

APPENDIX TO CHAPTER 2
AN INFORMATION-THEORETIC MEASURE OF ORDER IN OPERATING
SYSTEMS

A.1 The task of the appendix

An operating system, as characterized previously, is an open system of physical variables interacting through time. What distinguishes operating systems from static physical systems (e.g., stone bridges) and non-physical systems (e.g., systems of geometry) is the temporal change among their interacting features. Put otherwise, an operating system is an arrangement of physical entities in configurations that undergo regular patterns of temporal change.

According to the definition in section 2.5, a system's degree of order (1) varies *directly* with number of interacting features, and (2) varies *inversely* with the degree of independence (randomness) among those features as they change with time. What counts as a relevant feature in a given system will vary from case to case. While all operating systems have parts of one sort or another, what typically changes with time are not the parts but their interrelationships. For present purposes, we focus on the configurations in which parts of the system are arranged rather than on the parts themselves. The interacting features we will be concerned with are the configurations of parts as they undergo temporal change.

A simple example is a children's teeter-totter, consisting of a plank and a fulcrum over which it pivots (the parts). Let us say that these parts are configured in one of three ways during successive stages of normal operation: (a) both ends level with the fulcrum, (b) one end higher, and (c) the other end higher. We think of the apparatus operating in normal fashion when (b) and (c) alternatively are followed by (a) (b,a,c,a,b,a,c,...). Irregularities occur when one end approaches the level of the fulcrum but then falls again

before rising above it (as when one user is significantly heavier than the other).

For a somewhat more complex example, consider an ordinary (nondigital) alarm clock. While the number of parts will vary with design, the basic purpose of the device (let us say) is to show a different configuration of hands on the face for each passing second of a 12-hour period, and then to make a noise when a predetermined configuration is reached. This requires a total of at least $(12 \times 60 \times 60 =)$ 43,200 different configurations through which the system must pass, plus a few more when the alarm is engaged.

According to the preceding definition, both devices exhibit order in their operation to the extent that their successive configurations are interdependent. This amounts to each configuration's yielding to the next in regular sequence. The definition also provides that the clock's degree of order is much higher than that of the teeter-totter. This is because there are many more configurations relevant to the clock's proper operation. The clock's operation becomes disorderly when its configurations fail to succeed one another in proper sequence (as when the second hand starts to oscillate due to the run-down battery).

The present task is to find a measure enabling us to assign numbers to the degrees of order by which such systems might be characterized and hence to compare them on a quantitative basis. This measure should apply to operating systems generally (not just mechanical devices), including those like ecosystems with biological components. Among resources we have available for the task is the mathematical theory of communication.

A.2 Mathematical communication theory

Communication theory began as a mathematical study of the efficient transmission of messages through communication channels. Its first systematic

formulation came in 1948 with Claude Shannon's paper "A Mathematical Theory of Communication."¹ Since then, it has been employed extensively in the design of telephone systems, computer networks, and data links between space vehicles and their control stations. Less extensively, it has also been applied in certain branches of biology, economics, and psychology, and in humane disciplines like philosophy.² Another title for this study is "information theory."

The term 'information' here is used in a precisely defined sense. This sense has nothing to do with gaining knowledge or intelligence, or with ways in which a person might be well informed. It is not information *about* something, or information that might be passed from person to person. In one way or another, information of an ordinary sort always involves *reference* (e.g., of symbols to things), which is a semantic concept entirely foreign to information theory.

Information in the technical sense boils down to changes in an event's probability of occurrence. By way of illustration, consider the flip of an unbiased coin. Before the flip, the probability of a head's coming up is 50%. When a head actually shows up, its probability has increased to 100%. The difference between the probabilities before and after is the information associated with the head's actual occurrence.

For various technical reasons pertaining to the widespread use of digital (binary) encoding devices, information is usually measured in bits (*binary units*). One bit results when the probability of a given event is doubled, as in the illustration of the head's probability of occurrence changing from 50% to 100%. If the antecedent probability of an event is 25% (consider the spin of a 4-sided dreidel), its actual occurrence yields two bits of information, inasmuch as two doublings are required to reach 100% from 25%.

Information can also be thought off as removed uncertainty. The more uncertain an event antecedently, the more uncertainty is removed by its occurrence. Since an event

25% probable is twice as uncertain as one with antecedent probability of 50%, occurrence of the former yields twice the information yielded by occurrence of the latter.

In its most general form, information can also be conceived simply as increased probability. This enables the association of information with events that never actually occur. If the probability of an event E_1 is changed from 25% to 50% by the occurrence of a distinct event E_2 , then one bit of information regarding E_1 is produced by E_2 's occurrence. This effect is the basis of transactions across information channels.

Since the probability of most events prior to occurrence cannot be specified in terms of progressive halvings of 100, a more versatile function is needed for assigning numbers to quantities of information. The function adopted for this purpose by early information theorists is $\log 1/p(e)$ —that is, the logarithm of the reciprocal of the probability of event e conveying the information. Although logarithms to other bases have been used on occasion, use of base 2 went hand-in-hand with choice of the bit as the basic unit of information.

[Technical addendum. Inasmuch as the logarithm of $1/n$ is the negation of the logarithm of n , an equivalent expression is $-\log P(e)$. Given the values -1 and -2 of the logarithms (base 2) of 0.5 and 0.25 respectively, occurrence of an event e conveys 1 bit of information if $P(e)$ is 50% and 2 bits if $P(e)$ is 25%, as before. Similarly, approximately 1.6 bits are conveyed if $P(e)$ is 33% (since $-\log 0.33$ is 1.6), 0.6 bits if $P(e)$ is 67%, and so forth.]

A.3 Information channels

Communication theory is concerned primarily with the transmission of information across information channels. Considered in abstraction from particular physical embodiments (like an old-fashioned telegraph system), an information channel consists of a set of input symbols (e.g., dots, dashes, and spaces), a set of output symbols

(ditto), and for all *pairs* of input and output symbols a set of conditional probabilities giving the likelihood that a specific output symbol will be received when a specific input symbol is transmitted.

In simple and well designed physical communication systems, the probability of a faithful indication at the output of the symbol entered at the input should be fairly high. In a properly functioning telegraph system, for example, the probability of receiving a dash at the output when a dash is entered at the input is near 100%. But many communication systems are less reliable than this; and even telegraph systems are subject to static and other malfunctions. The conditional probabilities characteristic of a given information channel take uncertainties of this sort into account.

[Technical addendum. More exactly stated, an information channel consists of a set of input symbols $A (= \{a_i\}, i = 1, 2, \dots, r)$, a set of output symbols $B (= \{b_j\}, j = 1, 2, \dots, s)$, and a matrix of conditional probabilities $P(b_j/a_i)$ giving the probabilities for all i and j that b_j will be received at the output when a_i is entered at the input. In the case of the very simple information channel diagrammed below:

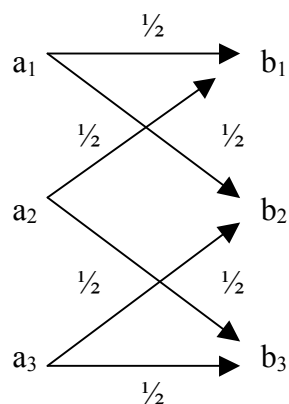


Figure A.1: Simple information channel

$P(b_1/a_1) = 1/2 = P(b_2/a_1) = P(b_1/a_2) = P(b_3/a_2) = P(b_2/a_3) = P(b_3/a_3)$, while $P(b_3/a_1)$, $P(b_2/a_2)$, and $P(b_1/a_3)$ all equal zero. The conditional probability matrix characterizing this channel accordingly is:

	b_1	b_2	b_3
a_1	$1/2$	$1/2$	0
a_2	$1/2$	0	$1/2$
a_3	0	$1/2$	$1/2$

Figure A.2: Simple information channel (probability matrix)]

A.4 The average information of a channel's source

The set of input symbols associated with a given information channel is sometimes known as the channel's *source*. A channel's overall capacity to communicate information depends directly upon the amount of information that can be entered via its source. The amount varies both (1) with the number of distinct symbol-events the source provides, and (2) with the probabilities of occurrence associated with those events.

Regarding (1), it is obvious that more uncertainty is removed by the occurrence of one out of ten equally uncertain events (each 10% probable before occurrence) than by the occurrence of one out of two equally uncertain events (each 50% probable). In general, the more distinct symbol-events it makes available, the more information a source can introduce into its channel.

Factor (2) comes into play when symbol-events at the source are not all equally probable. For purposes of comparison, imagine a source providing three equiprobable symbols-events (each approximately 33% probable). Each of these events provides the same amount of information upon occurrence (about 1.6 bits). Since each occurs equally

often, moreover, the source provides the same amount of information on the average as is provided by any individual occurrence (1.6 bits).

Now imagine a three-symbol source, one symbol-event of which is twice as probable as the other two (one 50% probable, the other two 25% each). Whereas occurrence of the more probable event results in just one bit of information, occurrence of the others results in two bits each. But the average information of the source is not just the average of these three amounts ($1/3^{\text{rd}}$ of $1 + 2 + 2$). The reason is that the event providing the least amount of information occurs more frequently than the others, and hence has proportionately more influence on the source's average information.

In this latter case, the source's average information is the sum of the information provided by the three symbol-events, each weighted by its individual probability of occurrence (1.5 bits, compared with about 1.6 when the three events are equally probable). Comparable provisions apply whenever the symbol-events of a source diverge from equiprobability.

In general, the closer symbol-events at the source approach equiprobability, the more information the source makes available for transmission through the channel. This quantity will be referred to as the *average information* of the source.³

[Technical addendum. The formal definition of *average information* is:

$$H(A) = \sum_A P(a^i) \log 1/P(a^i)$$

where 'A' indicates the summation of the quantity following it for all values a_i within A.

What this equation says, in effect, is that the average amount of information issuing from source A equals the amount of information represented by the occurrence of a symbol-event a_i multiplied by its probability of occurrence, summated over the entire membership of A.

When there are just two symbol-events, a_1 and a_2 , available at a source with probabilities $2/3$ and $1/3$ respectively, individual occurrences of the two convey approximately 0.6 ($= -\log 2/3$) and 1.6 ($= -\log 1/3$) bits of information respectively. In this case, $H(A) = (2/3)(0.6) + (1/3)(1.6) =$ (approximately) 0.9 bits of information.

Similar calculations show that when $P(a_1) = 3P(a_2)$ in a two-member source, then $H(A) = 0.8$; when $P(a_1) = 4P(a_2)$, then $H(A) = 0.7$; and so forth. When a_1 and a_2 are equally likely to occur, however, $H(A) = 1.0$. This illustrates the principle that the average information of an information source is maximal when its symbol-events are equally probable, and also shows that one bit is the maximum amount of information that can be introduced into a channel by a two-member source.]

A.5 Channel equivocation

The general purpose of an information channel is to make information available at its output pertaining to the occurrence of particular input-events. A channel may be more or less reliable in that regard. A given channel is 100% reliable if every event at its output gives an accurate and unambiguous indication of the corresponding event at its input.

If the identity of events at its input is sometimes left ambiguous by symbol-occurrences at its output, however, the channel is said to be characterized by a certain amount of equivocation. A channel without equivocation is sometimes called a noise-free channel, suggesting that no “noise” occurs to diminish the channel’s reliability as a transmitter of information.

[Technical addendum. Channel equivocation is the average amount of uncertainty at the output (B) regarding the identity of symbol-events at the input (A) after the reception of corresponding symbol-events at the output. Overall channel equivocation ($H(A/B)$) is a composite quantity resulting from (but not the same as) the equivocations associated with each output-event b_j ($H(A/b_j)$). This latter is determined according to the formula:

$$H(A/b_j) = \sum_A P(a_i/b_j) \log 1/P(a_i/b_j)$$

for all members a_i of the input set A .

It will be noted that the conditional probabilities in this formula are of input events (a_i) given output events (b_j)—thus, $P(a_i/b_j)$. This is opposite the sense of conditionalities ($P(b_j/a_i)$) defining a channel's probability characteristics (see Figure 3.2 for example). Probabilities $P(a_i/b_j)$ can be obtained from probabilities $P(b_j/a_i)$ according to what is known as Bayes's law:

$$P(a_i/b_j) = P(b_j/a_i)P(a_i)/P(b_j)$$

where $P(b_j)$ (the probability of b_j at the output) is the summation for all cases a_i of $P(b_j/a_i)P(a_i)$.

By definition, in a channel without equivocation there will be a unique input-event a_i associated with any given output-event b_j , such that $P(a_i/b_j)$ is unity and $-\log P(a_i/b_j)$ is zero. When these values are summated for all members of A , of course, $H(A/b_j)$ also turns out to be zero. For this quantity to equal zero means that when b_j occurs at the channel's output, it does so without equivocation.

By way of contrast, we may observe that there are two events (a_1 and a_2) at the input of the channel in Figure A.1 that lead to b_1 at the output. For simplicity, assume that the three input-events in this channel are equiprobable. It follows from the conditional probabilities of the channel matrix (Figure 3.2) and Bayes's Law that both $P(a_1/b_1)$ and $P(a_2/b_1)$ equal one-half. Since $\log (1/2) (= -\log 1/2) (= \log 2) = 1$, both $P(a_1/b_1) \log 1/P(a_1/b_1)$ and $P(a_2/b_1) \log 1/P(a_2/b_1)$ equal one-half. These two pairings together thus contribute ($1/2 + 1/2 =$) one bit of equivocation. Inasmuch as $P(a_3/b_1)$ is zero, no more equivocation comes from this source. $H(A/b_1)$ for this channel, accordingly, is one bit of equivocation.

The equivocation of an informational channel overall ($H(A/B)$) is the summation with respect to B of all values of $H(A/b_j)$ each multiplied by the probability of occurrence of b_j as follows:

$$H(A/B) = \sum_B P(b_j) \sum_A P(a_i/b_j) \log 1/P(a_i/b_j)$$

For any a_i and b_j for which $P(a_i/b_j)$ is unity—that is, for which b_j is an entirely reliable indicator of a_i —the quantity $\log 1/P(a_i/b_j)$ is zero, and the probability $P(b_j)$ is not a factor in the summation. For any b_j such that there is an a_i for which $P(a_i/b_j)$ is less than unity, on the other hand, $P(a_i/b_j) \log 1/P(a_i/b_j)$ will have a positive value, and its contribution to $H(A/B)$ will be a factor of $P(b_j)$ as well. When and only when $P(A/b_j)$ is zero for all members of B will the overall equivocation $H(A/B)$ of the channel itself equal zero.]

A.6 Mutual information

An information channel with low equivocation is characterized by a representation at the output of approximately the same events as those presented at the input. Conversely, the higher a channel's equivocation, the less reliable are events at its output as representations of input occurrences. The capacity of a channel to convey information thus varies inversely with the equivocation of its input with respect to its output.

As already noted in section A.4, a channel's capacity to communicate information also varies directly with the average amount of information that can be entered at its input. The difference between the average information of its input and the equivocation of its input with respect to its output thus measures a channel's capacity as a reliable conveyor of information. This quantity is referred to as the *mutual information* of the channel.

[Technical addendum. Given the preceding definitions of $H(A)$ and $H(A/B)$, the mutual information ($I(A;B)$) of an information channel A-B is defined as follows:

$$I(A;B) = H(A) - H(A/B)$$

The term ‘mutual information’ reflects the fact that $I(A;B)$ is the amount of information shared at output B with input A.]

A.7 Markov sources

A *Markov source* can be defined most directly in contrast with a zero-memory source. A zero-memory information source is one such that the symbol-occurrences coming from it are statistically independent. What this means is that the identity of an event emitted by the source has no bearing on the identities of events either preceding or following it.

A Markov source, by contrast, is one in which the symbol-occurrences emanating from it are statistically *interdependent*. In a first-order Markov source, the probability of a given event’s occurring at a given place in the sequence depends to some extent upon the identity of the immediately preceding event. When a letter Q appears in an English text, for example, it is highly probable that the next letter will be a U. In a second-order Markov source, a given probability of occurrence depends upon the identities of the two preceding events (as QU in English almost always is followed by a vowel). And so it goes for Markov sources of higher orders.

This concept can be adapted to our purposes by thinking of the statistical relations among successive events issuing from a Markov source as equivalent to statistical relations between input and output events in an information channel. What this amounts

to is easily illustrated by a specific sequence of symbol-occurrences—say, $s_1, s_2, s_3, s_4, \dots, s_n$ —issuing from a first-order Markov source. Given any pair of events in sequence (e.g., s_1 and s_2), the first (s_1) is conceived as occurring at the input of an information channel and the second (s_2) at the channel's output; and so on for other pairs in the sequence (s_2 and s_3 , then s_3 and s_4 , etc.).

The result is to convert the statistical interdependencies characteristic of a Markov source into statistical interdependencies characteristic of an information channel. As seen in the section following, this will enable us to conceive of successive stages of an operating system as constituting an information channel. We shall also see how the mutual information of that channel can serve as a measure of the corresponding system's degree of order.

[Technical addendum. Consider a sequence of event-occurrence e_n, e_{n+1}, e_{n+2} , etc. Now consider a series of pairs taken from this sequence, so ordered that the second member of a given pair (e.g., the pair e_n and e_{n+1}) becomes the first member of pair immediately following, e.g., e_{n+1} and e_{n+2}). In a first-order Markov source yielding the sequence of event-occurrences indicated above, the antecedent likelihood of the pair e_{n+1}, e_{n+2} following the pair e_n, e_{n+1} is $P(e_{n+2}/e_{n+1})$. In a second-order Markov source, this antecedent likelihood is $P(e_{n+2}/e_n, e_{n+1})$. And so on for Markov sources of higher order.

The conditional-probability matrix characteristic of a given information channel can be converted into a matrix characteristic of its Markov-source equivalent by equating output events of the former with appropriate input events. The matrix of Figure A.2, for example, is converted into the matrix of an equivalent first-order Markov source by relabelling the columns to match the rows, as follows:

	a ₁	a ₂	a ₃
a ₁	1/2	1/2	0
a ₂	1/2	0	1/2
a ₃	0	1/2	1/2

Figure A.3: First-order Markov source

Another way of displaying these relationships is illustrated by the state-diagram:

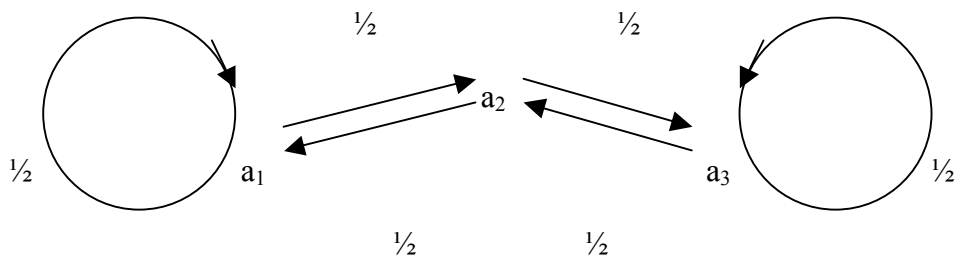


Figure A.4: First-order Markov source (flow diagram)

Figures A.3 and A.4 represent a first-order Markov source inasmuch as the probability of occurrence of a given event (heading the columns of Figure A.3) is shown conditional upon the occurrence of the immediately preceding event only (indicated at the left of Figure A.3).

A state-diagram representing a second-order Markov source (with only two members, for simplicity) is shown below

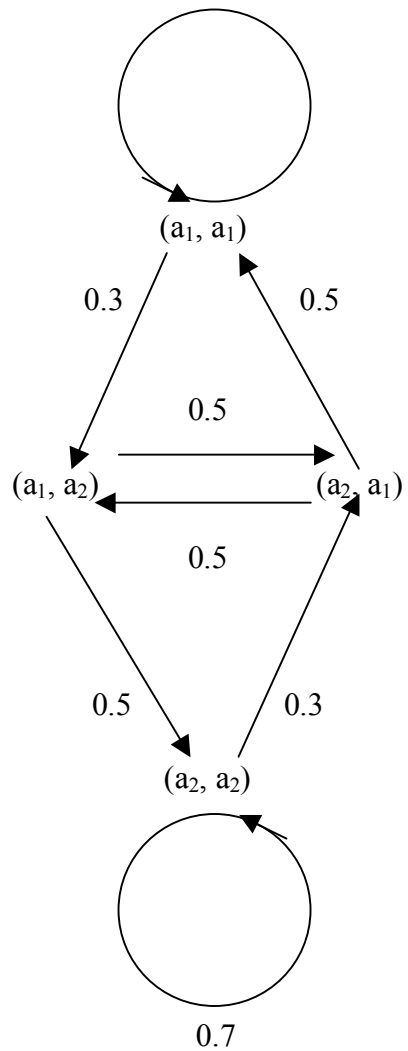


Figure A.5: Second-order Markov source

In this example, $P(a_1/a_1, a_1) = P(a_2/a_2, a_2) = 0.7$, $P(a_2/a_1, a_1) = P(a_1/a_2, a_2) = 0.3$, and $P(a_1/a_1, a_2) = P(a_1/a_2, a_1) = P(a_2/a_1, a_1) = P(a_2/a_2, a_1) = 0.5$. In interpreting this figure, it should be noted that if a_1 follows either a_1, a_2 or a_2, a_2 , then the next pair in sequence will be a_2, a_1 . The designation ‘ a_1, a_2 ’, of course, indicates two events in that particular order; and so on for the other pairings.]

A.8 Operating systems as Markov sources

The definition of orderliness in section 2.5 specifies that an operating system's degree of order (1) varies directly with number of interacting features, and (2) varies inversely with the degree of independence among those features. Working equivalents of (1) and (2) can be expressed in terms of (i) average information and (ii) equivocation respectively, leading to a definition of an operating system's degree of order in information theoretic terms. Let us consider how.

The average information of an information source is the summation of the amounts of information represented by its individual members, multiplied in each case by probability of occurrence (section A.4). This quantity generally increases with number of members. When the source is a set of all configurations exhibited by an operating system during normal operation, these configurations count as members of the source. Accordingly, the average information of such a source generally increases with number of configurations.

[Technical addendum. In the case of sources with equiprobable members, their average information increases with number of members in an entirely regular manner.

Qualification is required for sources whose members are not all equiprobable.

Probabilities of occurrence of an n -membered source might be distributed among members in such a fashion that it has lower average information than a source with $n-1$ members. In a six-member source, for example, five might occur 10% of the time each and the other 50% of the time. The source's average information in this case is 2.16 bits, less than that of a five-membered source with equiprobable members (2.32 bits).

Departures of this sort from a regular progression are unlikely to extend more than a step or two in the sequence. Thus a seven-membered source with one probability of 40% and the other six 10% has an average information of 2.52 bits, more than the 2.32 bits of an equiprobable five-membered source. Taking irregularities of this sort into account, it

remains accurate to say that average information of information sources generally increases with number of members.]

With regard to condition (1) in the definition of orderliness, we thus may say that an operating system's degree of order varies directly with the number of configurations it exhibits. This is to say that its degree of order varies directly with its average information as an information source.

We turn now to the second condition. An information channel A-B is reliable to the extent that events at output B provide unambiguous indications of corresponding events at input A. Ambiguity in that regard is referred to as equivocation. Channel equivocation is defined as the average amount of uncertainty left by the occurrence of events at B regarding the identity of corresponding events at A (section A.5).

To treat a first-order Markov source as an information channel is to treat each successive pair of events issuing from the source (events $s_1, s_2, s_3, s_4, \dots s_n$) as containing both an input and an output event (section A.7). In pair (s_1, s_2) , for instance, s_1 is input and s_2 output; in pair (s_2, s_3) , s_2 is input and s_3 output; and so forth. With higher order Markov sources, more preceding events are taken into account. In a second-order source, for example, the successive occurrences s_1, s_2, s_3 would contain s_3 as output and the pair s_1, s_2 as input. In any such case, equivocation of a Markov source is the average uncertainty regarding the identity of relevant antecedents after a given event has occurred.

When an operating system is treated as a Markov source, the successive configurations it exhibits through time are treated as a succession of events issuing from the source. The equivocation of an operating system so conceived is the average uncertainty left after the occurrence of a given configuration regarding the identity of

relevant configurations preceding it. This uncertainty increases with increasing independence among configurations involved, and vice versa.

With regard to condition (2) of the previous definition of orderliness, accordingly, we may say that an operating system's degree of order varies inversely with its equivocation. The more independence on the average between successive configurations, the lower its degree of order; and vice versa. Conceived in this fashion, an operating system's equivocation measures the extent to which its successive configurations are independent.

As a final step, let us recall that a channel's mutual information is defined as the difference between the average information of its input and its equivocation (section A.6). This definition carries over to Markov sources generally and hence to operating systems considered as Markov sources. Inasmuch as an operating system's degree of order (1) varies directly with its average information as a Markov source and (2) varies inversely with its equivocation, the system's degree of order varies directly with its mutual information. In brief, the mutual information of an operating system treated as a Markov source serves as a measure of its degree of order.

A.9 Examples

In its normal operation, the children's teeter-totter considered in section A.1 functions as a Markov source issuing a series of repeating configurations (b, a, c, a, b, a, c, ...). In this series, the probability of a's occurrence at a given point is 50%, and those of b's occurrence and of c's occurrence 25% each. Operating normally (i.e., with complete regularity), this system loses no information to equivocation. Given these statistics, its mutual information is $(1.5 - 0 =) 1.5$ bits of information.

The other example considered in section A.1 is that of an alarm clock with both minute and second hands. In normal operation, the clock issues at least $(12 \times 60 \times 60 =)$

43,200 distinct configurations (with a few more when the alarm sounds). Its average information as a Markov source thus is between 15 and 16 bits of information ($2^{15} = 32,768$, $2^{16} = 65,536$; see section A.2). Since no equivocation is present during normal operation, this is also the quantity of its mutual information.

These examples have been chosen for the ease with which their mutual information can be calculated. It would be considerably more complicated to determine the mutual information characterizing an automobile engine, and even more so in the case of a biological system. Our present purpose does not require actual calculations in such complex cases. The purpose of the Appendix is merely to show how, in principle, such calculations might be made. To show this is to show that the orderliness of biological systems, in principle, admits quantitative measurement.

Notes

1. Important preliminary work had been done by H. Nyquist and R.V.L. Hartely in the 1920s. Relevant papers of Nyquist, Hartely, and Shannon were all published in the *Bell System Technical Journal*. “The Mathematical Theory of Communication” was published in the July and October issues of 1948. Various reprints have since become available.
2. A brief survey of philosophic applications can be found in the present author’s “Information Theory,” *Routledge Encyclopedia of Philosophy*, 1998, vol. 4, pp. 782-286.
3. It is customary to refer to this quantity as the *entropy* of the source. This terminology will not be used here to avoid taking a gratuitous stand on the relation between this quantity and the entropy of thermodynamics, on which experts disagree. A discussion of the relation between entropy in communication

theory and that in thermodynamics may be found in the author's Cybernetics and the Philosophy of Mind, Ch. III, sect. 3.