Effective Complexity

Murray Gell-Mann
Seth Lloyd

SFI WORKING PAPER: 2003-12-068

SANTA FE INSTITUTE

Effective Complexity

Murray Gell-Mann
Seth Lloyd

It would take a great many different concepts—or quantities—to capture all of our notions of what is meant by complexity (or its opposite, simplicity). However, the notion that corresponds most closely to what we mean by complexity in ordinary conversation and in most scientific discourse is "effective complexity." In nontechnical language, we can define the effective complexity (EC) of an entity as the length of a highly compressed description of its regularities [6, 7, 8].

For a more technical definition, we need a formal approach both to the notion of minimum description length and to the distinction between regularities and those features that are treated as random or incidental.

We can illustrate with a number of examples how EC corresponds to our intuitive notion of complexity. We may call a novel complex if it has a great many different characters, scenes, subplots, and so on, so that the regularities of the novel require a long description. The United States tax code is complex, since it is very long and each rule in it is a regularity. Neckties may be simple, like those with regimental stripes, or complex, like some of those designed by Jerry Garcia.

From time to time, an author presents a supposedly new measure of complexity (such as the "self-dissimilarity" of Wolpert and Macready [17]) without recognizing that when carefully defined it is just a special case of effective complexity.

Like some other concepts sometimes identified with complexity, the EC of an entity is context-dependent, even subjective to a considerable extent. It depends on the coarse graining (level of detail) at which the entity is described, the language used to describe it, the previous knowledge and understanding that are assumed, and, of course, the nature of the distinction made between regularity and randomness.

Like other proposed "measures of complexity," EC is most useful when comparing two entities, at least one of which has a large value of the quantity in question.

Now, how do we distinguish regular features of an entity from ones treated as random or incidental? There is, as we shall see, a way to make a nearly absolute distinction between the two kinds of features, but that approach is of limited usefulness because it always assigns very low values of EC, attributing almost all information content to the random category rather than the regular one.

In most practical cases, the distinction between regularity and randomness—or between regular and random information content—depends on some judgment of what is important and what is unimportant, even though the judge need not be human or even alive.

Take the case of neckties, as discussed above. We tacitly assumed that effective complexity would refer to the pattern of the tie, while wine stains, coffee stains, and so on, would be relegated to the domain of the random or incidental. But suppose we are dry cleaners. Then the characteristics of the stains might be the relevant regularities, while the pattern is treated as incidental.

Often, regularity and randomness are envisaged as corresponding to signal and noise, respectively, for example in the case of music and static on the radio. But, as is well known, an investigation of sources of radio static by Karl Jansky et al. (at Bell Telephone Laboratories in the 1930s) revealed that one of those sources lies in the direction of the center of our galaxy, thus preparing the way for radio astronomy. Part of what had been treated as random turned into a very important set of regularities.

It is useful to encode the description of the entity into a bit string, even though the choice of coding scheme introduces another element of context dependence. For such strings we can make use of the well-known concept of algorithmic information content (AIC), which is a kind of minimum description length.

The AIC of a bit string (and, hence, of the entity it describes) is the length of the shortest program that will cause a given universal computer U to print out the string and then halt [3, 4, 11]. Of course, the choice of U introduces yet another form of context dependence.

For strings of a particular length, the ones with the highest AIC are those with the fewest regularities. Ideally they have no regularities at all except the

length. Such strings are sometimes called "random" strings, although the terminology does not agree precisely with the usual meaning of random (stochastic, especially with equal probabilities for all alternatives). Some authors call AIC "algorithmic complexity," but it is not properly a measure of complexity, since randomness is not what we usually mean when we speak of complexity. Another name for AIC, "algorithmic randomness," is somewhat more apt.

Now we can begin to construct a technical definition of effective complexity, using AIC (or something very like it) as a minimum description length. We split the AIC of the string representing the entity into two terms, one for regularities and the other for features treated as random or incidental. The first term is then the effective complexity, the minimum description length of the regularities of the entity [8].

It is not enough to define EC as the AIC of the regularities of an entity. We must still examine how the regularities are described and distinguished from features treated as random, using the judgment of what is important. One of the best ways to exhibit regularities is the method used in statistical mechanics, say, for a classical sample of a pure gas. The detailed description of the positions and momenta of all the molecules is obviously too much information to gather, store, retrieve, or interpret. Instead, certain regularities are picked out. The entity considered—the real sample of gas—is embedded conceptually in a set of comparable samples, where the others are all imagined rather than real. The members of the set are assigned probabilities, so that we have an ensemble. The entity itself must be a typical member of the ensemble (in other words, not one with abnormally low probability). The set and its probability distribution will then reflect the regularities.

For extensive systems, the statistical-mechanical methods of Boltzmann and Gibbs, when described in modern language, amount to using the principle of maximum ignorance, as emphasized by Jaynes [9]. The ignorance measure or Shannon information I is introduced. (With a multiplicative constant, I is the entropy.) Then the probabilities in the ensemble are varied and I is maximized subject to keeping fixed certain average quantities over the ensemble. For example, if the average energy is kept fixed—and nothing else—the Maxwell-Boltzmann distribution of probabilities results.

We have, of course,

$$I = -\sum_r P_r \log P_r \,, \tag{1}$$

where log means logarithm to the base 2 and the P's are the (coarse-grained) probabilities for the individual members r of the ensemble. The multiplicative constant that yields entropy is $k \ln 2$, where k is Boltzmann's constant.

In this situation, with one real member of the ensemble and the rest imagined, the fine-grained probabilities are all zero for the members of the ensemble other than e, the entity under consideration (or the bit string describing it). Of course, the fine-grained probability of e is unity. The typicality condition

previously mentioned is just

$$-\log P_e \lesssim I\,. \tag{2}$$

Here the symbol "$\lesssim$" means "less than or equal" to within a few bits.

We can regard the quantities kept fixed (while I is maximized) as the things judged to be important. In most problems of statistical mechanics, these are, of course, the averages of familiar extensive quantities such as the energy. The choice of quantities controls the regularities expressed by the probability distribution.

In some problems, the quantities being averaged have to do with membership in a set. (For example, in Gibbs's microcanonical ensemble, we deal with the set of states having energies in a narrow interval.) In such a case, we would make use of the membership function, which is one for members of the set and zero otherwise. When the average of that function over all the members of the ensemble is one, every member with nonzero probability is in the set.

In discussing an ensemble E of bit strings used to represent the regularities of an entity, we shall apply a method that incorporates the maximizing of ignorance subject to constraints. We introduce the AIC of the ensemble and call it Y. We then have our technical definition of effective complexity: it is the value of Y for the ensemble that is finally employed. In general, then, Y is a kind of candidate for the role of effective complexity.

Besides $Y = K(E)$, the AIC of the ensemble E (for a given universal computer U), we can also consider $K(r|E)$, the contingent AIC of each member r given the ensemble. The weighted average, with probabilities P_r, of this contingent AIC can be related to I in the following way.

We note that Rüdiger Schack [15] has discussed converting any universal computer U into a corresponding U' that incorporates an efficient recoding scheme (Shannon-Fano coding). Such a scheme associates longer bit strings with less probable members of the ensemble and shorter ones with more probable members. Schack has then shown that if K is defined using U', then the average contingent AIC of the members lies between I and $I + 1$. We shall adopt his procedure and thus have

$$\sum_r P_r K(r|E) \approx I\,, \tag{3}$$

where $\approx$ means equal to within a few bits (here actually one bit).

Let us define the total information Σ as the sum of Y and I. The first term is, of course, the AIC of the ensemble and we have seen that the second is, to within a bit, the average contingent AIC of the members given the ensemble.

To throw some light on the role of the total information, consider the situation of a theoretical scientist trying to construct a theory to account for a large body of data. Suppose the theory can be represented as a probability distribution over a set of bodies of data, one of which consists of the real data and the rest of which are imagined. Then Y corresponds to the complexity of the theory and I measures the extent to which the predictions of the theory are distributed widely over different possible bodies of data. Ideally, the theorist would like both

quantities to be small, the first so as to make the theory simple and the second so as to make it focus narrowly on the real data. However, there may be trade-offs. By adding bells and whistles to the theory, along with a number of arbitrary parameters, one may be able to focus on the real data, but at the expense of complicating the theory. Similarly, by allowing appreciable probabilities for very many possible bodies of data, one may be able to get away with a simple theory. (Occasionally, of course, a theorist is fortunate enough to be able to make both Y and I small, as James Clerk Maxwell did in the case of the equations for electromagnetism.) In any case, the first desideratum is to minimize the sum of the two terms, the total information Σ. Then one can deal with the possible trade-offs.

We shall show that to within a few bits the smallest possible value of Σ is $K \equiv K(e)$, the AIC of the string representing the entity itself. Here we make use of the typicality condition (2) that the log of the (coarse-grained) probability for the entity is less than or equal to I to within a few bits. We also make use of certain abstract properties of the AIC:

$$K(A) \lesssim K(A, B) \tag{4}$$

and

$$K(A, B) \lesssim K(B) + K(A|B)\,, \tag{5}$$

where again the symbol $\lesssim$ means "less than or equal to" up to a few bits. A true information measure would, of course, obey the first relation without the caveat "up to a few bits" and would obey the second relation as an equality.

Because of efficient recoding, we have

$$K(e|E) \lesssim -\log P_e\,. \tag{6}$$

We can now prove that $K = K(e)$ is an approximate lower bound for the total information $\Sigma = K(E) + I$:

$$\begin{aligned} K &= K(e) \lesssim K(e, E)\,, && \text{(7a)}\\ K(e, E) &\lesssim K(E) + K(e|E)\,, && \text{(7b)}\\ K(e|E) &\lesssim -\log P_e\,, && \text{(7c)}\\ -\log P_e &\lesssim I\,. && \text{(7d)} \end{aligned}$$

We see, too, that when the approximate lower bound is achieved, all these approximate inequalities become approximate equalities:

$$\begin{aligned} K &\approx K(e, E)\,, && \text{(8a)}\\ K(e, E) &\approx Y + K(e|E)\,, && \text{(8b)}\\ K(e|E) &\approx -\log P_e\,, && \text{(8c)}\\ -\log P_e &\approx I\,. && \text{(8d)} \end{aligned}$$

The treatment of this in Gell-Mann and Lloyd [8] is slightly flawed. The approximate inequality (7b), although given correctly, was accidentally replaced

later on by an approximate equality, so that condition (8b) came out as a truism. Thus (8b) was omitted from the list of new conditions that hold when the total information achieves its approximate lower bound. As a result, we gave only three conditions of approximate equality instead of the four quoted here in (8a)–(8d).

Also, in the discussion at the end of the paragraph preceding eq. (2) of Gell-Mann and Lloyd [8], we wrote $\log K_U(a)$ by mistake in place of $\log K_U(b) + 2 \log \log K_U(b)$, but that does not affect any of our results.

Clearly the total information Σ achieves its approximate minimum value K for the singleton distribution, which assigns probability one to the bit string representing our entity and zero probabilities to all other strings. For that distribution, Y is about equal to K and the measure of ignorance I equals zero.

There are many other distributions for which $\Sigma \approx K$. If we plot Y against I, the line along which $Y + I = K$ is a straight line with slope minus one, with the singleton at the top of the line. We are imposing on the ensemble—the one that we actually use to define the effective complexity—the condition that the total information approximately achieve its minimum. In other words, we want to stay on the straight line or within a few bits of it.

All ensembles of which e is a typical member lie, to within a few bits, above and to the right of a boundary. That boundary coincides with our straight line all the way from the top down to a certain point, where we run out of ensembles that have $Y + I \approx K$. Below that point the actual boundary for ensembles in the $Y - I$ plane no longer follows the straight line but veers off to the right.

Now, as we discussed, we maximize the measure of ignorance I subject to staying on that straight line. If we do that and impose no other conditions, we end up at the point where the boundary in the $I - Y$ plane departs from the straight line. As described in the paper of Gell-Mann and Lloyd (who are indebted to Charles H. Bennett for many useful discussions of this manner), that point always corresponds to an effective complexity Y that is very small. If we imposed no other conditions, every entity would come out simple! In certain circumstances, that is all right, but for most problems it is an absurd result. What went wrong? The answer is that, as in statistical mechanics, we must usually impose some more conditions, fixing the values of certain average quantities treated as important by a judge. If we maximize I subject to staying (approximately) on the straight line and to keeping those values fixed, we end up with a meaningful effective complexity, which can be large in appropriate circumstances.

The situation is made easier to discuss if we narrow the universe of possible ensembles in a drastic manner suggested by Kolmogorov, one of the inventors (or discoverers?) of AIC, in work reviewed in the books by Cover and Thomas [4] and by Li and Vitányi [11]. Instead of using arbitrary probability distributions over the space of all bit strings, one restricts the ensembles to those obeying two conditions. The set must contain only strings of the same length as the original bit string and all the nonzero probabilities must be equal. In this simplified situation, every allowable ensemble can be fully characterized as a subset of the set of all bit strings that have the same length as the original one. Here I is

just the logarithm of the number of members of the subset. Also, being a typical member of the ensemble simply means belonging to the subset.

Vitányi and Li describe how, for this model problem, Kolmogorov suggested maximizing I subject only to staying on the straight line. In that case, as pointed out above, one is led immediately to the point in the $I - Y$ plane where the boundary departs from the straight line. Kolmogorov called the value of Y at that point the "minimum sufficient statistic." His student L. A. Levin (now a professor at Boston University) kept pointing out to him that this "statistic" was always small and therefore of limited utility, but the great man paid insufficient attention [10].

In the model problem, the boundary curve comes near the I axis at the point where I achieves its maximum, the string length l. At that point the subset is the entire set of strings of the same length as the one describing the entity e. Clearly, that set has a very short description and thus a very small value of Y.

What should be done, whether in this model problem or in the more general case that we discussed earlier, is to utilize the lowest point on the straight line such that the average quantities judged to be important still have their fixed values. Then Y no longer has to be tiny and the measure of ignorance I can be much less than it was for the case of no further constraints.

We have succeeded, then, in splitting K into two terms, the effective complexity and the measure of random information content, and they are equal to the values of Y and I, respectively, for the chosen ensemble. We can think of the separation of K into Y and I in terms of a distinction between a basic program (for printing out the string representing our entity) and data fed into that basic program.

We can also treat as a kind of coarse graining the passage from the original singlet distribution (in which the bit string representing the entity is the only member with nonzero probability) to an ensemble of which that bit string is a typical member. In fact, we have been labeling the probabilities in each ensemble as coarse-grained probabilities P_r. Now it often happens that one ensemble can be regarded as a coarse graining of another, as was discussed in Gell-Mann and Lloyd [8]. We can explore that situation here as it applies to ensembles that lie on or very close to the straight line $Y + I = K$.

We start from the approximate equalities (8a)–(8d) (accurate to within a few bits) that characterize an ensemble on or near the straight line. There the coarse-graining acts on initial "singleton" probabilities that are just one for the original string and zero for all others. We want to generalize the above formulae to the case of an ensemble with any initial fine-grained probability distribution $p \equiv \{p_r\}$, which gets coarse grained to yield another ensemble with probability distribution $P \equiv \{P_r\}$ and approximately the same value of Σ. We propose the

following formulae as the appropriate generalizations:

$$\begin{aligned} K(p) &\approx K(p,P)\,, && \text{(9a)}\\ K(p,P) &\approx K(P)+K(p|P)\,, && \text{(9b)}\\ K(p|P) &\approx -\Sigma_r p_r \log P_r + \Sigma_r p_r \log p_r\,, && \text{(9c)}\\ -\Sigma_r p_r \log P_r &\approx -\Sigma_r P_r \log P_r\,. && \text{(9d)} \end{aligned}$$

These equations reduce to (8a) through (8d) respectively for the case in which the fine-grained distribution is the "singleton" distribution. Also, it is easy to see that Σ is approximately conserved by these approximate equalities, as a result of our including the last term in eq. (9c).

Equation (9a) tells us that, to within a few bits, the coarse-grained probability distribution P contains only algorithmic information that is in the fine-grained distribution p. Equation (9b) tells us that the ordinary relation between joint and conditional mutual information holds here to within a few bits even though that relation does not always hold for joint and conditional *algorithmic* information.

We can compare this discussion of coarse graining to the treatment in Gell-Mann and Lloyd [8]. There we required three properties of a coarse-graining transformation from p to P: that the transformation actually yield a probability distribution, that if iterated it produce the same set of P's, and that it obey eq. (9d) above. We attained these objectives by maximizing the ignorance associated with the P's while keeping some averages involving the P's equal to the corresponding averages involving the p's (linear constraint conditions).

Here we emphasize that we are generalizing that work to the case where Y is introduced and the sum of Y and I is kept approximately fixed at its minimum value while we maximize I subject to some constraint conditions linear in the probabilities.

Say we start with the singleton ensemble in which only the original string has a nonzero probability and move down the straight line in a succession of coarse grainings until we reach the ensemble for which Y is the effective complexity. The above equations are then applied over and over again for the successive coarse grainings, and they apply also between the original (singleton) probability distribution and the final one.

Alternatively, we can, if we like, regard the transition from P to p as a fine graining, using the same formulae. We can start at the point where the boundary curve departs from the straight line and move up the line in a sequence of fine grainings. In fact, we can utilize the linear constraints successively. We apply first one of them, then that one and another, then those two and a third, and so forth, until all the constraints have been applied to the maximization of I subject to staying on the straight line. Each additional constraint yields a fine graining.

There are at least four issues that we feel require discussion at this point, even though many questions about them remain. Two of these issues relate to certain generalizations of the notion of algorithmic information content.

AIC as it stands is technically uncomputable, as shown long ago by Chaitin [3]. That is not so if we modify the definition by introducing a finite maximum execution time T within which the program must cause the modified universal computer U' to print out the bit string. Such a modification has another, more important advantage. We can vary T and, thus, explore certain situations where apparent complexity is large but effective complexity as defined above (for $T \to \infty$) is small.

Take the example [6] of energy levels of heavy nuclei. Fifty years ago, it seemed that any detailed explanation of the pattern involved would be extremely long and complicated. Today, however, we believe that an accurate calculation of the positions of all the levels is possible, in principle, using a simple theory: QCD, the quantum field theory of quarks and gluons, combined with QED, the quantum field theory of photons and electromagnetic interactions, including those of quarks. Thus, for T very large or infinite, the modified AIC of the levels is small—they are simple. But the computation time required is too long to permit the calculations to be performed using existing hardware and software. Thus, for moderate values of T the levels appear complex.

In such a case, the time around which the modified AIC declines from a large value to a small one (as T increases) is related to "logical depth" as defined by Charles H. Bennett [2]. Roughly, logical depth is the time (or number of steps) necessary for a program to cause U to print out the coded description of an entity and then halt, averaged over programs in such a way as to emphasize short ones.

There are cases where the modified AIC declines, as T increases, in a sequence of steps or plateaus. In that case we can say that certain kinds of regularities are buried more deeply than others.

While it is very instructive to vary T in connnection with generalizing K—the AIC of the bit string describing our entity—we encounter problems if we try to utilize a finite value of T in our whole discussion of breaking up K into effective complexity and random information. Not all the theorems that allow us to treat AIC as an approximate information measure apply to the generalization with variable T.

In addition to logical depth, we can utilize a quantity that is, in a sense, inverse to it, namely Bennett's "crypticity," [2] which is, in rough terms, the time necessary to go from the description of an entity to a short program that yields that description. As an example of a situation where crypticity is important, consider a discussion of pseudorandomness. These days, when random numbers are called for in a calculation, one often uses instead a random-looking sequence of numbers produced by a deterministic process. Such a pseudorandom sequence typically has a great deal of crypticity. A lengthy investigation of the sequence could reveal its deterministic nature and, if it is generated by a short program, could correctly assign to it a very low AIC. Given only a modest time, however, we could fail to identify the sequence as one generated by a simple deterministic process and mistake it for a truly random sequence with a high value of AIC.

The concept of crypticity can also be usefully applied to situations where a bit string of modest AIC appears to exhibit large AIC in the form of effective complexity rather than random information. We might call such a string "pseudocomplex." An example of a pseudocomplex string would be one recording an image, at a certain scale, of the Mandelbrot set. Another would be an apparently complex pattern generated by a simple cellular automaton from a simple initial condition. Note that a pseudorandom string, which has passed many of the usual statistical tests for randomness, is not appreciably compressed by conventional data compression algorithms, such as the one known as LZW [4]. By contrast, a pseudocomplex string typically possesses a large number of obvious statistical regularities and is, therefore, readily compressible to some extent by LZW, but not all the way to the very short program that actually generated the string.

We should mention that a number of authors have considered mutual information as a measure of complexity in the context of dynamical systems [1, 5, 12]. Without modification, that idea presents a conflict with our intuitive notion of complexity. Consider two identical very long bit strings consisting entirely of ones. The mutual information between them is very large, yet each is obviously very simple. Moreover, the statement that they are the same is also very simple. The pair of strings is not at all complex in any usual sense of the word.

Typically, the authors in question have recognized that a more acceptable quantity in a discussion of complexity is mutual algorithmic information, defined for two strings as the sum of their AIC values minus the AIC of the two taken together. If two strings are simple and identical, though very long, their mutual AIC is small.

Of course, identical long strings could be "random," in which case their very large mutual algorithmic information does not correspond to what we usually mean by complexity. EC is still the best measure of complexity.

We can easily generalize the definition of mutual information to the case of any number of strings (or entities described by them). For example, for three strings we have

$$K_{\rm mut} = K(1) + K(2) + K(3) - K(1,2) - K(2,3) - K(1,3) + K(1,2,3)\,. \quad (10)$$

Under certain conditions we can see a connection between mutual algorithmic information and effective complexity. For example, suppose we are presented not with a single entity but with N entities that are selected at random from among the typical members of a particular ensemble. The mutual algorithmic information content among these entities is then a good estimate of the AIC of the ensemble from which they are selected, and that quantity is, under suitable conditions, equal to the effective complexity candidate Y attributed to each of the entities.

The way the calculation goes is roughly the following. On average the K value for m arguments is approximately $Y + mI$, and the sum in eq. (10) then comes out equal to Y. It is easily shown that such an equality yielding Y holds not just for three entities but for any number N, with the appropriate generalization

of eq. (10). The elimination of the I term produces the connection of $K_{\rm mut}$ with the effective complexity candidate.

At last we arrive at the questions relevant to a nontraditional measure of ignorance. Suppose that for some reason we are dealing, in the definition of I, not with the usual measure given in eq. (1), but rather with the generalization discussed in this volume, namely

$$I_q = -\frac{[\Sigma_r (P_r)^q - 1]}{(q-1)}\,, \tag{11}$$

which reduces to eq. (1) in the limit where q approaches 1. Should we be maximizing this measure of ignorance—while keeping certain average quantities fixed—in order to arrive at a suitable ensemble? (Presumably we average using not the probabilities P_r but their qth powers normalized so as to sum to unity—the so-called Escort probabilities.) Do we, while maximizing I, keep a measure of total information at its minimum value? Is a nonlinear term added to $I + Y$? What happens to the lower bound on $I + Y$? Can we make appropriate changes in the definition of AIC that will preserve or suitably generalize the relations we discuss here? What happens to the approximate equality of I and the average contingent AIC (given the ensemble)? What becomes of the four conditions in eqs. (8a) to (8d)? What happens to the corresponding conditions (9a) to (9d) for the case where we are coarse graining one probability distribution and thus obtaining another one?

As is well known, a kind of entropy based on the generalized information or ignorance of eq. (11) has been suggested [16] as the basis for a full-blown alternative, valid for certain situations, to the "thermostatistics" (thermodynamics and statistical mechanics) of Boltzmann and Gibbs. (The latter is, of course, founded on eq. (1) as the formula for information or ignorance.) Such a basic interpretation of eq. (11) has been criticized by authors such as Luzzi et al. [13] and Nauenberg [14]. We do not address those criticisms here, but should they prove justified—in whole or in part—they need not rule out, at a practical level, the applicability of eq. (11) to a variety of cases, such as systems of particles attracted by $1/r^2$ forces or systems at the so-called "edge of chaos."

ACKNOWLEDGMENTS

This research was supported by the National Science Foundation under the Nanoscale Modeling and Simulation initiative. In addition, the work of Murray Gell-Mann was supported by the C.O.U.Q. Foundation and by Insight Venture Management. The generous help provided by these organizations is gratefully acknowledged.

REFERENCES

[1] Adami, C., C. Ofria, and T. C. Collier. "Evolution of Biological Complexity." *PNAS (USA)* **97** (2000): 4463–4468.

[2] Bennett, C. H. "Dissipation, Information, Computational Complexity and the Definition of Organization." In *Emerging Syntheses in Science*, edited by D. Pines, 215–234. Santa Fe Institute Studies in the Sciences of Complexity, Proc. Vol. I. Redwood City: Addison-Wesley, 1987.

[3] Chaitin, G. J. *Information, Randomness, and Incompleteness*. Singapore: World Scientific, 1987.

[4] Cover, T. M., and J. A. Thomas. *Elements of Information Theory*. New York: Wiley, 1991.

[5] Crutchfield, J. P., and K. Young. "Inferring Statistical Complexity." *Phys. Rev. Lett.* **63** (1989): 105–108.

[6] Gell-Mann, M. *The Quark and the Jaguar*. New York: W. H. Freeman, 1994.

[7] Gell-Mann, M. "What is Complexity?" *Complexity* **1/1** (1995): 16–19.

[8] Gell–Mann, M., and S. Lloyd. "Information Measures, Effective Complexity, and Total Information." *Complexity* **2/1** (1996): 44–52.

[9] Jaynes, E. T. *Papers on Probability, Statistics and Statistical Physics*, edited by R. D. Rosenkrantz. Reidel: Dordrecht, 1982.

[10] Levin, L. A. Personal communication, 2000.

[11] Li, M., and P. M. B. Vitanyi. *An Introduction to Kolmogorov Complexity and Its Applications*. New York: Springer-Verlag, 1993.

[12] Lloyd, S., and H. Pagels. "Complexity as Thermodynamic Depth." *Ann. Phys.* **188** (1988): 186–213.

[13] Luzzi, R., A. R. Vasconcellos, and J. G. Ramos. "On the Question of the So-Called Non-Extensive Thermodynamics." IFGW-UNICAMP Internal Report, Universidade Estadual de Campinas, Campinas, Sao Paulo, Brasil, 2002.

[14] Nauenberg, M. "A Critique of Nonextensive q-Entropy for Thermal Statistics. Dec. 2002. lanl.gov e-Print Archive, Quantum Physics, Cornell University. ⟨http://eprints.lanl.gov/abs/cond-mat/0210561⟩.

[15] Schack, R. "Algorithmic Information and Simplicity in Statistical Physics." *Intl. J. Theor. Phys.* **36** (1997): 209–226.

[16] Tsallis, C. "Possible Generalization of Boltzmann-Gibbs Statistics." *J. Stat. Phys.* **52** (1988): 479–487.

[17] Wolpert, D. H., and W. G. Macready. "Self-Dissimmilarity: An Empirically Observable Measure of Complexity." In *Unifying Themes in Complex Systems: Proceedings of the First NECSI International Conference*, edited by Y. Bar-Yam, 626–643. Cambridge, Perseus, 2002.