Toward an Information Morality: Imperatives Derived from a Statistical Mechanics of Meaning

Robert Adámy Duisberg

March 2, 2009

1 Introduction

The evolution of the cosmos describes the great arc from the initial singularity onward to the ultimate end of the universe. The Big Bang in its extreme improbability is effectively the zero point of entropy from which the universe unwinds, entropy ever increasing. However, within the closed system of our planet and its diaphanous aura, the steady influx of our suns energy is able to locally drive the entropy down, consistent with thermodynamics second law, leading to the improbable, myriad order of life on earth. This ordering of life is manifest not only physically, but also in the cognitive capacities of sentient beingsin their creation of mental representations of their experience. These cognitive patterns too constitute diminutions in the ineluctable flux of entropy.

We strive to make sense of the world, to find patterns and discern form in our perceptions of experience. Our understanding of how we manage to do this, to construct meaningful relations out of incoming streams of sensations, has deepened in recent years through research in neurology and artificial neural networks of the kind which now informs cognitive psychology. This paper proposes a measure of such "meaningful relations" as represented in cognitive systems. Having such a means of evaluation suggests implications of applying such a measure to our actions in this world, that is it can provide a basis for normative, objective values, a form of "moral realism" informing, for example, choices between preservation versus consumption.

2 The Mathematical Theory of Information

To understand how we make sense out of information, we may begin with our best understanding of information itself. Information Theory, as formalized by Claude Shannon[6], has valuable applications in communications. For example, its theorems are useful in the design of optimal digital encoding and compression algorithms for data transmission. In this theory, information is defined in terms of the statistics of symbols in messages, in a way that turns out to be mathematically isomorphic to the definition of entropy that comes out of statistical thermodynamics. Specifically, information is defined to be equal to the negative entropy—the "surprise value"—of symbols in messages.

To understand what this means, consider a source of messages M, sending out messages consisting of symbols chosen from a given alphabet of n symbols x_i, where $i = 1, ..., n$, each with a probability $p(x_i)$ of occurring anywhere in a message. For example, in English $p(x =$'**e**'$) > p(x =$'**q**'$)$. When we read any new symbol in a message, the "surprisal" of the new symbol is taken to be just the inverse of its probability, $1/p(x_i)$. The more unlikely a symbol is to occur, the more surprised we are when it does. The information entropy of the message source as a whole is then taken to be the expectation value for this sense of surprise averaged over all symbols in a large sample of messages, thus:

$$H(M) = \sum_{i=1}^{n} p(x_i) \log(1/p(x_i)) = -\sum_{i=1}^{n} p(x_i) \log p(x_i).$$

This expresses a value of the average surprise imparted by each symbol in the stream as a whole. It is usually measured in "bits" when the logarithms are chosen to be of base 2, and the conventional symbol H is borrowed from thermodynamics where it signifies entropy.

So for example, if the data stream comprises an endless repetition of one character, 'b' say, there is no surprise at all when the next character in the stream arrives. The arrival of 'b' is a complete certainty; the probability of reading 'b' is one. Since $\log(1) = 0$, and since $p(x_i) = 0$ for all other symbols but 'b,' the value of the sum above is zero—this corresponds to the *maximum* level of entropy (general sameness), from which it decreases (going negative) for more unpredictable signals. Surprisal and information then increase according to the minus sign above.

For an example at the other extreme, if in another message source the occurrence of each character in the ASCII alphabet is equiprobable, (i.e. if each character were found $(1/128)^{th}$ of the time), then each successive character will convey a full seven bits of information (an ASCII character is an 8-bit byte with 1 bit reserved as a parity bit). As we can see, the value of the sum is:

$$H = -\sum_{i=1}^{128} \frac{1}{128} \log_2 \left(\frac{1}{128} \right) = \log_2 128 = 7.$$

But consideration of these two extrema of information content point up a sharp disjunction between the information in a message and how meaningful it is, since clearly the upper bound of maximal information corresponds to a condition of gibberish—what might be expected from the proverbial monkeys at typewriters. In fact Shannon estimated the information rate of English text to be between 0.6 and 1.3 bits per character, far below our 7 bit maximum. This low information rate for meaningful messages reflects the redundancy that necessarily arises from the internal coherence and linguistic structure that are related to meaning. This limitation of information theory, that the measure of information says nothing about its import or significance, suggests that it may be useful to devise a measure, analogous to the statistical measure of information, that can begin to assess and evaluate elements involved in creating meaning in communications. Let us consider some systems that display abilities relating to this question.

3 Artificial Neural Networks

Research in a relatively new branch of Artificial Intelligence has patterned an entire family of mathematical models after what is known about biological neural networks.[2, 5] These offer an alternative methodology to that of "classical AI," with its emphasis on *a priori* rule-bases or "expert systems," providing particularly efficacious applications aimed at pattern recognition in variable, noisy or incomplete data sets. Systems based on artificial neural networks find application in such areas as machine recognition of handwritten characters, and feature detection in photographs or voice recordings.

Common to all members of this family of models is the notion from graph theory of an array of N nodes, often suggestively referred to as "neurons," formed into a network by a set of connections or directed arcs, also called "axons" or "synapses" in keeping with the metaphor. The arcs are labeled with values called "weights" that represent the strength of synaptic connections, which can be either excitatory or inhibitory. The weights compose a matrix $\mathbf{W}$ of values w_{ij} between each pair of nodes i and j, where $0 < i, j \leq N$. Each node i has an associated "activation level," a_i, representing its state of excitation. Each node's excitation is in turn considered to be dependent upon its incoming "stimulus," computed in a straightforward way as the aggregate of the activations of all the other nodes connected to it, weighted by the strength of each such connection, w_{ji} from the j-th node to the i-th. This weighted sum is typically passed through some thresholding function to give the activation level of the node in question.

$$a_i(t) = F\left(\sum_{j \neq i}^{N} w_{ji} a_j(t-1)\right)$$

The time dependence, (t), of the activation levels indicates the manner in which activations tend to propagate through the network, in that a current activation is derived from previous activations of the input nodes. In artificial networks, the propagation dynamics are often synchronously "clocked" through the network in discrete time increments, while in truly distributed implementations, as well as natural systems, the dynamics are likely to be asynchronous. The thresholding function $F(x)$ is sometimes taken to be a step function, as if the nodes were simple switches, but more often it is chosen to be a sigmoid function (so named for its S-like shape, smoothing out the corners of the step function) of the form $F(x) = 1/(1 + e^{-x})$, the differentiability of which is advantageous for algorithms by which these networks "learn." A variety of learning algorithms exist, typically involving the propagation of some feedback from experience or training back through the network to enhance the connection weights that have contributed to correct or advantageous decisions, and to diminish the weights on connections that contribute to errors in recognition.

Whatever the process of learning in either artificial or natural cognitive networks, its results are the modification of connection strengths between neurons so that the resulting networks are characterized by clusterings of nodes that are mutually excitatory. It is this quality in trained artificial neural networks that affords them behaviors we value, such as being able to correctly identify partial or variable inputs; if only part of a mutually excitatory cluster is stimulated, the whole will nonetheless tend to become active. Marvin Minsky, in the model he describes as a "society of mind," roughly associates such mutually excitatory clustering with representations of concepts or fundamental percepts[4]. We may recognize in this our own ability to respond to synecdoche, the poetic device in which reference to a part evokes the whole.

Similar phenomena can be observed as learning occurs in artificial networks. For example, in a system that has learned to recognize handwritten characters, a common subset may exists among the internal nodes that become activated during the recognition of the letters 'A' and 'H.' This common cluster may be associated with the perception of the middle cross-bar in both characters. Thus it is as if, through its exposure to the training data set, the learning algorithm has enabled discernment of the salient components in the data stream upon which to base its decisions, essentially expressing logic of the form "if the character has a cross-bar with vertical side-bars, it is probably an 'H,' but a cross-bar with side-bars that slant in toward the top is more likely an 'A'. "

Note well in the above example that the "cross-bar percept" was not pre-programmed into the system. Rather, the salient structure in the information stream impressed itself into the network through learning recognition, and finds itself represented in the network as a self-exciting cluster of nodes.

Notice also that the clustering of nodes in a network can be considered as a statistical property of the associated graph.[1] If we imagine an initial state as a fully connected graph in which all connection weights are equal, an effective *tabula rasa* with maximal entropy, and compare this to a well tutored network, we expect to see that salient concepts in the tutorial information stream have become represented as changes in the distribution of weights, effectively implementing the self-excitation of the cluster. This ordering of the weights will be manifest as a reduction in a form of entropy.

Thus the passage of information through a responsive cognitive network can be seen to physically lower the entropy of the network itself, as concepts become represented in mutually excitatory sub-networks. This is analogous to the reduction in entropy of a thermodynamic system as energy is passed through it, indeed just as our stream of solar energy has made possible the ordering of life on earth.

4 The Concept of Salience in a Cognitive Network

We saw that information theory gives us a useful measure of information content, but tells us nothing about the meaning in the messages. A measure of the entropy reduction over a matrix of weights can reflect the structuring of the cognitive network so represented, when it has learned to recognize what is meaningful in the information environment it experiences. This reduction in entropy is due to the clusterings of weights associated with the learning of salient concepts from experience, and provides a means of quantifying that quality. We may define a quantity, call it “salience,” which is a direct measure of this organizing of a network as it learns to represent the salient elements in its information environment via mutually excitatory weighting clusters.

Having a measure of a quantity establishes the basis for the ability to evaluate, literally to associate a value with that property. So the importance of a measure of this kind is that it gives us a quantitative way to value the encoding of relationships associated with the development of meaningful understanding. Thus ascribing a value to order and our understanding of it, in a consideration of values or ethics may stand as a counterweight to, for example, mere economic evaluation.

Formally, a definition of “salience” grows out of an extension to Shannon’s approach, by looking at the probability distributions not of the frequencies of symbols in a message, but of the values of the synaptic strengths in a cognitive network that has learned to understand (i.e. correctly categorize) elements of an information stream. Also our new entropy measure must be over a two dimensional matrix rather than over Shannon’s one dimensional channel. For simplicity let the allowed values for the w_{ij} be restricted to a domain of K discrete values $\{w_1, ..., w_K\}$ (otherwise the outer sum below could be replaced by an integral). Then we have,

$$S(\mathbf{W}) = -\sum_{k=1}^{K}\sum_{i\neq j}^{N} p(w_{ij} = w_k)\log p(w_{ij} = w_k)$$

where $p(w_{ij} = w_k)$ is the probability that a given matrix element in $\mathbf{W}$ has a particular value w_k, in the domain of allowable values for synaptic weights. S is the common symbol for thermodynamic entropy, and is suggested here to stand for "Salience."

According to this measure then, the aforementioned *tabula rasa* case will have value of $S = 0$, again because the certainty that value of $w_{ij} = 1, \forall i \neq j$, by definition.

Consider by contrast, an end case in which the network has condensed into a set of loosely connected, internally tight sub-clusters, in what Barabási calls a "small worlds model." With a suitable change of basis, the matrix can therefore be arranged so that it has an appearance resembling that sketched below (in which we make the simplification that the domain of weights is $\{0, 1\}$).

$$\mathbf{W} = \begin{pmatrix} 0 & 1 & 1 & \ldots & 0 & \ldots \\ 1 & 0 & 1 & \ldots & & \ldots \\ 1 & 1 & 0 & 1 & 1 & \ldots \\ & \ldots & 1 & 0 & 1 & \ldots \\ 0 & \ldots & 1 & 1 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots \end{pmatrix}.$$

In this fragment there is shown a cluster size three nodes per cluster. The matrix has a total of $(N^2 - N)$ relevant weights (excluding the diagonal elements, since nodes are not considered to connect to themselves). We can easily see in this case, by counting six ones in each cluster around three diagonal zeros, that the number of nodes with value 1 is $(6 \times \frac{N}{3}) = 2N$. Therefore,

$$p(w_{ij} = 1) = \frac{2N}{N^2 - N} = \frac{2}{N-1}, \quad p(w_{ij} = 0) = 1 - \frac{2}{N-1},$$

and, for large N,

$$\log p(w_{ij} = 1) = 1 - \log(N-1), \quad \log p(w_{ij} = 0) = \log(1 - \frac{2}{N-1}) \to 0.$$

Then in this simplified case, again for large N

$$S(\mathbf{W}) = -(N^2 - N)\left(\frac{2}{N-1}\right)(1 - \log(N-1)) = 2N\left(\log(N-1) - 1\right)$$
$$\approx 2N \log N.$$

This displays a marked increase in the value of salience as the matrix has condensed, through learning, into a set of self-exciting clusters.

5 Conclusions

It follows from a statistical definition of meaning that the reductions in physical entropy associated with sentience and the cognitive awareness of meaningful relations are fundamental, universal, and measurable values. If we lose these ordered states assembled in their myriad diversity, whether through extinctions or environmental destruction or book burning or museum looting, it is a universal loss. We can assert a fundamental and measurable value associated with that loss, and this measurement affords such losses a moral weight in ethical questions of value and conservation.

References

[1] Albert, R., Barabási, A.-L. (2002), “The Statistical Mechanics of Complex Networks,” *Reviews of Modern Physics*, 74, pp. 47-97.

[2] De Wilde, P. (1997), Neural Network Models, 2nd ed., Springer Verlag, NY.

[3] Lewis, T., Amini, F., Lannon, R. (2000), A General Theory of Love, Random House, NY.

[4] Minsky, M. (1986), The Society of Mind, Simon & Schuster, NY.

[5] Rumelhart, D.E., McClelland, J.L., *et al* (1986), Parallel Distributed Processing: Explorations in the Microstructure of Cognition, MIT Press, Cambridge MA.

[6] Shannon, C.E. (1948), “A Mathematical Theory of Communication,” *Bell System Technical Journal*, 27, pp. 379–423 & 623–656, July & October, 1948.