Probability, algorithmic complexity, and subjective randomness

Thomas L. Griffiths
gruffydd@psych.stanford.edu
Department of Psychology
Stanford University

Joshua B. Tenenbaum
jbt@mit.edu
Brain and Cognitive Sciences Department
Massachusetts Institute of Technology

Abstract

We present a statistical account of human randomness judgments that uses the idea of algorithmic complexity. We show that an existing measure of the randomness of a sequence corresponds to the assumption that non-random sequences are generated by a particular probabilistic finite state automaton, and use this as the basis for an account that evaluates randomness in terms of the length of programs for machines at different levels of the Chomsky hierarchy. This approach results in a model that predicts human judgments better than the responses of other participants in the same experiment.

The development of information theory prompted cognitive scientists to formally examine how humans encode experience, with a variety of schemes being used to predict subjective complexity (Leeuwenberg, 1969), memorization difficulty (Restle, 1970), and sequence completion (Simon & Kotovsky, 1963). This proliferation of similar, seemingly arbitrary theories was curtailed by Simon's (1972) observation that the inevitable high correlation between measures of information content renders them essentially equivalent. The development of algorithmic information theory (see Li & Vitanyi, 1997, for a detailed account) has revived some of these ideas, with code lengths playing a central role in recent accounts of human concept learning (Feldman, 2000), subjective randomness (Falk & Konold, 1997), and the role of simplicity in cognition (Chater, 1999). Algorithmic information theory avoids the arbitrariness of earlier approaches by using a single universal code: the complexity of an object (called the Kolmogorov complexity after Kolmogorov, 1965) is the length of the shortest computer program that can reproduce it.

Chater and Vitanyi (2003) argue that a preference for simplicity can be seen throughout cognition, from perception to language learning. Their argument is based upon the important constraints that simplicity provides for solving problems of induction, which are central to cognition. Kolmogorov complexity gives a formal means of addressing "asymptotic" questions about induction, such as why anything is learnable at all, but the constraints it imposes are too weak to support the rapid inferences that characterize human cognition. In order to explain how human beings learn so much from so little, we need to consider accounts that can express the strong prior knowledge that contributes to our inferences. The structures that people find simple form a strict (and flexible) subset of those easily expressed in a computer program. For example, the sequence of heads and tails TTHTTTHHTH appears quite complex to us, even though, as the parity of the first 10 digits of π, it is easily generated by a computer. Identifying the kinds of regularities that contribute to our sense of simplicity will be an important part of any cognitive theory, and is in fact necessary since Kolmogorov complexity is not computable (Kolmogorov, 1965).

There is a crucial middle ground between Kolmogorov complexity and the arbitrary encoding schemes to which Simon (1972) objected. We will explore this middle ground using an approach that combines rational statistical inference with algorithmic information theory. This approach gives an intuitive transparency to measures of complexity by expressing them in terms of probabilities, and uses computability to establish meaningful differences between them. We will test this approach on judgments of the randomness of binary sequences, since randomness is one of the key applications of Kolmogorov complexity: Kolmogorov (1965) suggested that random sequences are irreducibly complex, a notion that has inspired several psychological theories (eg. Falk & Konold, 1997). We will analyze subjective randomness as an inference about the source of a sequence X, comparing its probability of being generated by a random source, $P(X|\text{random})$, with its probability of generation by a more regular process, $P(X|\text{regular})$. Since probabilities map directly to code lengths, $P(X|\text{regular})$ uniquely identifies a measure of complexity. This formulation allows us to identify the properties of an existing complexity measure (Falk & Konold, 1997), and extend it to capture more of the statistical structure detected by people. While Kolmogorov complexity is expressed in terms of programs for a universal Turing machine, many of the regularities people detect are computable by simpler devices. We will use Chomsky's (1956) hierarchy of formal languages to organize our analysis, testing a set of nested models that can be interpreted in terms of the length of programs for automata at different levels of the hierarchy.

Complexity and randomness

The idea of using a code based upon the length of computer programs was independently proposed by Solomonoff (1964), Kolmogorov (1965), and Chaitin (1969), although it has come to be associated with Kolmogorov. A sequence X has Kolmogorov complexity $K(X)$ equal to the length of the shortest program p for a (prefix) universal Turing machine U that produces X and then halts,

$$K(X) = \min_{p:U(p)=X} l(p), \tag{1}$$

where $l(p)$ is the length of p in bits. Kolmogorov complexity can be used to define algorithmic probability, with the probability of X being

$$R(X) = 2^{-K(X)} = \max_{p:U(p)=X} 2^{-l(p)}. \tag{2}$$

There is no requirement that $R(X)$ sum to one over all sequences; many probability distributions that correspond to codes are unnormalized, assigning the missing probability to an undefined sequence.

Kolmogorov complexity can be used to mathematically define the randomness of sequences, identifying a sequence X as random if $l(X) - K(X)$ is small (Kolmogorov, 1965). While not necessarily following the form of this definition, psychologists have preserved its spirit in proposing that the perceived randomness of a sequence increases with its complexity. Falk and Konold (1997) consider a particular measure of complexity they call the "difficulty predictor" (DP), calculated by counting the number of runs (sub-sequences containing only heads or tails), and adding twice the number of alternating sub-sequences. For example, the sequence TTTHHHTHTH is a run of tails, a run of heads, and an alternating sub-sequence, $DP = 4$. If there are several partitions into runs and alternations, DP is calculated on the partition that results in the lowest score.[1]

Falk and Konold (1997) showed that DP correlates remarkably well with subjective randomness judgments. Figure 1 shows the results of Falk and Konold (1997, Experiment 1), in which 97 participants each rated the apparent randomness of ten binary sequences of length 21, with each sequence containing between 2 and 20 alternations (transitions from heads to tails or vice versa). The mean ratings show the classic preference for overalternating sequences: the sequences perceived as most random are those with 14 alternations, while a truly random process would be most likely to produce sequences

---

[1]We modify DP slightly from the definition of Falk and Konold (1997), who seem to require alternating subsequences to be of even length. The equivalence results shown below also hold for their original version, but it makes the counter-intuitive interpretation of HTHTH as a run of a single head, followed by an alternating subsequence, $DP = 3$. Under our formulation it would be parsed as an alternating sequence, $DP = 2$.
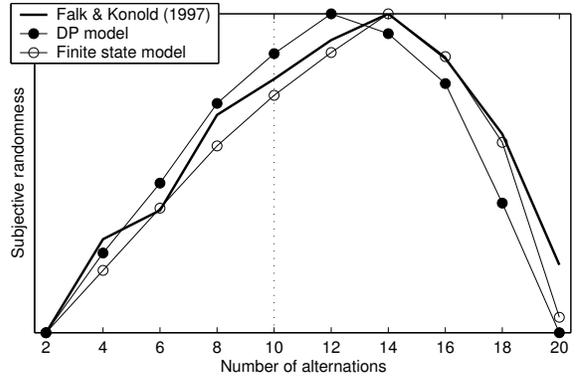

Figure 1: Mean randomness ratings from Falk and Konold (1987, Experiment 1), shown with the predictions of DP and the finite state model.

with 10 alternations. The mean DP has a similar profile, achieving a maximum at 12 alternations and giving a correlation of $r = 0.93$.

Subjective randomness as a statistical inference

Psychologists have claimed that the way we think about chance is inconsistent with probability theory (eg. Kahneman & Tversky, 1972). For example, people are willing to say that X_1=HHTHT is more random than X_2=HHHHH, while they are equally likely to arise by chance: $P(X_1|\text{random}) = P(X_2|\text{random}) = (\frac{1}{2})^5$. However, many of the apparently irrational aspects of human judgments can be understood by considering the possibility that people are assessing a different kind of probability – instead of $P(X|\text{random})$, we evaluate $P(\text{random}|X)$ (Griffiths & Tenenbaum, 2001).

The statistical basis of subjective randomness becomes clear if we view randomness judgments in terms of a signal detection task (cf. Lopes, 1982; Lopes & Oden, 1987). On seeing a stimulus X, we consider two hypotheses: X was produced by a random process, or X was produced by a regular process. Finding regularities is an important part of identifying predictable processes, a fundamental component of induction (Lopes, 1982). The decision about the source of X can be formalized as a Bayesian inference,

$$\frac{P(\text{random}|X)}{P(\text{regular}|X)} = \frac{P(X|\text{random})}{P(X|\text{regular})} \frac{P(\text{random})}{P(\text{regular})}, \tag{3}$$

in which the posterior odds in favor of a random generating process are obtained from the likelihood ratio and the prior odds. The only part of the right hand side of the equation affected by X is the likelihood ratio, which led Griffiths and Tenenbaum (2001) to define the subjective randomness of X as

$$\text{random}(X) = \log \frac{P(X|\text{random})}{P(X|\text{regular})}, \tag{4}$$

being the evidence that X provides towards the conclusion that it was produced by a random process.

When evaluating binary sequences, it is natural to set $P(X|\text{random}) = (\frac{1}{2})^{l(X)}$. Taking the logarithm in base 2, $\text{random}(X)$ is $-l(X) - \log_2 P(X|\text{regular})$, depending entirely on $P(X|\text{regular})$. We obtain $\text{random}(X) = K(X) - l(X)$, the difference between the complexity of a sequence and its length, if we choose $P(X|\text{regular}) = R(X)$, the algorithmic probability defined in Equation 2. This is identical to the mathematical definitions of randomness discussed above. However, the key point of this statistical approach is that we are not restricted to using $R(X)$: we have a measure of the randomness of X for any choice of $P(X|\text{regular})$, reducing the problem of specifying a measure of complexity to the more intuitive task of determining the probability with which sequences are produced by a regular process.

A statistical model of randomness perception

DP does a good job of predicting the results of Falk and Konold (1997), but it has some counterintuitive properties, such as independence of length: HHTHT and HHHHHHHHHHHHHHHHHHHTHT both have $DP = 3$, but the former seems more random. In this section we will show that using DP is equivalent to specifying $P(X|\text{regular})$ with a hidden Markov model (HMM), providing the basis for a statistical model of randomness perception.

An HMM is a probabilistic finite state automaton, associating each symbol x_i in the sequence $X = x_1 x_2 \ldots x_n$ with a "hidden" state z_i. The probability of X under this model is obtained by summing over all possible hidden states, $P(X) = \sum_Z P(X, Z)$, where $Z = z_1 z_2 \ldots z_n$. The model assumes that each x_i is chosen based only on z_i and each z_i is chosen based only on z_{i-1}, allowing us to write $P(X, Z) = P(z_0) \prod_{i=2}^{n} P(z_i|z_{i-1}) \prod_{i=1}^{n} P(x_i|z_i)$. An HMM can thus be specified by a transition matrix $P(z_i|z_{i-1})$, a prior $P(z_0)$, and an emission matrix $P(x_i|z_i)$. Hidden Markov models are widely used in statistical approaches to speech recognition and bioinformatics.

Using DP as a measure of randomness is equivalent to specifying $P(X|\text{random})$ with an HMM corresponding to the finite state automaton shown in Figure 2. This HMM has six states, and we can define a transition matrix

$$P(z_i|z_{i-1}) = \begin{pmatrix} \delta & C\alpha & C\alpha^2 & 0 & 0 & C\alpha^2 \\ C\alpha & \delta & C\alpha^2 & 0 & 0 & C\alpha^2 \\ C\alpha & C\alpha & 0 & \delta & 0 & C\alpha^2 \\ C\alpha & C\alpha & \delta & 0 & 0 & C\alpha^2 \\ C\alpha & C\alpha & C\alpha^2 & 0 & 0 & \delta \\ C\alpha & C\alpha & C\alpha^2 & 0 & \delta & 0 \end{pmatrix} \quad (5)$$

where each row is a vector of (unnormalized) transition probabilities (ie. the first row is $P(z_i|z_{i-1} = 1)$), and a prior $P(z_0) = (C\alpha \ C\alpha \ C\alpha^2 \ 0 \ 0 \ C\alpha^2)$, with $C = \frac{1-\delta}{2\alpha+2\alpha^2}$. If we then take $P(x_i = \mathsf{H}|z_i)$ to be

1 for $z_i = 1, 3, 5$ and 0 for $z_i = 2, 4, 6$ we have a regular generating process based on repeating four "motifs": state 1 repeats H, state 2 repeats T, states 3 and 4 repeat HT, and states 5 and 6 repeat TH. δ is the probability of continuing with a motif, while α defines a prior over motifs, with the probability of producing a motif of length k proportional to α^k.

Having defined this HMM, the equivalence to DP is straightforward. For a choice of Z indicating n_1 runs and n_2 alternating sub-sequences, $P(X, Z) = \delta^{n-n_1-n_2} (\frac{1-\delta}{2\alpha+2\alpha^2})^{n_1+n_2} \alpha^{n_1+2n_2}$. Taking $P(X|\text{regular})$ to be $\max_Z P(X, Z)$, it is easy to show that $\text{random}(X) = -DP \log \alpha$ when $\delta = 0.5$ and $\alpha = \frac{\sqrt{3}-1}{2}$. By varying δ and α, we obtain a more general model: as shown in Figure 1, taking $\delta = 0.525$, $\alpha = 0.107$ gives a better fit to the data of Falk and Konold (1997), $r = 0.99$. This also addresses some of the counterintuitive predictions of DP: if $\delta > 0.5$, increasing the length of a sequence but not changing the number of runs or alternating sub-sequences reduces its randomness, since $P(X|\text{regular})$ grows faster than $P(X|\text{random})$. With the choices of δ and α given above, $\text{random}(\mathsf{HHTHT}) = 3.33$, while $\text{random}(\mathsf{HHHHHHHHHHHHHHHHHHHTHT}) = 2.61$. The effect is greater with larger values of δ.

Just as the algorithmic probability $R(X)$ is a probability distribution defined by the length of programs for a universal Turing machine, this choice of $P(X|\text{random})$ can be seen as specifying the length of "programs" for a particular finite state automaton. The output of an automaton is determined by its state sequence, just as the output of a universal Turing machine is determined by its program. However, since the state sequence is the same length as the sequence itself, this alone does not provide a meaningful measure of complexity. In our model, probability imposes a metric on state sequences, dictating a greater cost for moves between certain states. Since we find the state sequence Z most likely to have produced X, we have an analogue of Kolmogorov complexity defined on a finite state automaton.
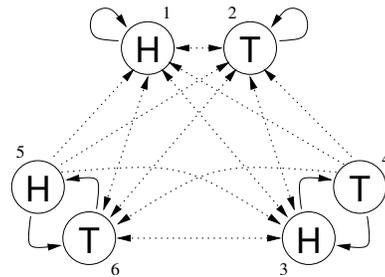

Figure 2: The finite state automaton corresponding to the HMM described in the text. Solid lines indicate motif continuation, dotted lines are permitted state changes, and numbers label the states.
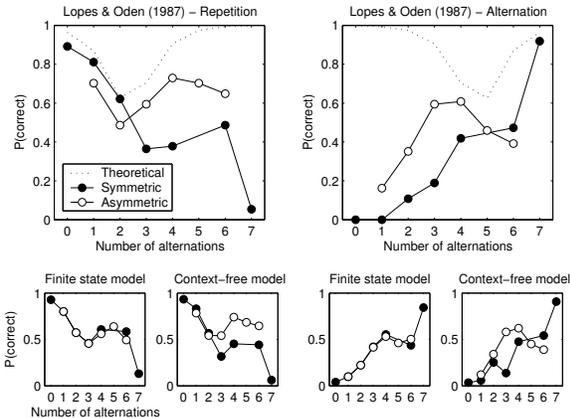
Figure 3: Results of Lopes and Oden (1987, Experiment 1) together with predictions of finite state and context-free models, illustrating effects of symmetry.

Ascending the Chomsky hierarchy

Solomonoff's (1964) contemplation of codes based upon computer programs was initiated by Noam Chomsky's talk at the 1956 Dartmouth Summer Study Group on Artificial Intelligence (Li & Vitanyi, 1997, p. 308). Chomsky was presenting a formal hierarchy of languages based upon the kinds of computing machines required to recognize them (Chomsky, 1956). Chomsky identified four types of languages, falling into a strict hierarchy: Type 3, the simplest, are regular languages, recognizable by a finite state automaton; Type 2 are context-free languages, recognizable by a push-down automaton; Type 1 are context-sensitive languages, recognizable by a Turing machine with a tape of length linearly bounded by the length of the input; Type 0 are recursively enumerable languages, recognizable by a standard Turing machine. Kolmogorov complexity is defined with respect to a universal Turing machine, capable of recognizing Type 0 languages.

There are features of regular sequences that cannot be recognized by a finite state automaton, belonging to languages on higher levels of the Chomsky hierarchy. One such feature is symmetry: the set of symmetric sequences is a classic example of a context-free language, and symmetry is known to affect subjective randomness judgments (Lopes & Oden, 1987). Here we will develop "context-free" (Type 2) and "context-sensitive" (Type 1) models that incorporate these regularities.

Lopes and Oden (1987, Experiment 1) illustrated the effects of symmetry on subjective randomness using a signal detection task, in which participants classified sequences of length 8 as being either random or non-random. Half of the sequences were generated at random, but the other half were generated by a process biased towards either repetition or alternation, depending on condition. The proportion of sequences correctly classified was ex-

amined as a function of the number of alternations and whether or not a sequence was symmetric, including both mirror symmetry (all sequences for which $x_1x_2x_3x_4 = x_8x_7x_6x_5$) and "cyclic" symmetry (HTHTHTHT, HHTTHHTT, HHHHTTTT and their complements). Their results are shown in Figure 3, together with the theoretical optimal performance that could be obtained with perfect knowledge of the processes generating the sequences. Deviations from optimal performance reflect a difference between the $P(X|\text{regular})$ implicitly used by participants and the distribution used to generate the sequences.

Our Bayesian model naturally addresses this decision problem. By the relationship between log odds and probabilities, we have

$$P(\text{random}|X) = \frac{1}{1 + \exp\{-\lambda \, \text{random}(X) - \psi\}}$$

where λ scales the effect of $\text{random}(X)$, with $\lambda = 1$ a correctly weighted Bayesian inference, and $\psi = \log \frac{P(\text{random})}{P(\text{regular})}$ is the log prior odds. Fitting the finite state model outlined in the previous section to this data gives $\delta = 0.638$, $\alpha = 0.659$, $\lambda = 0.128$, $\psi = -2.75$, and a correlation of $r = 0.90$. However, as shown in Figure 3, the model does not predict the effect of symmetry.

Accommodating symmetry requires a "context-free" model for $P(X|\text{regular})$. This model allows sequences to be generated by three methods: repetition, producing sequences with probabilities determined by the HMM introduced above; symmetry, where half of the sequence is produced by the HMM and the second half is produced by reflection; and complement symmetry, where the second half is produced by reflection and exchanging H and T. We then take $P(X|\text{regular}) = \max_{Z,M} P(X, Z|M)P(M)$, where M is the method of production. Since the two new methods of producing a sequence go beyond the capacities of a finite state automaton, this can be viewed as imposing a metric on the programs for a push-down automaton. Applying this model to the data from Lopes and Oden (1987), we obtain $\delta = 0.688, \alpha = 0.756, \lambda = 0.125, \psi = -2.99$, and probabilities of 0.437, 0.491, 0.072, for repetition, symmetry, and complement symmetry respectively, with a fit of $r = 0.98$. As shown in Figure 3, the model captures the effect of symmetry.

Duplication is a context-sensitive regularity: the set of all sequences generated by repeating a subsequence exactly twice forms a context-sensitive language. This kind of regularity can be incorporated into a "context-sensitive" model in the same way as symmetry, but the results of Lopes and Oden (1987) are at too coarse a grain to evaluate such a model. Likewise, these results do not allow us to identify whether our simple finite state model captures enough regularities: since only motifs of length 2 are included, random(THHTHHTH) is quite large.

Table 1: Log likelihood (correlation) for models

Model	4 motifs	22 motifs
Finite state	-1617.95 (0.65)	-1597.47 (0.69)
Context-free	-1577.41 (0.74)	-1553.59 (0.79)
Context-sensitive	-1555.47 (0.79)	-1531.05 (0.83)

To evaluate the contribution of these factors, we conducted an experiment testing two sets of nested models. The experiment was based on Lopes and Oden (1987, Experiment 1), asking participants to classify sequences as regular or random. One set of models used the HMM equivalent to DP, with 4 motifs and 6 states. The second used an HMM extended to allow 22 motifs (all motifs up to length 4 that were not repetitions of a smaller motif), with a total of 72 states. In each set we evaluated three models, at different levels in the Chomsky hierarchy. The finite state (Type 3) model was simplest, with four free parameters: δ, α, λ and ψ. The context-free (Type 2) model adds two parameters for the probabilities of symmetry and complement symmetry, and the context-sensitive (Type 1) model adds one more parameter for the probability of duplication. Because the three models are nested, the simpler being a special case of the more complex, we can use likelihood ratio tests to determine whether the additional parameters significantly improve the fit of the model.

Method

Participants

Participants were 20 MIT undergraduates.

Stimuli

Stimuli were sequences of heads (H) and tails (T) presented in 130 point fixed width sans-serif font on a 19" monitor at 1280×1024 pixel resolution.

Procedure

Participants were instructed that they were about to see sequences which had either been produced by a random process (flipping a fair coin) or by other processes in which the choice of heads and tails was not random, and had to classify these sequences according to their source. After a practice session, each participant classified all 128 sequences of length 8, in random order, with each sequence randomly starting with either a head or a tail. Participants took breaks at intervals of 32 sequences.

Results

We optimized the log-likelihood for all six models, with the results shown in Table 1. The model with 4 motifs consistently performed worse than the model with 22 motifs, so we will focus on the results of the latter. The context-free model gave a significant improvement over the finite state model, $\chi^2(2) = 87.76, p < 0.0001$, and the context-sensitive model
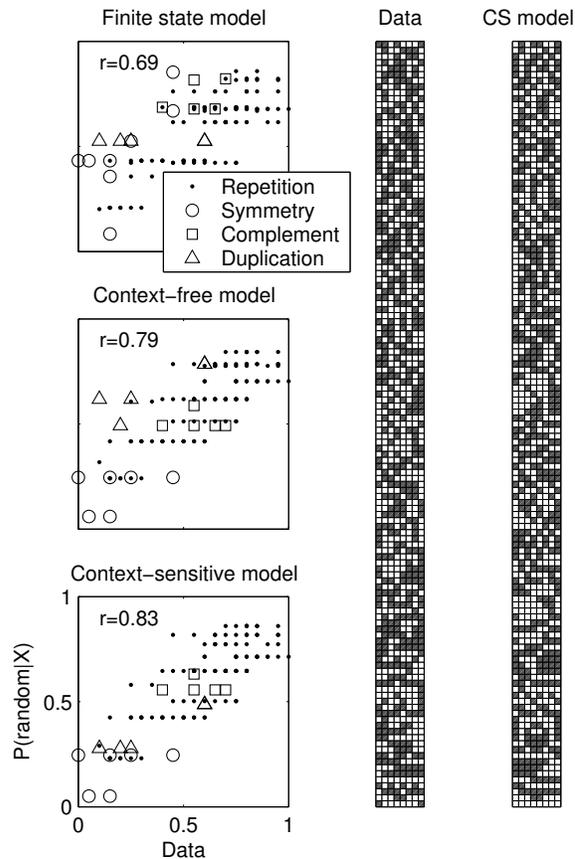

Figure 4: Scatterplots show the relationship between model predictions and data, with markers according to the process the context-sensitive model identified as generating the sequence. The arrays on the right show the sequences used in the experiment, ordered from most to least random by both human responses and the context-sensitive model.

gave a further significant improvement, $\chi^2(1) = 45.08, p < 0.0001$. Scatterplots for these three models are shown in Figure 4, together with sequences ordered by randomness based on the data and the context-sensitive model. The parameters of the context-sensitive model were $\delta = 0.706, \alpha = 0.102, \lambda = 0.532, \psi = -1.99$, with probabilities of $0.429, 0.497, 0.007, 0.067$ for repetition, symmetry, complement, and duplication. This model also accounted well for the data sets discussed earlier in the paper, obtaining correlations of $r = 0.94$ on Falk and Konold (1997) and $r = 0.91$ on Lopes and Oden (1987) using exactly the same parameter values, showing that it can give a good account of randomness judgments in general and is not just overfitting this particular data set.

The likelihood ratio tests suggest that the context-sensitive model gives the best account of human randomness judgments. Since this model has several

free parameters, we evaluated its generalization performance using a split-half procedure. We randomly split the participants in half ten times, computed the correlation between halves, and fit the context-sensitive model to one half, computing its correlation with both that half and the unseen data. The mean correlation (obtained via a Fisher z transformation) between halves was $r = 0.73$, while the model gave a mean correlation of $r = 0.77$ on the fit half and $r = 0.76$ on the unseen half. The fact that the correlation with the unseen data is higher than the correlation between halves suggests that the model is accurately extracting information about the statistical structure underlying subjective randomness.

Discussion

The results of our experiment and model-fitting suggest that the perceived randomness of binary sequences is sensitive to motif repetition, symmetry and symmetry in the complement, and duplication of a sub-sequence. In fact, these regularities can be incorporated into a statistical model that predicts human judgments better than the responses of other participants in the same experiment. These regularities can be recognized by a Turing machine with a tape of length linearly bounded by the length of the sequence, corresponding to Type 1 in the Chomsky hierarchy. Our statistical model provides a computable measure of the randomness of a sequence in the spirit of Kolmogorov complexity, but defined on a simpler computing machine.

The probabilistic approach presented in this paper provides an intuitive method for developing measures of complexity. However, we differ from existing accounts of randomness that make claims about complexity (Chater, 1999; Falk & Konold, 1997) in viewing probability as primary, and the relationship between randomness and complexity as a secondary consequence of a statistical inference comparing random generation with generation by a more regular process. This approach emphasizes the interpretation of subjective randomness in terms of a rational statistical inference, and explains why complex sequences should seem more random in terms of $P(X|\text{regular})$ being biased towards simple outcomes: random sequences are those that seem too complex to have been produced by a simple process.

Chater and Vitanyi (2003) argue that simplicity may provide a unifying principle for cognitive science. While simplicity undoubtedly plays an important role in guiding induction, being able to use these ideas in cognitive science requires developing a means of quantifying simplicity that can accommodate the kind of strong prior knowledge that human beings bring to bear on inductive problems. Kolmogorov complexity provides a universal, objective measure, and a firm foundation for this endeavour, but is very permissive in the kinds of structures it identifies as simple. We have described an approach that uses the framework of rational statistical inference to explore measures of complexity that are more restrictive than Kolmogorov complexity, while retaining the principles of algorithmic information theory by organizing these measures in terms of computability. This approach provides us with a good account of subjective randomness, and suggests that it may be possible to develop restricted measures of complexity applicable elsewhere in cognitive science.

References

Chaitin, G. J. (1969). On the length of programs for computing finite binary sequences: statistical considerations. *Journal of the ACM*, 16:145–159.

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

Chater, N. and Vitanyi, P. (2003). Simplicity: a unifying principle in cognitive science. *Trends in Cognitive Science*, 7:19–22.

Chomsky, N. (1956). Threee models for the description of language. *IRE Transactions on Information Theory*, 2:113–124.

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

Feldman, J. (2000). Minimization of boolean copmlexity in human concept learning. *Nature*, 407:630–633.

Kahneman, D. and Tversky, A. (1972). Subjective probability: A judgment of representativeness. *Cognitive Psychology*, 3:430–454.

Kolmogorov, A. N. (1965). Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1:1–7.

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

Li, M. and Vitanyi, P. (1997). *An introduction to Kolmogorov complexity and its applications*. Springer Verlag, London.

Lopes, L. L. (1982). Doing the impossible: A note on induction and the experience of randomness. *Journal of Experimental Psychology*, 8:626–636.

Lopes, L. L. and Oden, G. C. (1987). Distringuishing between random and nonrandom events. *Journal of Experimental Psychology: Learning, Memory and Cognition*, 13:392–400.

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

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

Simon, H. A. and 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 i. *Information and Control*, 7:1–22.