between object and binary representation, values of D_{sum} will change by a multiplicative constant, which can be interpreted as parameter B.

Moreover, of course, our generalization of the Universal Law of Generalization does not hold for *all* items a and b but for *almost all* items a and b (in the sense of concentration of Pr-probability).[11]

The second point is that it might appear that the outcome of this result provides some reason to prefer D_{sum} over D_{max} as a preferred measure of information distance in psychological contexts. But note that they give the same values up to a multiplicative factor 2, since we have noted above (1) that $D_{\text{max}} \leqslant D_{\text{sum}} \leqslant 2D_{\text{max}}$. But even so, this apparent preference between the two measures is merely a consequence of the specific way in which Shepard defined G.

3.4. An alternative Universal Law of Generalization

It turns out that there are mathematical reasons to choose a slightly different measure of the confusability between items a, b than initially chosen by Shepard. Define a new measure of confusability as

$$G'(a,b) = \frac{\min\{\Pr(R_b|S_a), \Pr(R_a|S_b)\}}{\max\{\Pr(R_a|S_a), \Pr(R_b|S_b)\}},$$

where we consider the ratio of (i) to (ii), such that (i) is the minimum of the two probabilities that the stimulus for a elicits the response for b or the stimulus for b elicits the response for a, and (ii) is the maximum of the two probabilities that the stimulus for a elicits the response for a and the stimulus for b elicits the response for b. Then analogous analysis to that above leads to a similar result. Thus, from the earlier analysis argument we have $-\log \Pr(R_b|S_a) \overset{+}{=} K(b|a)$ (by noting $K(\Pr) \overset{+}{=} 0$). And moreover $K(b|a) \overset{+}{=} 0$ for $b = a$ so that the precise form of the denominator—whether min, max, square root of product—does not matter since it will be a constant independent of a and b. The important part of the formula is the numerator: note that the minimum for the conditional probabilities in the formula translates into the maximum for the related conditional Kolmogorov complexities. Thus, for almost all a, b, in the sense of concentration of Pr-probability, we obtain $\log G'(a,b) \overset{+}{=} -D_{\text{max}}(a,b)$, and therefore

$$\ln G'(a,b) \overset{+}{=} -(\ln 2)D_{\text{max}}(a,b).$$

Straightforward substitution of the log-expressions of G and G' in the relation (1) yields $-\log G'(a,b) \overset{+}{<} -2\log G(a,b) \overset{+}{<} -2\log G'(a,b)$. That is, there are positive constants C_1, C_2 independent of a, b such that

$$G'(a,b) \leqslant C_1 G(a,b) \leqslant C_2 \sqrt{G'(a,b)}.$$

for almost all a, b, in the sense of concentration of Pr-probability.[12] It seems likely that the two measures $G'(a,b)$ and $G(a,b)$ will be so strongly positively correlated, in the empirical data, that the empirical fits derived by Shepard for the Universal Law using 'G'

---

[11] It is, of course, logically possible that the very subset of items for which our generalization does not hold just happens to correspond to representations of items that naturally occur in the environment. If this were the case, then this would pose problems for the present result. In the absence of any reason to suppose that this is the case, however, we do not consider this further. We thank Peter van der Helm for noting this point.

[12] Note that $\sqrt{G'(a,b)} > G'(a,b)$ since $0 < G'(a,b) < 1$.

would be roughly equally strong using 'G'', although we do not assess this directly.

There is a formal reason to prefer the $G'(a, b)$-version as the proper measure of confusability over the $G(a, b)$ version, since it appeared above that the negative logarithm of the $G'(a, b)$ is precisely (up to the $\overset{+}{=}$ relation) the information distance $D_{\max}(a, b)$. As we observed above, the latter has been shown in Bennett et al. (1998) to be the *universal* (that is, optimal) cognitive distance. Viewing a cognitive distance D as defined in Bennett et al. (1998) as a code-length this means the following: If we fix b and let a run over the possible items, we can then define the probability $P(a|b)$ of a given b by $P(a|b) = 2^{-D(a,b)}$.

It was shown in Bennett et al. (1998) that $\sum_{a:a \neq b} 2^{-D(a,b)} \leqslant 1$ so that $P(a|b)$ is a proper probability. In fact, the cognitive distance code of length $D(a, b)$, the shortest binary program that serves to compute a from b and also to compute b from a, is length-equivalent to the Shannon–Fano code associated with $P(a|b)$, and hence achieves the optimal (minimal) expected code word length (the entropy of P, by Shannon's Noiseless Coding Theorem (Cover & Thomas, 1991) among all prefix-codes.

Now let us go to the punch line: Since $D_{\max}(a, b)$ is the minimal cognitive distance, minorizing all other cognitive distances, up to a constant additive term, its associated probability distribution $P_{G'}(a|b) := G'(a, b) = 2^{-D_{\max}(a,b)}$, with b fixed, majorizes, up to a constant multiplicative factor, *every* probability distribution $P_D(a|b) = 2^{-D(a,b)}$ with $D(a, b)$ a cognitive distance. This follows because if $D_{\max}(a, b)$ is smaller than any other $D(a, b)$ up to an additive constant, where D is a cognitive distance, then $2^{-D_{\max}(a,b)}$ is larger than $2^{-D(a,b)}$ up to a multiplicative constant.

That is, if we fix b and consider the probability of confusing any item a with item b, according to some semi-computable cognitive similarity criterion, as the negative exponent of the cognitive distance according to that similarity criterion, then the confusion measure $G'(a, b)$ is the largest such probability incorporating confusability according to *all* semi-computable (including all computable) cognitive similarity criteria. It is particularly attractive that the universal law can be demonstrated when the underlying distance measure between items is information distance—because, as we have stressed, information distance captures similarities between items that are similar in any other semi-computable cognitive distance. Moreover, unlike many specific cognitive distances, which are typically only defined over specific representational formats such as sets of features or locations in a multidimensional space, information distance is well-defined over any type of representation. Hence, information distance provides a maximally general, and maximally neutral measure of the distance between items.

3.5. Normalized Universal Law

The universal laws above are formulated in terms of absolute information distance. But, from a psychological point of view, this might seem to be inappropriate. If two enormously complex images, each containing 10^9 bits of information, are separated by an information distance of 1000, then this would seem to indicate that they are remarkably similar, because they share almost all of their structure (although the nature of the bits that differed would, of course, determine whether this similarity is perceptually salient or not). But if two much simpler images, each containing 1000 bits of information are separated by an information distance of 1000, then this indicates that they share essentially no important structure at all. Therefore, it would seem that we should classify them as highly dissimilar.

This suggests that, from a psychological point of view, that we may need some kind of relative measure of the information distance between objects, normalized for the absolute complexity of the objects involved. We do this by dividing it by the greater of the two lengths of the shortest programs that compute the strings concerned from scratch. Define a new measure of normalized G'-confusability by

$$g(a, b) = \left(\frac{\min\{\Pr(R_b|S_a), \Pr(R_a|S_b)\}}{\max\{\Pr(R_a|S_a), \Pr(R_b|S_b)\}} \right)^{-1/\min\{\log \Pr(S_a), \log \Pr(S_b)\}}.$$

Then, a quite analogous analysis to that above leads to the following result. For almost all a, b, in the sense of concentration of Pr-probability, we obtain $-\log g(a, b) \overset{+}{=} D_{\max}(a, b)/\max\{K(a), K(b)\}$ and therefore

$$-\ln g(a, b) \overset{+}{=} (\ln 2) \frac{D_{\max}(a, b)}{\max\{K(a), K(b)\}}.$$

Write $d(a, b) = D_{\max}(a, b)/\max\{K(a), K(b)\}$. This $d(a, b)$ is a *normalized information distance*, satisfies the metric requirements, is always in between 0 and 1, and can be viewed as a 'percentage-wise similarity'. It turns out that $d(x, y)$ is universal (always gives the smallest distance) in a wide class of sensible and computable normalized similarity metrics.

This measure (or at least a close variant) has been used to in non-psychological, but somewhat related contexts of text analysis and bioinformatics (Li et al., 2001; Li & Vitányi, 2001). Just as in the psychological context the percentage of shared information may be the important measure, it is a convenient way to measure English text or DNA sequence similarity. In those areas normalized information distance (or rather, the less perfect close relative based on the sum distance) has been experimentally applied. Using a compression program called *GenCompress* we

heuristically approximate $K(x)$ and $K(x|y)$. With the caveat that the program is 'heuristic', that is, without mathematical closeness-of-approximation guarantees (C.H. Bennett, M. Li, B. Ma, in an article to appear in *Scientific American*) took 33 chain letters—collected by Charles Bennett from 1980 to 1997—and approximated their pairwise distance $d(x, y)$. Then, we used standard phylogeny building programs from bioinformatics research to construct a tree of these chain letters. The resulting tree gives a perfect phylogeny for all notable features, in the sense that each notable feature is grouped together in the tree (so that the tree is parsimonious). This fundamental notion can be applied in many different areas. One of these concerns a major challenge in bioinformatics: to find good methods to compare genomes. Traditional approaches of computing the phylogeny use so-called 'multiple alignment.' They would not work here since chain letters contain swapped sentences and genomes contain translocated genes and noncoding regions. Using the chain letter method, a more serious application in Li et al. (2001) and Li and Vitányi (2001) automatically builds correct phylogenies from complete mitochondrial genomes of mammals. This work corroborated a biological conjecture that ferungulates—placental mammals that are not primates, including cats, cows, horses, whales—are closer to the primates—monkeys, humans—than to rodents.

4. Discussion

We have shown that Shepard's Universal Law of generalization can be derived, if we assume that psychological distance is modelled as information distance. We have also indicated that information distance is a highly general notion of distance, which may be of broader psychological interest. Thus, we have here addressed the specific relationship evident in the data that Shepard (1987) encapsulates as the Universal Law. But an interesting open question is whether the notion of information distance can be used to address the question of generalization, as tackled by the results in Shepard (1987) and Tenenbaum and Griffiths (2001). Given the rich mathematical connections between the theory of Kolmogorov complexity and inductive inference and statistics (e.g., Rissanen, 1986, 1989, 1996; Solomonoff, 1964, 1978; Wallace & Boulton, 1968; Wallace & Freeman, 1987), it may be hoped some relationship between information distance and generalization might be established. On the other hand, as we noted earlier, in tasks where people deliberately make generalizations from examples, it is possible that wide variation in people's strategies and background knowledge may mean that there are few robust regularities in the empirical data (Stewart & Chater, 2002).

The high level of abstraction used to derive the results above is in some ways attractive, because it allows progress to be made, without having to rely on detailed proposals about the nature of the representations and processes used by the cognitive system. But, of course, developing a quantitative and detailed model of which confusions people do make, and which they do not, would required specifying such proposals. An account of this kind would, at minimum, be able predict that people would not detect a regularity between, say, an random 'chequerboard' providing a binary encoding of the initial digits of π and a binary encoding of the initial digits of π^2, even though the information distance between these two stimuli is small. On such an account, the failure to detect the regularity might arise either because of the representations used, or the available processes defined over those representations. For example, a low level (e.g., retinal level) representation of these stimuli might preserve the detailed information that allows the short program to be constructed between them. But it seems inconceivably unlikely that the cognitive system has an operation for squaring large binary numbers that could possibly transform one representation to the other—and hence the regularity, although present, will be cognitively irrelevant. On the other hand, a higher level representation may throw away such details. For example, such a representation might convey merely the pattern: 'random chequerboard'. If so, then all random arrays of pixels will seem equally similar to each other, and again the special relationship between the binary encodings of π and π^2 will be lost—hence the regularity is lost in the very choice of representation, whatever processes might subsequently act upon those representations.

The possibility of representing the same stimulus at various levels of representation also raises interesting psychological issues. For example, two random chequerboards (or, more extremely, two TV screens with white noise on them) are likely to be quite confusable, even though the information distance between a detailed representation of those stimuli is very large (because they share no useful structure, the length of a program to transform one into the other will be long). In the present framework, this indicates that the representations that are cognitive relevant in determining such confusability might be quite high-level. Formally, such problems may be related to notions of 'Kolmogorov sufficient statistic' as treated in Gács, Tromp, and Vitányi (2001), where one may choose the level of representation by selecting the model class. But we do not pursue that issue here.

Finally, note that the specific representations and processes employed by the cognitive system may be highly selective. In particular, we might expect that, through pressures of either learning and/or evolution, these representations and processes will be geared to

finding the regularities that are actually typically present in the natural world (it seems safe to assume that regularities concerning the squaring of binary numbers are not of this kind), and also regularities that are relevant to that person's *actions* and *goals*. One might reasonably suspect, therefore, that a detailed psychological account of confusability would reveal, for example, that items which differ on some matter of great importance to survival (e.g., two animals which differ regarding their teeth or claws) might be less frequently confused than items which differ on some matter of marginal importance to survival (e.g., two clouds that differ in some aspect of shape). The present analysis, however, is sufficiently abstract that it does not need to make any commitment on such issues.

Overall, we note that the generalization of the Universal Law that we have outlined in this paper is attractive, because it applies in such a general setting. Specifically, it does not presuppose that items correspond to points in an internal multidimensional psychological space. But, as we have noted, there is presently relatively little empirical evidence for the truth of the Universal Law outside this context. Thus, the present analysis suggests that there is a need for empirical research to determine whether the Universal Law does indeed hold in these more general circumstances. Such research might investigate whether the Universal Law still holds, as we would predict, even for stimuli, such as complex visual or linguistic material, that seems unlikely to embed naturally into a multidimensional psychological space. We hope that the present paper will serve as a stimulus to empirical research of this kind.

References

Akerlof, G., & Yellen, J. (1985). Can small deviations from rationality make significant differences to economic equilibria? *American Economic Review, 75*, 708–720.

Ashby, F. G., & Townsend, J. T. (1986). Varieties of perceptual independence. *Psychological Review, 93*, 154–179.

Attneave, F., & Frost, R. (1969). The determination of perceived tridimensional orientation by minimum criteria. *Perception and Psychophysics, 6*, 391–396.

Bak, P. (1997). *How nature works: The science of self-organized criticality*. New York: Copernicus Press.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge: Cambridge University Press.

Bennett, C. H., Gács, P., Li, M., Vitányi, P. M. B., & Zurek, W. (1998). Information distance. *IEEE Transactions on Information Theory, IT-44*, 1407–1423.

Biedermam, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review, 94*, 115–147.

Brent, M. R., & Cartwright, T. A. (1996). Distributional regularity and phonotactic constraints are useful for segmentation. *Cognition, 61*, 93–125.

Buffart, H., Leeuwenberg, E., & Restle, F. (1981). Coding theory of visual pattern completion. *Journal of Experimental Psychology: Human Perception and Performance, 7*, 241–274.

Carnap, R. (1952). *The continuum of inductive methods*. Chicago: University of Chicago Press.

Cartwright, N. (1983). *How the laws of Physics lie*. Oxford: Oxford University Press.

Chater, N. (1996). Reconciling simplicity and likelihood principles in perceptual organization. *Psychological Review, 103*, 566–581.

Chater, N. (1997). Simplicity and the mind. *The Psychologist*, 495–498.

Chater, N. (1999). The search for simplicity: A fundamental cognitive principle? *Quarterly Journal of Experimental Psychology, 52A*, 273–302.

Chater, N., & Hahn, U. (1997). Representational distortion, similarity, and the universal law of generalization. In: *Proceedings of the interdisciplinary workshop on similarity and categorization, Sim-Cat'97*. Department of Artificial Intelligence, University of Edinburgh.

Cheng, K. (2000). Shepard's universal law supported by honeybees in spatial generalization. *Psychological Science, 11*, 403–408.

Cover, T. M., & Thomas, J. A. (1991). *Elements of Information Theory*. New York: Wiley.

Cunningham, J. P., & Shepard, R. N. (1974). Monotone mapping of similarities into a general metric space. *Journal of Mathematical Psychology, 11*, 335–363.

Ekman, G. (1954). Dimensions of color vision. *Journal of Psychology, 38*, 467–474.

Ennis, D. M. (1989). Toward a Universal Law of Generalization. *Science, 242*, 944.

Falk, R., & Konold, C. (1997). Making sense of randomness: Implicit encoding as a basis for judgment. *Psychological Review, 104*, 301–318.

Falmagne, J. C. (1986). Psychophysical measurement and theory. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance* (pp. 1–66), New York: Wiley.

Fechner, G. T. (1966). *Elements of psychophysics*. In D. H. Howes & E. C. Boring (Eds.) (H. E. Adler, trans.). New York: Holt, Reinhart & Winston (originally published, 1860).

Feldman, J. (2000). Minimization of Boolean complexity in human concept learning. *Nature, 407*, 630–633.

Fodor, J. A., Bever, T. G., & Garrett, M. F. (1974). *The psychology of language*. York: McGraw Hill.

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

Friedman, M. (1953). *Essays in positive economics*. Chicago: University of Chicago Press.

Gács, P., Tromp, J. T., & Vitányi, P. M. B. (2001). Algorithmic statistics. *IEEE Transactions in Information Theory, 47*, 2443–2463.

Gao, Q., Li, M., & Vitányi, P. M. B. (2000). Applying MDL to learning best model granularity. *Artificial Intelligence, 121*, 1–29.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-completeness*. New York: Freeman.

Goldsmith, J. A. (2001). Unsupervised learning of the morphology of a natural language. *Computational Linguistics, 27*, 153–198.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive Psychology, 23*, 222–262.

Guttman, N., & Kalish, H. I. (1956). Discriminability and stimulus generalization. *Journal of Experimental Psychology, 51*, 79–88.

Hahn, U., Chater, N., & Henley, R. (1996). Weighting in similarity judgments: An investigation of the 'MAX' hypothesis. *Proceedings of the 18th annual conference of the cognitive science society*. Mahwah, NJ: Erlbaum.

Hahn, U., Chater, N., & Richardson, L. B. (2001). Similarity: A transformational approach. *Proceedings of the 23rd annual conference of the cognitive science society*. Mahwah, NJ: Erlbaum.

Hahn, U., Chater, N., & Richardson, L. B. (2003). Similarity as transformation. *Cognition*.

van der Helm, P. A. (2000). Simplicity versus likelihood in visual perception: From surprisals to precisals. *Psychological Bulletin*, *126*, 770–800.

van der Helm, P. A., & Leeuwenberg, E. L. J. (1996). Goodness of visual regularities: A nontransformational approach. *Psychological Review*, *103*, 429–456.

Hochberg, J., & McAlister, E. (1953). A quantitative approach to figure 'goodness'. *Journal of Experimental Psychology*, *46*, 361–364.

Johansson, G. (1950). *Configurations in event perception*. Stockholm: Almqvist & Wiksell.

Kemeny, J. G. (1953). The use of simplicity in induction. *Philosophical Review*, *62*, 391–408.

Koffka, K. (1962). *Principles of Gestalt psychology* (5th ed.). London: Routledge and Kegan Paul (Original work published in 1935).

Kolmogorov, A. N. (1965). Three approaches to the definition of the concept 'quantity of information'. *Problems in Information Transmission*, *1*, 1–7.

Kreps, D. (1990). *A course in microeconomic theory*. Princeton, NJ: Princeton University Press.

Laming, D. (1997). *The measurement of sensation*. Oxford: Oxford University Press.

Leeuwenberg, E. (1969). Quantitative specification of information in sequential patterns. *Psychological Review*, *76*, 216–220.

Leeuwenberg, E. (1971). A perceptual coding language for perceptual and auditory patterns. *American Journal of Psychology*, *84*, 307–349.

Leeuwenberg, E., & Boselie, F. (1988). Against the likelihood principle in visual form perception. *Psychological Review*, *95*, 485–491.

Li, M., Badger, J. H., Chen, X., Kwong, S., Kearney, P., & Zhang, H. (2001). An information-based sequence distance and its application to whole mitochondrial genome phylogeny. *Bioinformatics*, *17*, 149–154.

Li, M., & Vitányi, P. M. B. (1997). *An introduction to Kolmogorov complexity and its applications* (2nd ed.). New York: Springer.

Li, M., & Vitányi, P. M. B. (2001). Normalized information distance with an application to whole genome phylogeny analysis. Computer Science, University of California, Santa Barbara, Manuscript.

Mach, E. (1959). *The analysis of sensations* (Translated by C. M. Williams & S. Waterlow, original work published, 1886). New York: Dover.

Marr, D. (1982). *Vision*. San Fransisco: W. H. Freeman & Co.

Martin-Löf, P. (1966). The definition of random sequences. *Information and Control*, *9*, 602–619.

McGuire, W. J. (1961). A multiprocess model for paired-associate learning. *Journal of Experimental Psychology*, *62*, 335–347.

Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.

Miller, G. A., & Nicely, P. E. (1955). An analysis of perceptual confusions among some English consonants. *Journal of the Acoustical Society of America*, *27*, 338–352.

Minsky, M. (1977). Frame system theory. In P. N. Johnson-Laird, & P. C. Wason (Eds.), *Thinking: Readings in cognitive science* (pp. 355–376). Cambridge: Cambridge University Press (originally published 1975).

von Mises, R. (1919). Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematisches Zeitschrift*, *5*, 52–99.

Myung, I. J., Balasubramanian, V., & Pitt, M. A. (2000a). Counting probability distributions: Differential geometry and model selection. *Proceedings of the National Academy of Sciences USA*, *97*, 11170–11175.

Myung, I. J., Forster, M. R., & Browne, M. W. (Eds.) (2000b). Special issue on model selection. *Journal of Mathematical Psychology*, *44*, 1–231.

Myung, I. J., & Pitt, M. A. (1997). Applying Occam's razor in modeling cognition: A Bayesian approach. *Psychological Bulletin and Review*, *4*, 79–95.

Myung, I. J., & Shepard, R. N. (1996). Maximum entropy inference and stimulus generalization. *Journal of Mathematical Psychology*, *40*, 342–347.

Newton, I. (1687). *Philosophiae Naturalis Principia Mathematica* also known as the *Principia*.

Nosofsky, R. M. (1985). Overall similarity and the identification of separable-dimension stimuli: A choice model analysis. *Journal of Experimental Psychology: Perception and Psychophysics*, *38*, 415–432.

Nosofsky, R. M. (1988a). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, Cognition*, *14*, 700–708.

Nosofsky, R. M. (1988b). Similarity, frequency, and category representation. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *14*, 54–65.

Odifreddi, P. (1989). *Classical recursion theory*. Amsterdam: North-Holland.

Palmer, S. E. (1982). Symmetry, transformation, and the structure of perceptual systems. In J. Beck (Ed.), *Organization and representation in perception* (pp. 95–144). Hillsdale NJ: Erlbaum.

Pitt, M. A., Myung, I. J., & Zhang, S. (2002). Toward a method of selecting among computational models of cognition. *Psychological Review*, *109*, 472–491.

Pothos, E., & Chater, N. (2002). A simplicity principle in human unsupervised categorization. *Cognitive Science*, *26*, 303–343.

Pruzansky, S., Tversky, A., & Carroll, J. D. (1982). Spatial versus tree representations of proximity data. *Psychometrika*, *47*, 3–24.

Restle, F. (1970). Theory of serial pattern learning: Structural trees. *Psychological Review*, *77*, 481–495.

Restle, F. (1979). Coding theory of the perception of motion configurations. *Psychological Review*, *86*, 1–24.

Reznikova, Zh. I., & Ryabko, B. Ya. (1986). Analysis of the language of ants by information-theoretical methods. *Problems in Information Transmission*, *22*, 245–249.

Rissanen, J. J. (1986). Stochastic complexity and modelling. *The Annals of Statistics*, *14*, 1080–1100.

Rissanen, J. J. (1989). *Stochastic complexity and statistical inquiry*. Singapore: World Scientific Publishers.

Rissanen, J. J. (1996). Fisher information and stochastic complexity. *IEEE Transactions on Information Theory*, *IT-42*, 40–47.

Rothkopf, E. Z. (1957). A measure of stimulus similarity and errors in some paired associate learning tasks. *Journal of Experimental Psychology*, *53*, 94–101.

Schank, R. C., & Abelson, R. P. (1977). *Scripts, plans, and understanding*. Hillsdale, NJ: Erlbaum.

Shannon, C. E. (1948). The mathematical theory of communication. *Bell System Technical Journal*, *27*, 379–423, 623–656.

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

Shepard, R. N., Hovland, C. L., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological Monographs*, *75*, 1–42.

Simon, H. A. (1959). Theories of decision-making in economics and behavioural science. *American Economic Review*, *49*, 253–283.

Simon, H. A. (1972). Complexity and representation of patterned sequences of symbols. *Psychological Review*, *79*, 369–382.

Simon, H. A., & Kotovsky, K. (1963). Human acquisition of concepts for sequential patterns. *Psychological Review*, *70*, 534–546.

Solomonoff, R.J. (1964). A formal theory of inductive inference, Part 1 and Part 2. *Information and Control*, *7*, 1–22, 224–254.

Solomonoff, R. J. (1978). Complexity-based induction systems: comparisons and convergence theorems. *IEEE Transactions on Information Theory*, *IT-24*, 422–432.

Stevens, S. S. (1961). To honor Fechner and repeal his law. *Science*, *133*, 80–86.

Stewart, N., & Chater, N. (2002). The effect of category variability in perceptual categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*, 893–907.

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–641.

Turing, A. M. (1936). On computable numbers with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, Series 2* **42**, 230–265; Correction, *Proceedings of the London Mathematical Society, Series 2*, **43** (1937), 544–546.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*, 327–352.

Ullman, S. (1996). *High-level vision: object recognition and visual cognition*. Cambridge, MA: MIT Press.

Vitányi, P. M. B., & Li, M. (2000). Minimum description length induction, Bayesianism, and Kolmogorov complexity. *IEEE Transactions on Information Theory*, *IT-46*, 446–464.

Vitz, P. C., & Todd, T. C. (1969). A coded element of the perceptual processing of sequential stimuli. *Psychological Review*, *76*, 433–449.

Wagemans, J. (1995). Detection of visual symmetries. *Spatial Vision*, *9*, 9–32.

Wallace, C. S., & Boulton, D. M. (1968). An information measure for classification. *Computing Journal*, *11*, 185–195.

Wallace, C. S., & Freeman, P. R. (1987). Estimation and inference by compact coding, *Journal of the Royal Statistical Society, Series B*, *49*, 240–251. Discussion: *Journal of the Royal Statistical Society, Series B*, 252–265.