1989

Information Theory

Klaus Krippendorff
University of Pennsylvania, kkrippendorff@asc.upenn.edu

Information Theory

INFORMATION THEORY

A calculus capable of accounting for variation and information flow within systems regardless of whether they are biological, social, or technical. Information theory is characterized by a few axioms from which many measuring functions, accounting equations, theorems, limits, and, above all, its notion of information and communication can be derived. The information theorist treats quantities of information much like a physicist traces energy uses and losses within a mechanical system or an accountant measures cash flows and capital distributions within a company. Although quantities of information do not behave like energy and matter and have little to do with truth or value, once information flows are assessed they can be related to and shed light on other organizational features of the system in which such flows are observed.

Origins

The idea of information theory emerged in the late 1940s and came to several researchers virtually independently. NORBERT WIENER, the founder of CYBERNETICS (the theory of communication and control in humans and machines), came to it while working on statistical aspects of communication engineering. Soviet mathematician A. N. Kolmogoroff came to it from probability theory, and CLAUDE SHANNON of the Bell Telephone Laboratories in the United States developed it while working on problems of coding and deciphering messages. Earlier, British statistician R. A. Fisher, known for his analysis of variance, suggested a quantitative expression for the amount of information an experiment provides. Nearly a century before all four of them, Austrian physicist Ludwig Boltzmann had measured thermodynamic entropy by a function that resembles the one now used in information theory. However, it was Shannon who published the most elaborate account of the theory in 1948, offering proof of the uniqueness of its form and twenty-one theorems of considerable generality. WARREN WEAVER anticipated that any theory clarifying the understanding of information and communication was certain to affect all fields of knowledge. He gave a popular account of Shannon's work and coauthored with him *The Mathematical Theory of Communication* (1949). Subsequently, U.S. statistician Solomon Kullback linked information theory to statistics, and British cybernetician W. Ross Ashby generalized it to many variables.

Historically, information theory was a major stimulus to the development of communication research. It made the heretofore vague notions of information mathematically tractable, liberated it from the conflicting claims by diverse disciplines concerned with knowledge and communication technology, and legitimized research on communication and information processes whether they occurred in society, in electronic information systems, or within the human brain.

Three versions of the theory are discussed here: the possibilistic and semantic theory of information, the probabilistic or statistical theory of communication, and its extension to a method for testing complex models of qualitative data.

Semantic Information

The semantic theory quantifies information in ways similar to ordinary uses of the term: we might judge one report to be *more* informative than another, we might experience *how little* we can say in a telegram, and we might admit to having *not enough* information to decide how to resolve an issue. To obtain information we may ask questions. Questions admit uncertainty and are designed to elicit answers that help the questioner decide among several uncertain possibilities. The knower selects an answer from a repertoire of possible responses. The questioner decides what that answer means and which uncertain alternatives it thereby excludes. Information is always selective among a set of preconceived alternatives, and the theory quantifies this selectivity in terms of the number of questions we need to have answered.

The semantic theory presupposes a distinction between two sets of elements, languages, or symbol repertoires, connected by a code. One contains the set of messages, answers to questions, statements, or meaningful actions exchanged; the other contains the set of meanings, referents, things, people, ideas, concepts, or consequences the former refer to, indicate, or are about. The semantic theory suggests that information is manifest in what the elements in one set imply about those in the other set. From the point of view of the questioner or receiver the theory

expresses the amount of information, I, a message conveys as the *difference between two states of uncertainty, U,* before and after that message became known:

$$I\left(\text{message}\middle|\begin{matrix}\text{state of}\\ \text{knowledge}\end{matrix}\right) = U\left(\begin{matrix}\text{before receipt}\\ \text{of message}\end{matrix}\right) - U\left(\begin{matrix}\text{after receipt}\\ \text{of message}\end{matrix}\right)$$

The message is an element in one set; the uncertainties concern elements in the other set, for example, the interpretations such messages could have; and the amount of information indicates the selectivity that a message induces within the domain of possible interpretations.

Accordingly, information is *positive* when a message, answer, or report reduces the receiver's uncertainty about what he or she wishes to know. A sequence of informative messages, such as would be received during an interview or a conversation, reduces the receiver's uncertainty or enhances his or her state of knowledge stepwise and results in *additive* quantities of information associated with each message. A message whose content is already known does not alter the receiver's uncertainty and is *redundant,* simple repetition being one example. A message that says something unrelated to what the receiver needs to know is *irrelevant.* A message that denies what previously appeared certain and thus increases the receiver's uncertainty conveys *negative* amounts of information. Except for some syntactic limitations, the formal complexity or material composition of the message does not enter the definition of information and does not affect what or how much it conveys. Semantic information measures not what a message is but what it does in someone's cognitive system of distinctions.

The unit of measurement in information theory equals the amount the answer to a yes-or-no question conveys and is called one *bit* (for *bi*nary dig*it*). Since N alternatives can be exhaustively distinguished by $\log_2 N$ yes-or-no questions, the state of uncertainty becomes simply $U = \log_2 N$ bits. Thus, if U is an integer, U equals the number of times N alternatives can be divided in half until only one alternative remains. The remainder is elementary algebra:

$$\begin{aligned} I\left(\text{message}\middle|\begin{matrix}\text{state of}\\ \text{knowledge}\end{matrix}\right) &= \log_2 N_{\text{before message}} - \log_2 N_{\text{after message}} \\ &= -\log_2 \frac{N_{\text{after message}}}{N_{\text{before message}}} \\ &= -\log_2 P_{\text{after|before}} \end{aligned}$$

Thus information—the difference between two states of uncertainty—is seen to be a *measure of the constraint* a message imposes by singling out a subset of the initial number of uncertain possibilities N. With P as the logical probability of this subset, it may also be interpreted as a *measure of the difficulty of selecting* among a set of alternatives *by chance* and thus becomes equated with that message's surprise value. For example, because ignorant students can answer 50 percent of all yes-or-no questions correctly merely by choosing at random, teachers expect that knowledgeable students will perform significantly above that logical probability. Therefore, the semantic theory can also be seen to equate information with choices that deviate from what would be expected under conditions of ignorance.

When the alternatives are enumerable, information theory offers a precise instrument for quantification. The answer to the question "Did she have a boy or a girl?" conveys one bit of information. To make appropriate choices among eight different subway trains requires three bits of information. To locate one criminal among, say, a million Bostonians requires nearly twenty bits of information, which is the minimum amount that Boston's police department has to process per individual crime. A Hollerith card with eighty columns by twelve rows, whose positions may be either punched or not, can store up to 960 bits of information. Two such cards can store twice as much. According to Bremmermann's Limit, which states that no computer can do better than 10^{47} bits per second and per gram of its mass, the limit on computability on earth is about 10^{72} bits and is not achievable in practice.

When the alternatives are less clear or known only in relation to each other, the theory offers possibilities of quantitative comparisons. The statement "She plays a stringed instrument" conveys three to four bits less information than one asserting that "She plays the viola," because the former leaves uncertain which stringed instrument she plays. For the same reason, "about noon" conveys less than "at 12:03 P.M.," although the additional quantity conveyed by minutes may be irrelevant in a particular situation. Information quantities can also be associated with the logical structure of complex messages. For example, two statements connected by an inclusive *or* convey less information than either statement does by itself; when the logical conjunction *and* is used to connect them, they are more informative together than either is alone.

Note several properties of the semantic theory. First, quantities of information are not tied to physical entities. The length of the silence between the signals of the Morse code is as critical as the absence of a letter from a friend is informative to the usual receiver.

Second, quantities of information are always expressed *relative to someone's cognitive system of*

distinctions, including the distinctions an "objective" measuring instrument makes for a scientific observer. An X-ray photograph may be more informative to the physician than it is to the patient precisely because the former tends to have a more elaborated conceptual system and language for interpreting such images. It follows that one can observe that person A said X to person B, where X is a vehicle of communication or the material form of a message, but it is only when the codes relating the cognitive systems of A and B to X are known that one can assert how much semantic information A communicated to B.

Third, quantities of information are always *contextual* measures. They are not attributable to a single message but express what this message does in the context of all possible messages or conditions. The larger this repertoire, the greater the amount of information a particular message may convey and its receiver needs to process in order to make an appropriate selection. Where there are no options there is no information. When the context of communication is not clearly understood, quantities of information may become at best approximate.

Fourth, a valuable oddity of the theory is that paradoxical or contradictory messages turn out to convey quantities of information that are infinite, indicating the logical inadequacy or powerlessness of a cognitive system to cope with such messages. This is particularly true when messages are *self-referential.* For example, "Ignore this command" asserts how it is to be taken, its illocutionary force or its truth value, and is impossible in a system that insists on the distinction or asymmetry between language and action or between statements and what these statements refer to. *See also* SEMANTICS.

Statistical Theory of Communication

In the mathematical theory of communication the statistical analog of uncertainty is called *entropy* and is defined by the famous Shannon-Wiener formula:

$$H(A) = -\sum_{a \in A} p_a \log_2 p_a$$

where the variable A consists of mutually exclusive categories, values, or symbols a, and p_a is the probability with which a is observed in A. The entropy is a measure of variability or diversity not unlike the statistical concept of variance, except that it does not require variables to express magnitudes and is hence entirely general. When all observations fall into one category the entropy is zero; otherwise it is a positive quantity whose maximum depends on the number of distinctions drawn within a sample.

In social research, entropy measures have served to assess occupational diversity in cities, the variability of television programming, the consensus on preferences for political candidates, the specificity of financial reports, the diversity of opinions, and the richness of vocabularies. Entropy measures may be used comparatively, for example, to differentiate between different genres of literature (newspaper English is low in entropy compared with avant-garde poetry); or they may be correlated with other variables, for example, to ascertain how diversity of opinion is related to number of newspapers serving a community or to predict the reading ease of a text. However, taking full advantage of the additivity of entropy and information quantities, the theory's most important contribution is the calculus it defines on top of such entropies. Already the relationship between entropy and the aforementioned uncertainty is instructive in this regard.

When there are N_A alternatives a and each is observed the same number of times, that is, $p_a = 1/N_A$, then in this special case, the entropy equals the uncertainty, $H(A) = U(A) = \log_2 N_A$. When n individual observations are differentiated into mutually exclusive classes $a = 1,2, \ldots,$ so that $n = n_1 + n_2 + \ldots$ and $p_a = n_a/n$, then

$$H(A) = \sum_{a \in A} \frac{n_a}{n} (\log_2 n - \log_2 n_a)$$

in which $\log_2 n$ is the quantity of uncertainty in the sample of size n with each observation considered unique, $\log_2 n_a$ is the quantity to which the uncertainty reduces after knowing an observation to be of type a, and $\Sigma\, n_a/n$ renders the expression as an average reduction of uncertainty. Thus the entropy $H(A)$ is the average uncertainty or diversity in a sample when its n observations are considered in categories. The entropy formula is the same whether one considers the entropy in one variable, A, in a matrix of two variables, say, A and B, or in a cross-tabulation of many variables $A, B, C, \ldots, Z$:

$$H(ABC \ldots Z) = -\sum_a \sum_b \sum_c \ldots \sum_z p_{abc\ldots z} \log_2 p_{abc\ldots z}$$

The mathematical theory of communication relates a sender, who emits symbols a from a set A with a certain entropy $H(A)$, to a receiver, who receives symbols b from a set B with a certain entropy $H(B)$, by means of a channel that converts input symbols a into output symbols b and associates a probability with each transition. In the ideal channel, symbols sent and symbols received are related one-to-one (Figure 1*a*). Variation at the receiver for which the sender does not account is called *noise* and is manifest in one-to-many relations (Figure 1*b*). Variation at the sender omitted by the receiver is called *equivocation* and is manifest in many-to-one relations

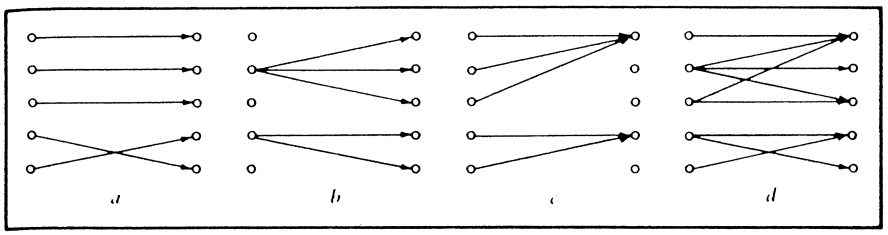

Figure 1. *(Information Theory)* Four examples of symbol-transition diagrams: *(a)* error free; *(b)* noise only; *(c)* equivocation only; *(d)* mixed.

(Figure 1*c*) with the most typical example being a mixture of these (Figure 1*d*). Noise and equivocation distract from perfect communication but in different ways. The term *noise* is borrowed from acoustical distortions and is generalized here to cover all kinds of random alterations, blurred images, and uncertainties about how a sent symbol is received. Equivocation shows up in a receiver's simplification of what has been sent or the ambiguity about the sender's intentions. The theory has three ways of expressing the *amount of information transmitted*, $T(A:B)$, through a channel:

$$(1)\ T(A:B) = H(B) - H_A(B)$$
$$(2)\ T(A:B) = H(A) - H_B(A)$$
$$(3)\ T(A:B) = H(A) + H(B) - H(AB)$$

The first expresses communication as the difference between the entropy at the receiver and that part of its entropy that is noise, $H_A(B)$. The second expresses communication as the difference between the entropy at the receiver and that part of its entropy lost as equivocation, $H_B(A)$. Both formally resemble the expression for the semantic information by being the difference between the entropy without and the entropy with reference to a second variable. The third expresses communication as the difference between the entropy that the sender and the receiver would exhibit if they were entirely unrelated and the joint entropy, $H(AB)$, that is in fact observed. It follows that noise and equivocation can be obtained algebraically by $H_A(B) = H(AB) - H(A)$ and $H_B(A) = H(AB) - H(B)$, respectively. Communication is symmetrical, $T(A:B) = T(B:A)$, can be interpreted as *shared variation*, and the quantities involved may be depicted as in Figure 2.

Although communication always involves some kind of covariation, it speaks for the generality of the theory that senders and receivers need not share the same symbol repertoire. Indeed much of communication proceeds by conversions of mental images into verbal assertions, of sound into electrical impulses, of temporal representations into spatial ones, of expressions in one language into those of another, and so forth, during which some patterns are retained.

Regardless of the nature of the media involved, *the amount of communication possible is limited by the number of options available.* More specifically, no channel can transmit more information than its weakest component. For the simple channel between a sender and a receiver $T(A:B)_{max} = \min[H(A),H(B)]$.

Considering that messages can take many material forms and information can be carried by rather different symbols, much of early information theory was concerned with the construction and evaluation of appropriate codes for efficient and/or error-free communication. The coding function may be part of the communicator (e.g., a natural language) or part of the medium (e.g., a microphone or loudspeaker) (Figure 3).

In his fourth theorem Shannon shows that, given enough time, it is always possible to encode a message for transmission even through a very limited channel. However, with C as the channel capacity (in bits per second) and H as the entropy in the source (in bits per symbol) no code can achieve an average rate greater than C/H (symbols per second). In other words, different languages, different signaling alphabets, and different media may make communication more or less efficient, but none can exceed C/H.

Redundancy is another important concept provided by the theory. Redundancy is measured as the difference between the amount that could be and the amount that is in fact transmitted:

$$R = T_{max} - T$$

Redundancy may be caused by duplication of channels of communication, repetition of messages sent, or a priori restrictions on the full range of symbols or symbol combinations used for forming messages (by a GRAMMAR, for example). Although redundancy appears to measure the inefficiency of transmission, in human communication it is a valuable quantity because it can compensate for transmission errors and the effects of selective inattention. For example, the detection of misspellings in a written text, the simplifications used in forming a telegram, and speed reading are all possible only because of redundancy. Shannon estimated that the English language is about 50 percent redundant; subsequent researchers revised his calculation to nearly 70 percent. Shannon's tenth theorem states that the effect of noise in a channel of communication can be compensated for by an amount of redundancy equal to or exceeding the amount of noise in that channel. This redundancy

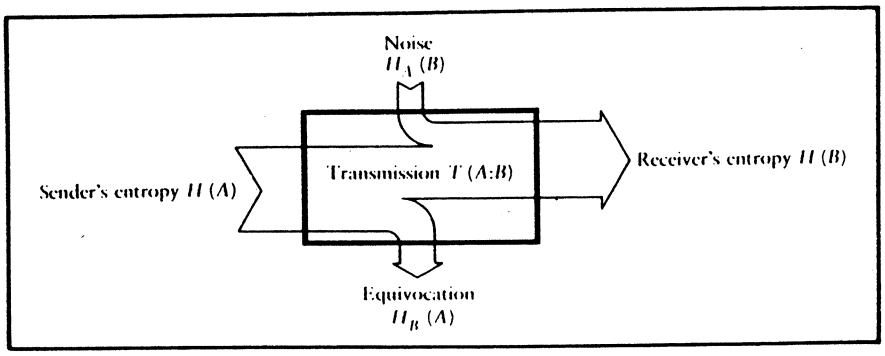

Figure 2. *(Information Theory)* Informational account for simple communication channels.

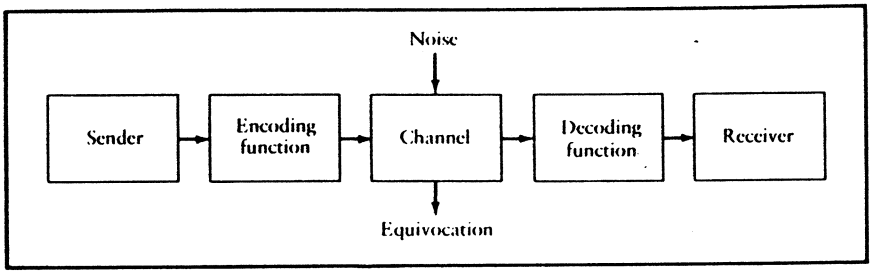

Figure 3. *(Information Theory)* Multicomponent communication process. (After Claude Shannon.)

may stem either from an additional correction channel or from a suitable coding of the messages transmitted.

In complex systems of many variables the total amount of information transmitted within it is

$$T(A:B:...:Z) = H(A) + H(B) + ... + H(Z) - H(AB...Z)$$

To analyze this quantity, various equations are available. For example,

$$T(A:B:...:K:L:...:Z) = T(A:B:...:K) + T(L:M:...:Z) + T(AB...K:LM...Z)$$

decomposes this total into two quantities within and one quantity between the subsystems $AB...K$ and $LM...Z$. Or

$$T(A:B:...:Z) = T(A:B) + T(AB:C) + T(ABC:D) + ... + T(AB...Y:Z)$$

expresses the total as the sum of the amounts transmitted between two variables plus the amount between the two and a third, the amount between the three and a fourth, and so on.

$$H(Z) = T(A:Z) + T_A(B:Z) + T_{AB}(C:Z) + ... + T_{AB...X}(Y:Z) + H_{AB...Y}(Z)$$

explains the entropy in Z in terms of the amount of information transmitted from A plus the amount of information transmitted from B controlled for by A, and so on, plus the unexplainable noise in Z. In this manner complex information flows within a system may be analyzed.

Structural Models

Structural modeling searches for models of qualitative data that represent an optimum balance between structural simplicity and the insignificance of their errors of information omission. Thus models may be found that fit the data best and model the flow of information throughout a system with the least amount of error. Shannon's originally chainlike conception is just one such model.

In the previous examples the total amount of information found in the multivariate data about a system is seen as defined by two kinds of quantities. The sum $H(A) + H(B) + ... + H(Z) = H(m_{ind})$ can be interpreted as the maximum entropy that a model m_{ind} exhibits whose variables $A,B,...,Z$ are statistically independent. The quantity $H(AB...Z) = H(m_o)$ is the entropy actually observed within a model m_o capable of representing all complexities contained in the data. If the two quantities were equal, the data could be said to fit the model of independent variables and show no structure. The total amount, $T(m_{ind}) = H(m_{ind}) - H(m_o)$, can be seen to express the amount of information by which the model m_{ind} is in error. Between the two models, m_o and m_{ind}, on which classical information theory is based, a host of other models could be constructed and tested. Consider four structurally different models within six variables each (Figure 4).

Just as for m_{ind}, each model m_i can be used to generate its own maximum entropy distribution, yielding $T(m_i)$ (for models with loops, as in m_1, this quantity must be obtained by iterative computation,

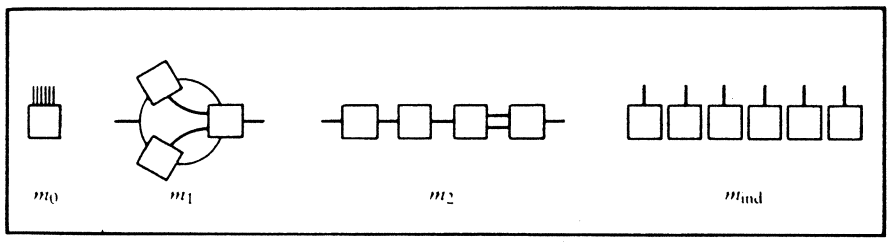

Figure 4. *(Information Theory)* Four examples of structural communication models in block diagrams: *(from left to right)* undifferentiated whole; with circularities; linear chain; with independent components.

whereas for other models algebraic techniques are readily available). In these terms the total amount of information in the data can be decomposed by

$$T(m_{ind}) = T(m_i) + [T(m_{ind}) - T(m_i)]$$

where $T(m_i)$ is the amount of information the model m_i fails to capture, whereas $[T(m_{ind}) - T(m_i)]$ is the amount of information represented by m_i.

Limitations of Information Theory

Some writers have argued that information theory is biased by its early applications in engineering, that it is unable to account for semantic aspects of communication, and that it is limited to linear models (allowing no feedback). None of these arguments is correct.

According to Shannon's second theorem, whose proof is corroborated by many others, the form of the entropy and the information functions is unique, given the axioms of the theory. This puts the theory on a rather unquestionable basis. The critics' burden is to reveal possible inadequacies of the theory by showing the unreasonableness of its axioms, which may be stated here as follows:

(1) $H_2(p,1-p)$
is continuous for $0 \leq p \leq 1$ and $H_2(\frac{1}{2},\frac{1}{2}) = 1$

(2) $H_r(p_1,p_2, \ldots ,p_r)$
is a symmetrical function of its arguments,
$\sum_{a=1}^{a=r} p_a = 1$, and

(3) for any $0 \leq \lambda \leq 1$:
$H_3(\lambda p,(1-\lambda)p,1-p) = H_2(p,1-p) + pH_2(\lambda,1-\lambda)$.

Inapplicabilities of the theory could be encountered, for example, when probabilities do not add to one—a condition that would already fail the first axiom. This condition may arise when the universe of events is undefined, observations in a sample are nonenumerable, or distinctions are fuzzy (do not yield mutually exclusive categories). Information theory presupposes the applicability of the theory of probability (logical possibility, relative frequency, proportion or percent), which is a rather basic demand. The second axiom would become inappropriate, for example, when the ordering of the events 1,2,...,r would make a difference in the amounts the whole set carries. This condition may arise when data are nonqualitative (magnitudinal, for example), in which case the information contained in these proximities is ignored. The third axiom would fail when information quantities are nonadditive and/or probabilities are not multiplicative—for example, when two messages jointly convey more information than the sum of what they convey separately. This situation may arise in irony or when metacommunications and communications are mixed up. The fact that information theory cannot reflect its own context and is, hence, *morpheostatic* in character is common to most social theories and not unique to this one.

Extensions of Information Theory

The basic idea of information theory—equating information with selectivity—may be extended. Effective decisions, one could argue, organize the world, create unusual material arrangements. Messages ranging from blueprints, computer programs, and DNA to political speeches and votes convey information to the extent that they bring about thermodynamically nonentropic pattern, like the assembly of a piece of equipment, a network of computations, the biological structure of an organism, or new forms of social organization. Thus information could be conceived as a measure of the organizational work a message can do, selection being a simple case of this. Information in this sense can be processed (combined, transformed, or encoded in different media) or duplicated at comparatively little cost. Information creates its own context of application. When it organizes an information-processing system it may become amplified, elaborated, and expanded beyond its original scope. Information also becomes part of any living organization, social or biological, that maintains its structure against natural processes of decay or organizational infringements from its environment. Because the thermodynamic laws and the economic costs of production and dissemination apply only to its material carriers, which are largely arbitrary, information is not a commodity. It provides relatively independent accounts for the escalating organizational changes in contemporary society. Information controls a society's rate of thermodynamic decay and directs its economic developments

while escaping many of the traditional socioeconomic constraints.

See also MODELS OF COMMUNICATION.

Bibliography. Klaus Krippendorff, *Information Theory: Structural Models for Qualitative Data*, Beverly Hills, Calif., 1986; Claude E. Shannon and Warren Weaver, *The Mathematical Theory of Communication*, Urbana, Ill., 1949, reprint 1964.

KLAUS KRIPPENDORFF

International Encyclopedia of Communications

ERIK BARNOUW
Editor in Chief

GEORGE GERBNER
Chair, Editorial Board

WILBUR SCHRAMM
Consulting Editor

TOBIA L. WORTH
Editorial Director

LARRY GROSS
Associate Editor

Volume 2

Published jointly with
THE ANNENBERG SCHOOL OF COMMUNICATIONS,
University of Pennsylvania

OXFORD UNIVERSITY PRESS
New York Oxford