Semantic Information

Yehoshua Bar-Hillel; Rudolf Carnap

The British Journal for the Philosophy of Science, Vol. 4, No. 14. (Aug., 1953), pp. 147-157.

Stable URL:
http://links.jstor.org/sici?sici=0007-0882%28195308%294%3A14%3C147%3ASI%3E2.0.CO%3B2-7

The British Journal for the Philosophy of Science is currently published by Oxford University Press.

SEMANTIC INFORMATION *

YEHOSHUA BAR-HILLEL and RUDOLF CARNAP

I

THE Mathematical Theory of Communication, often referred to also as Theory (of Transmission) of Information, as practised nowadays, is not interested in the content of the symbols whose information it measures. The measures, as defined, for instance, by Shannon,[1] have nothing to do with what these symbols symbolise, but only with the frequency of their occurrence.[2] The probabilities which occur in the definientia of the definitions of the various concepts in Communication Theory are just these frequencies, absolute or relative, sometimes perhaps estimates of these frequencies.

This deliberate restriction of the scope of Statistical Communication Theory was of great heuristic value and enabled this theory to reach important results in a short time. Unfortunately, however, it often turned out that impatient scientists in various fields applied the terminology and the theorems of Communication Theory to fields in which the term 'information' was used, presystematically, in a semantic sense, that is, one involving contents or designata of symbols,

* Received 23. x. 52. This is a slightly revised version of a paper read before the Symposium on Applications of Communication Theory, London, 26th September 1952. A more detailed and systematic treatment of the same topic appears simultaneously in *Technical Report* No. 247 of the Research Laboratory of Electronics, Massachusetts Institute of Technology, 1953.

[1] See, e.g., C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, 1949. E. Colin Cherry's 'A History of the Theory of Information', *Proceedings of the Institution of Electrical Engineers*, 1951, **98**, part iii, pp. 383-393, gives an excellent account of the development of this theory and contains also an extensive bibliography up to 1950.

[2] A notable exception to this general trend of communication engineers is D. M. MacKay, who recognised as early as 1948 that the concept of information treated in Communication Theory, which he proposes now to call 'selective information' should be supplemented by a concept of 'scientific information'. It seems, however, that this concept does not coincide with what we call here 'semantic information'. A clarification of the exact relationship will be undertaken elsewhere.

The clearest presentation of MacKay's ideas may be found in his contribution to the *Transactions of the Eighth Conference on Cybernetics*, 'In Search of Basic Symbols', New York, 1952, pp. 181-235, where an earlier paper of his, 'The Nomenclature of Information Theory', is also reprinted.

or even in a pragmatic sense, that is, one involving the users of these symbols. There can be no doubt that the clarification of these concepts of information is a very important task. However, the definitions of information and amount of information given in present Communication Theory do not constitute a solution of this task. To transfer these definitions to the fields in which those semantic or pragmatic concepts are used, may at best have some heuristic stimulating value but at worst be absolutely misleading.

In the following, the outlines of a Theory of Semantic Information will be presented. The contents of the symbols will be decisively involved in the definition of the basic concepts of this theory and an application of these concepts and of the theorems concerning them to fields involving semantics thereby warranted. But precaution will still have to be taken not to apply prematurely these concepts and theorems to fields like psychology and other social sciences, in which users of symbols play an essential role. It is expected, however, that the semantic concept of information will serve as a better approximation for some future explanation of a psychological concept of information than the concept dealt with in Communication Theory.

2

The fundamental concepts of the theory of semantic information can be defined in a straightforward way on the basis of the theory of inductive probability that has been recently developed by one of us.[1] Unfortunately, the space at our disposal does not permit us to develop the full terminological background on which our presentation is based. We have to refer once and for all to the extensive presentation of this background given in [Prob.] or to the more concise one given in [Cont.]. (In the Appendix, an even more concise summary is offered for the convenience of the reader.) Let us state only that what follows refer to a fixed language system L_n^π, by which we mean, approximately, an applied first-order functional semantical system with n individuals, say a_1, a_2 . . . , a_n, and π primitive properties, say P_1, P_2, . . . , P_π. A disjunction which, for each of the πn atomic statements, contains either this statement or its negation (but not both) as a component, will be called a *content-element*. The content-elements are the weakest factual statements of L_n^π inasmuch as the

[1] R. Carnap, *Logical Foundations of Probability*, University of Chicago Press, 1950, cited hereafter as [Prob.] and *The Continuum of Inductive Methods*, University of Chicago Press, 1952, cited hereafter as [Cont.].

only factual statement L-implied by a content-element is this content-element itself. One of the 64 content-elements in L_3^2, for instance, is

$$P_1a_1v \sim P_2a_1v \sim P_1a_2vP_2a_2vP_1a_3vP_2a_3.$$

The class of all content-elements L-implied by any statement i (in L_n^π) is called the *content* of this statement and denoted by ' Cont (i) '. It can be easily verified that the content of any atomic statement contains exactly half of all content-elements, that of an L-true statement none, and that of an L-false statement all of them. The last property may look slightly artificial but is no more so than the use of, say, the null-set in set-theory.

We offer Cont (i) as an explicatum for the ordinary concept ' the information conveyed by the statement i ', taken in its semantic sense. We have no time to show at length that Cont (i) is an adequate explicatum. But it can be immediately verified that it fulfils at least the condition that Cont (i) includes Cont (j) if i L-implies j. This condition should be regarded as a necessary, though certainly not sufficient, condition of adequacy of any proposed explanation of the mentioned concept.

Since Cont (i) is equal to the class of the negations of the state-descriptions contained in the range of $\sim$ i, the properties of Cont (i) can be easily derived from the properties of the concept ' range of i ' which has been treated at length in [Prob.]. We shall say nothing more here.

It is often important not only to know what is the information conveyed by some statement but also to attach a measure to this information. We need not start afresh looking for appropriate measure-functions ranging over contents since measure-functions over ranges have been extensively discussed in [Prob.]. For each of the latter m-functions, as they are called in [Prob.], a corresponding content-measure-function is defined simply by

$$\text{cont (i)} = \text{m} (\sim \text{i}).$$

cont (i) (read : the content-measure of i) is offered as one (not *the*) explicatum of the ordinary concept ' amount of information conveyed by i ' in its semantic sense. Among the most important properties of cont (i), immediately derivable from the corresponding properties of m(i) treated in [Prob.] we have

$$0 \leqq \text{cont (i)} \leqq 1,$$

where the extremes are reserved for L-true and L-false statements, respectively, and

$$\text{cont}\ (i\,.\,j) = \text{cont}\ (i) + \text{cont}\ (j) - \text{cont}\ (i \vee j).$$

From the last theorem follow immediately :

$$\text{cont}\ (i\,.\,j) \leqq \text{cont}\ (i) + \text{cont}\ (j),$$

and the interesting additivity theorem :

cont (i . j) = cont (i) + cont (j) if and only if i and j are L-disjunct.

Here, however, an inconsistency in the intuitions of many of us becomes apparent. Though it is indeed, after some reflection, quite plausible that the content of a conjunction should be equal to the sum of the contents of its components if and only if these components are L-disjunct or content-exclusive, in other words, if they have no factual consequences in common, it is also plausible, without much reflection, that the content of the conjunction of two basic statements, say 'P_1a_1' and '$\sim P_2a_3$' should be equal to the sum of the contents of these statements since they are independent, and this not only in the weak deductive sense of this term, but even in the much stronger sense of initial irrelevance. But no two basic statements with different predicates are L-disjunct, since they have their disjunction, which is a factual statement, as a common consequence. Our intuitions here, as in so many other cases, are in conflict and it is best to solve this conflict by assuming that there is not *one* explicandum 'amount of information' but at least two, for one of which cont is a suitable explicatum, whereas the explicatum for the other has still to be found.

So far we have dealt with the information conveyed by some statement separately. At times, however, we are as much, or even more, interested in the information conveyed by a statement j in excess to that conveyed by some other statement i or a class of statements. We therefore define the concepts 'content of j relative to i' and 'content-measure of j relative to i' by

$$\text{Cont}\ (j/i) = \text{Cont}\ (i\,.\,j) - \text{Cont}\ (i)$$

and

$$\text{cont}\ (j/i) = \text{cont}\ (i\,.\,j) - \text{cont}\ (i)$$

respectively. (Notice that the '—' in the first of these definitions is the symbol of class-difference, in the second that of ordinary numerical difference.) The maximum value of cont (j/i) is obviously cont (j) and this value is obtained if and only if i and j are L-disjunct. The minimum value of cont (j/i) is 0 and this value is obtained if, and only if, i L-implies j. Of special interest is that

$$\text{cont }(j/t) = \text{cont }(j),$$

where t stands for any L-true statement (any 'tautology'), since this allows us to define cont (j) in terms of cont (j/i), thereby reversing the definition procedure followed by us before. Even more interesting is that

$$\text{cont }(j/i) = \text{cont }(i \supset j),$$

from which it follows that, given i, j conveys no more additional information than $i \supset j$, by itself a much weaker statement.

If we stipulate now, for the second explicatum of 'amounts of information', that all basic statements shall convey the same amount of information, and this independently of whether these statements appear alone or as components in some non-contradictory conjunction, and if we stipulate, in addition, for the purpose of normalisation, that the amount of information conveyed by a basic statement shall be 1, it can easily be seen, along well-known lines of computation, that these stipulations are fulfilled if we define this second function, to be called 'measure of information' and denoted by 'inf', as

$$\text{inf}(i) = \text{Log}\,\frac{1}{1\text{-cont }(i)}$$

(where 'Log' stands for $\log_2$), from which we obtain, by simple substitution,

$$\text{inf}(i) = \text{Log}\,\frac{1}{m\ (i)} = -\ \text{Log } m\ (i).$$

The last equation is analogous to a definition of amount of information in Communication Theory but with inductive probability instead of statistical probability.

Among the various theorems regarding inf let us mention

$$0 \leqq \text{inf}(i) \leqq$$

and the theorem of additivity, which, however, involves now a quite different condition.

$\text{inf}(i\,.\,j) = \text{inf}(i) + \text{inf}(j)$ if, and only if, i is initially irrelevant to j (with respect to that m-function on which inf is based).

Sometimes $\text{inf}(i\,.\,j)$ is greater than $\text{inf}(i) + \text{inf}(j)$. If anyone should find this strange that might be due to the fact that he has subconsciously switched to some other explicatum, such as cont, for which this can indeed not happen.

Another theorem of great importance deals with inf(j/i). It states that

$$\mathrm{inf}(j/i) = \mathrm{Log}\,\frac{1}{c(j,i)} = -\,\mathrm{Log}\,c(j,i)$$

where c(j , i) is the degree of confirmation of (the hypothesis) j on (the evidence) i, defined in [Prob.] as $\frac{m(i\,.\,j)}{m(i)}$.

The statistical correlate of inf has found a large field of application in communication engineering. Neither cont nor its statistical correlate have found any useful application so far. It is, however, to be expected that the facet of the amount of information which is measured by cont, should find its fields of application too, especially so since cont is a mathematically simpler function of m than inf.[1]

3

Among the various m-functions on which cont and inf may be based, there are two groups of special importance. The first group consists of just one member, to be designated here by 'm_D'—its symbol in [Prob.] is '$m^\dagger$' — : the second group has infinitely many members denoted collectively by 'm_I'. m_D assigns to each content-element the same value. This makes the computations with this function especially easy, in general, and the preference given to it understandable. It suffers, however, from the great disadvantage that it does not allow us, roughly speaking, to learn from experience. 'P_1a_4', for instance, will have a c_D-value of $\frac{1}{2}$, on no evidence at all or, in other words, on the tautological evidence, and the same c_D-value on the evidence '$P_1a_1 \,.\, P_1a_2 \,.\, P_1a_3$'. In spite of this defect, there are situations in which m_D , c_D, and the information-functions based upon them may be of importance. Situations in which we intend to use only deductive reasoning are of this type, hence the subscript 'D' for 'deductive'.

In those situations in which inductive reasoning is to be applied, only such m-functions may be used which allow us to learn from experience, in other words, which fulfil the Requirement of Instantial

[1] Indeed, in a paper by John G. Kemeny and Paul Oppenheim, 'Degree of Factual Support', *Philosophy of Science*, 1952, **19**, 288-306 (published after the present paper was read in London), a concept of Strength of a statement was used whose definition corresponds closely to that of our cont_D.

Relevance.[1] These are the functions m_I; the 'I' stands for 'inductive'.

All the theorems that hold for cont and inf in general hold, of course, also for $cont_D$ and inf_D and all the $cont_I$ and inf_I functions. For these more specific functions, however, additional theorems can be proven. For lack of space, this will not be done here. Let us only remark that certain inconsistencies in our intuitive requirements with regard to information-functions may be due to a subconscious switching from D-type functions to I-type functions and vice versa.

4

Situations often arise in which we do not know whether a certain event has occurred or will occur, but only that exactly one event out of a class of mutually exclusive events has occurred or will occur. The statements describing these events convey each a certain amount of information on the available evidence. It makes therefore good sense to ask for some average of the amount of information conveyed by these statements. If these statements refer to future events, one talks about the amount of information that may be expected to be conveyed, on the average. In [Prob.] it is shown that in many similar situations the c-mean estimate of the function in question will be a satisfactory measure for this expected value. Confining ourselves here, for the sake of easy comparability with prevailing Communication Theory, to inf and using 'exhaustive system' to denote a class of statements of the above-mentioned character, we define *the (c-mean) estimate of the measure of information conveyed by (the members of the exhaustive system) H on (the evidence) e*, in symbols : est (inf, H, e), as follows :

$$\text{est}(\text{inf}, H, e) = \sum_{p=1}^{n} c(h_p, e) \times \text{inf}(h_p/e).$$

From this definition and from a prior theorem on $\text{inf}(h_p/e)$ the theorem

$$\text{est}(\text{inf}, H, e) = -\sum_{p} c(h_p, e) \times \text{Log}\, c(h_p, e)$$

immediately follows. The communicational correlate of this theorem is well known. We see no reason, so far, to attach any special significance to the formal similarity of its right side to certain entropy-type expressions in statistical thermodynamics.

[1] See R. Carnap, 'On the Comparative Concept of Confirmation', this *Journal*, 1953, 3, p. 314

To give a simple illustration : If on the basis of available evidence, mainly prior observations, the c-value of the hypothesis, h_1, 'There will be warm weather in London on the 23rd of September 1953',[1] is $\frac{1}{2}$, the c-value of h_2, 'There will be temperate weather . . .', is $\frac{1}{4}$, and the c-value of h_3, 'There will be cold weather . . .', is $\frac{1}{4}$, then

$$\begin{aligned} \mathrm{est}(\mathrm{inf}, H, e) &= -\sum_p c(h_p, e) \times \mathrm{Log}\, c(h_p, e) \\ &= \tfrac{1}{2} \times 1 + \tfrac{1}{4} \times 2 + \tfrac{1}{4} \times 2 = 1{\cdot}5 \end{aligned}$$

(where $H = \{h_1, h_2, h_3\}$).

If $H = \{h_1 \ldots h_n\}$ and $K = \{k_1 \ldots k_m\}$ are (deductively) independent exhaustive systems (on e) (i.e. no $h_p . k_q$ is L-false (on e)), then $H . K$, defined as $\{h_1 . k_1, h_1 . k_2 \ldots h_1 . k_m, h_2 . k_1 \ldots h_n . k_m\}$ is exhaustive too (on e), hence

$$\mathrm{est}(\mathrm{inf}, H . K, e) = \sum_{p=1}^{n} \sum_{q=1}^{m} c(h_p . k_q, e) \times \mathrm{inf}(h_p . k_q/e).$$

We have

$$\mathrm{est}(\mathrm{inf}, H . K, e) \leqq \mathrm{est}(\mathrm{inf}, H, e) + \mathrm{est}(\mathrm{inf}, K, e),$$

with equality holding if, and only if, the h's and the k's are mutually irrelevant.

If the statement k is added to our evidence, *the posterior estimate of the measure of information conveyed by H on e and k* will, in general, be different from *the prior estimate of the measure of information conveyed by H on e alone*. This difference is often of great importance and will therefore receive a special name, *amount of specification of H through k on e* and be denoted by 'sp (inf, H, k, e)'. The formal definition is

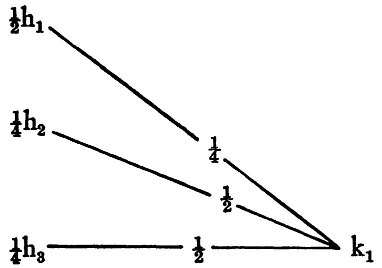
$$\mathrm{sp}(\mathrm{inf}, H, k, e) = \mathrm{est}(\mathrm{inf}, H, e) - \mathrm{est}(\mathrm{inf}, H, e . k).$$

It is easy to see that $\mathrm{sp}(\mathrm{inf}, H, k, e) = 0$ if (but not only if) k is irrelevant to the h's on e. sp may have positive and negative values with its maximum obviously equal to the prior estimate itself. This value will be obtained when $e . k$ L-implies some h_p. In this case, H is completely specified through k (on e).

Let, to continue our previous illustration, k_1 be a certain report of weather-instrument-readings. Let $c(k_1, e . h_1) = \frac{1}{4}$, $c(k_1, e . h_2) = c(k_1, e . h_3) = \frac{1}{2}$. The following diagram will help to visualise the situation :

[1] The formulation of this statement exceeds already the potentialities of the language-systems envisaged here. However, this is of no importance in this connection.

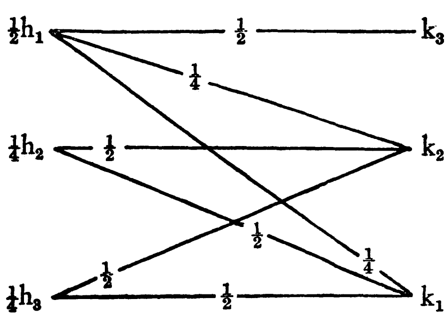
It is easy to compute, or to read from the diagram, that $c(h_1, e.k_1) = c(h_2, e.k_1) = c(h_3, e.k_1) = \frac{1}{3}$. Hence $est(inf, H, e.k_1) = Log\ 3 = 1{\cdot}585$ and $sp(inf, H, k_1, e) = -0{\cdot}085$.



Situations often arise in which the event stated in k has not yet occurred, or, at least, in which it is not known whether it has occurred but in which it is known that either it or some other event belonging to a certain class of events will occur or has occurred. In such circumstances, it makes sense to ask for some average of the posterior estimate of the measure of information conveyed by H on e and (some member of) K (the exhaustive system of the k's). We are led to *the (c-mean) estimate of this posterior estimate* denoted by '$est(inf, H/K, e)$' and defined as

$$est(inf, H/K, e) = \sum_{q=1}^{m} c(k_q, e) \times est(inf, H, e.k_q).$$

Let us complete our illustration in the following diagram :



Then $est(inf, H/K, e) = \sum_q c(k_q, e) \times est(inf, H, e.k_q)$

$$= \tfrac{3}{8} \times Log\ 3 + \tfrac{3}{8} \times Log\ 3 = 1{\cdot}189.$$

The estimate of the amount of specification of H through K on e is, of course, equal to the difference between the prior estimate of the measure of information conveyed by H on e and the estimate of the posterior estimate of the measure of information conveyed by H on e and K, in symbols :

$$\text{sp}(\text{inf}, H, K, e) = \text{est}(\text{inf}, H, e) - \text{est}(\text{inf}, H/K, e).$$

In our example, sp(inf, H, K, e) = 1·5 — 1·189 = 0·311.

Let us mention only three theorems in this connection, the communicational correlates of which are well-known :

$$\text{est}(\text{inf}, H/K, e) = \text{est}(\text{inf}, H\,.\,K, e) - \text{est}(\text{inf}, K, e),$$
$$\text{sp}(\text{inf}, H, K, e) = \text{sp}(\text{inf}, K, H, e),$$
$$\text{sp}(\text{inf}, H, K, e) \geqq 0.$$

5

Lack of space prevents us from going any deeper into the significance of the concepts and theorems indicated in the last section. It seems that the theory of semantic information might be fruitfully applied in various fields, for instance in the Theory of Design of Experiments [1] and in Test Theory.[2]

In view of the many misunderstandings and misapplications, in which Communication Theory has been involved, it would be desirable to undertake a clarification of its foundations. So would be a comparison between the theory outlined here and Communication Theory. These two tasks will be undertaken elsewhere by one of the authors (B.-H.).

Appendix

The language systems dealt with in this paper contain a finite number of *individual constants* which stand for *individuals* (things, events, or positions) and a finite number of *primitive one-place predicates* which designate primitive properties of the individuals. In an *atomic statement* e.g., 'P_1a_1' ('the individual a_1 has the property P_1'), a primitive property is asserted to hold for an individual. Atomic statements and statements formed out of one or more of them with the help of the customary connectives of negation,

[1] Indeed, R. A. Fisher defined in *The Design of Experiments*, Edinburgh and London, 1935 (in a less developed form already in papers dating back to 1922), a concept of Amount of Information which is, however, only distantly related to that developed here. Fisher's concept is certainly not a communicational one, but it is, like the communicational one, of a statistical, and not of a semantical, nature.

[2] Recent work done by Lee J. Cronbach at the University of Illinois seems to point in this direction. See, e.g., his preliminary report 'A Generalised Psychometric Theory Based on Information Measure', mimeographed, March 1952.

' $\sim$ ' (' not '), of disjunction, ' v ' (' or '), of conjunction, ' . ' (' and '), of (material) implication, ' $\supset$ ' (' if . . . then '), and of (material) equivalence, ' $\equiv$ ' (if and only if), are *molecular statements*. All atomic statements and their negations are *basic statements*. It is known that with the help of these tools numerical statements can be formed. Hence absolute frequencies (cardinal numbers of classes or properties) and relative frequencies can be expressed in them (but not measurable quantities like length and mass).

Any sentence is either *L-true* (logically true, analytic, e.g. ' $P_1a_1 \vee \sim P_1a_1$ ') or *L-false* (logically false, self-contradictory, e.g. ' $P_1a_1 . \sim P_1a_1$ ') or *factual* (logically indeterminate, synthetic, e.g. ' $P_1a_1 \vee \sim P_2P_3$ '). Logical relations can be defined, e.g. ' The statement i *L-implies* the statement j ' for ' i $\supset$ j is L-true ', ' i is *L-equivalent* to j ' for ' i $\equiv$ j is L-true ' ' i is *L-disjunct* with j ' for ' ivj is L-true '.

A *state-description* is a conjunction containing as components for every atomic statement either this statement or its negation, but not both, and no other statements. Thus a state-description completely describes a possible state of the universe in question. For any statement j of the system the class of those state-descriptions in which j holds, i.e. each of which L-implies j, is called the *range* of j. The range of j is null if, and only if, j is L-false ; in any other case j is L-equivalent to the disjunction of the state-descriptions in its range.

Research Laboratory of Electronics,
Massachusetts Institute of Technology,
Cambridge, Mass., U.S.A.

YEHOSHUA BAR-HILLEL

Department of Philosophy,
University of Chicago,
Chicago, Ill., U.S.A.

RUDOLF CARNAP