37

SEMANTIC INFORMATION*

YEHOSHUA BAR-HILLEL
Research Laboratory of Electronics, Massachusetts Institute of Technology

and

RUDOLF CARNAP
University of Chicago, U.S.A.

SEMANTIC INFORMATION AND ITS AMOUNTS

THEORY 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 WIENER and SHANNON, have nothing to do with what these symbols symbolize, but only with the frequency of their occurrence. The probabilities which occur in the definientia of the definitions of the various concepts in information theory are just these frequencies, absolute or relative, sometimes perhaps estimates of these frequencies.

This deliberate restriction of the scope of information theory was of great heuristic value and enabled the 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 statistical information theory to fields in which the term 'information' was used, presystematically, in a semantic sense, *i.e.* one involving contents or designata of symbols, or even in a pragmatic sense, *i.e.* one involving the users of these symbols. Important as the clarification of the function of the term 'information' in these senses may be, and there can hardly be a doubt as to this importance, 'information', as defined in present information theory, is not a suitable explicatum for these presystematic concepts and any transfer of the properties of this explanation to the fields in which these concepts are of importance 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 the 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 such as 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 explication of a psychological concept of information, than the statistical concept of present day theory.

We shall attempt to show that the fundamental concepts of the theory of

* This paper will appear in Brit. F. Phil. Sci., Aug. (1953).

semantic information can be defined in a very straightforward way on the basis of the theory of inductive probability that has been recently developed by CARNAP[1]. We shall also show that, in spite of an apparent strong restriction of the arguments over which the various information functions we shall propose are ranging, the whole theory of statistical information can be mapped onto our theory of semantic information, and finally that many theorems of our theory for which formally corresponding theorems exist in the statistical theory shed new light on the latter and provide also for a much wider field of application.

For an extensive presentation of the terminological background, reference may be made to CARNAP[1] or, for a more concise presentation, to CARNAP[2]. In the Appendix an even more concise summary is offered.

Let us state only that what follows refers to a fixed language system L_n^π, by which we mean, approximately, an applied first order functional semantical system with identity with n individuals say, $a_1, a_2, \ldots, a_n$, and with π primitive properties, say $P_1, P_2, \ldots, P_\pi$. 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 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 easily be 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 certainly be regarded as a necessary, though certainly not sufficient, condition of adequacy of any proposed explicatum 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 great length[1].

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[1]. For each of the latter m-functions, as they are called[1], a corresponding content measure function is defined simply by cont (i) = $m(\sim i)$. Cont (i) (the content measure of i) is offered as only one explanation 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 Carnap[1], we have

$$0 \leq \text{cont}\ (i) \leq 1$$

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 which

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

and the interesting additivity theorem

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

if and only if i and j are *L*-disjunct immediately follow.

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 as 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 i by itself. At times, however, we are as much, or even more, interested in the information conveyed by this statement in excess of that conveyed by some other statement or 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 minus sign 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 coat (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 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 'amount 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 normalization, that the amount of information conveyed by a basic statement shall be 1, it can easily be seen, along well known lines or computation, that these stipulations are fulfilled if we define this second function, to be called 'measure of information' and denoted by 'inf', as

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

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

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

The last equation is of course analogous to a form well known to every information theoretician.

Among the various theorems regarding inf let us mention

$$0 \leq \text{inf}\,(i) \leq$$

and the theorem of additivity, which, however, meets 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).

If anyone should find it strange that inf $(i \,.\, j)$ may be greater than inf (i) + inf (j), this is probably 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

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

where $c(j, i)$ is the degree of confirmation of (the hypothesis) j on (the evidence) i, defined by Carnap[1] as $\dfrac{m(i \cdot j)}{m(i)}$.

The statistical correlate of inf has found a large field of application in communication engineering and (unjustifiedly, as we tried to show before) in many other fields. Neither cont nor its statistical correlate have found any useful application so far. It is, however, to be expected that that facet of the amount of information which is measured by cont, to wit, a different one from that measured by inf, should find its fields of application too, especially so since cont is a mathematically simpler function of m than inf.

DEDUCTIVE AND INDUCTIVE *m*-FUNCTIONS

Among the various *m* functions on which cont and inf are based, there are two groups of special importance. The first group consists of just one member, to be designated here by 'm_D'; 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 1, on no evidence at all or, in other words, on the tautological evidence, and the same c_D-value on the evidence '$P_1a_1 . P_1a_2 . P_1a_3$' . Pias'. 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 deductive logic only are of this type, hence the subscript '*D*' for 'deductive'.

In those situations in which inductive logic is to be applied, only such *m*-functions may be used as allow us to learn from experience, in other words, fulfil the Principle of Instantial Relevance. The '*I*' in '*m*,', 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 specific theorems can be proven. For lack of time, 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.

ESTIMATIONS OF AMOUNTS OF INFORMATION

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 *n* 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. CARNAP[1] shows that in many similar situations the *c*-mean estimate of the function in question will be satisfactory. Confining ourselves, for the sake of simplicity and easy comparability with prevailing statistical information 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

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

From this definition and from a prior theorem on inf (h_p/e) the theorem

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

immediately follows. The statistical correlate of this theorem is, of course, 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.

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 1952' is ½, the c-value of h_2, 'There will be temperate weather . . . ' is ¼, and the c-value of h_3, 'There will be cold, weather . . . ' is ¼, then

$$\text{est (inf, } H, e) = - \sum_p \; c(h_p, e) \times \text{Log } c(h_p, e) = \tfrac{1}{2} \times 1 + \tfrac{1}{4} \times 2 + \tfrac{1}{4} \times 2 = 1 \cdot 5$$

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

If $H = \{h_1, \ldots, h_n\}$ and $K = \{k_1, \ldots, k_m,\}$ are exhaustive systems, 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, hence

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

We have, of course

$$\text{est (inf, } H \,.\, K, e) \leq \text{est (inf, } H, e) + \text{est (inf, } K, e),$$

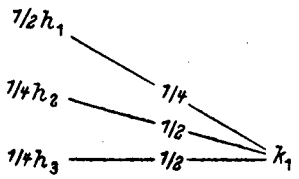
with equality holding if and only if the *h*'s and the *k*'s are mutually irrelevant.

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(*H, k, e*)'. The formal definition is

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

It is easy to see that sp(*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 $c(k_1, e \,.\, h_1) = \tfrac{1}{4}$, $c(k_1, e \,.\, h_2) = c(k_1, e \,.\, h_3) = \tfrac{1}{2}$. The following diagram will help to visualize the situation

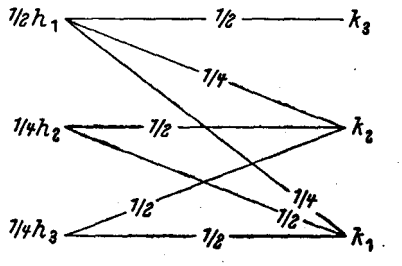


It is easy to compute, or to read from the diagram, that $c(h_l, e \,.\, k_1) = c(h_2, e \,.\, k) = c(h_3, e \,.\, k_1) = \frac{1}{3}$. Hence est(inf, $H, e \,.\, k_1$) = Log 3 = 1 · 585 and sp(inf, H, k_1, e) = −0·085.

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 by

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

Let us complete our illustration in the following diagram



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

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

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

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

In our example, est(sp, *H, K, e)* = 1·5 − 1·189 = 0·311.

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

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

$$\text{est(sp, } H, K, e) = \text{est(sp, } K, H, e)$$

$$\text{est(sp, } H, K, e) \geq 0$$

CONCLUSION

Lack of time prevents us from going any deeper into the significance of the concepts and theorems developed in the last section. Let it be said only that the applicability of the semantic theory of information goes apparently far

beyond that of the statistical theory. One immediately obvious additional field is that connected with design of experiments. It seems clear that, *ceteris paribus,* an experimental situation should be so designed that the estimate of the amount of information which is to be conveyed by the outcome of the experiment should be a maximum.

We would like to stress, finally, that the statistical theory can be mapped without remainder on to the semantic theory, but not vice versa. It might look, at first sight, as if the statistical theory is much more liberal than the semantic one since the former allows, as arguments for its amount-of information function, not only statements but symbols of any kind whatsoever. But this advantage is only apparent. To the expression 'the amount of information conveyed by the symbol *s*' the expression, 'the amount of information conveyed by the statement, "The symbol s is transmitted"', can be correlated. The statement, 'The symbol *s* is transmitted', belongs, *nota bene,* to the metalanguage of the language to which *s* itself belongs. But this is exactly as it should be. Notice that for the evaluation of the measure of information of our statement 'The symbol *s* is transmitted', the meaning of the object linguistic symbol is completely irrelevant but not, of course, the meaning of this metalinguistic statement itself. Only due to this meaning, an *m*-value is assigned to it, on the basis of which its inf-value is then determined. For the prevailing statistical information theory there exists a complete (and richer) counterpart within the semantic theory, *viz*., within the semantic theory of the corresponding metalanguage, but to another part of the semantic theory, namely to that dealing with the object-language itself, no statistical counterpart exists. This assertion is not meant as a criticism, since this limitation of the statistical theory was deliberately planned and, historically speaking, fully justified in the context of its origination. The theory presented here intends to give the outlines of a more comprehensive structure.

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_l' (`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 '~' ('not'), of disjunction, 'v' ('or'), of conjunction, ' . ' ('and'), of (material) implication, '$\supset$ ' ('if . . . then '), and of (material) equivalence, '$\equiv$' ('if . . . then'), and of (material) equivalence, `D' ('if and only if'), are molecular statements. All atomic statements and their negations are basic statements. The systems contain also variables with universal and existential quantifiers and the sign of identity, '='. It is well 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_la_l v ~P_1a_1') of *L*-false (logically false, self-contradictory,. *e.g.* 'P_1a_1 . ~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', and '*i* is *L*-exclusive of *j*' for '$i \,.\, j$ is *L*-false'.

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.

For a system with n individual constants and π primitive predicates, there are obviously πn atomic sentences and $2^{\pi n}$ state descriptions.

REFERENCES

[1] CARNAP, R., *Logical Foundations of Probability,* Chicago, 1950

[2] —, *The Continuum of Inductive Methods,* Chicago, 1952

[3] A more detailed and systematic study of semantic information appears in *Tech. Rep. Electron. Mass. Inst. Tech.,* No. 247

DISCUSSION

D. M. MACKAY: Contributions to the semantic hygiene of the field of information theory are badly needed, and it is good to find some trans-Atlantic interest in the matter. It is a pity that Dr. Bar-Hillel was not at our 1950 symposium, as I should have liked to ask him about the relation of his and Prof. Carnap's approach to that which I outlined there and in an earlier paper*. In discussion with Prof. Carnap last year it seemed that the concepts of metron content and content measure had much in common, though I am not clear that the additivity conditions are the same.

It should perhaps be repeated for the sake of clarity that the notion of metron content, outlined as it was first in a lecture at King's College in January, 1948, is in no sense an extension of the concepts of statistical communication theory, which had not then been published. In the 1950 symposium paper on Nomenclature I tried to indicate the connection of the two concepts, and I would certainly agree with the authors (as Shannon does) that 'information' in the sense of Shannon's '*H*' is quite useless as an explicatum for the semantic content of a representation. It might, incidentally, have been helpful if the authors had applied their discipline of rigour to their own usage, and spoken of Shannon's statistical theory (as he does) as the theory of communication. Their opening sentences are true in relation to the latter, but quite false in relation to general information theory as understood and practised on this side of the Atlantic at least, where quantitative concepts of information were, indeed, first introduced in connection with the design of experiments. Since the observation of a communication signal is only a particular class of experiment, it seems obvious that statistical communication theory could be 'mapped without remainder' on to the general theory of information but not vice versa. But has this ever been doubted, except perhaps by those who have tried to ride a wagon hitched to the wrong horse?

* MACKAY, *D. M., Phil. Mag.,* 41 (1950) 289.

In view of the importance of the present paper as a rigorous contribution towards what I have called the analysis of representations, it would be specially helpful if the authors could try to relate their terminology to that which has been used in earlier and undoubtedly less rigorous work. Is it permissible to regard their 'number of primitive properties', for example, as the equivalent of logon content, in the widened sense in which I have used it for the dimensionality of representational space?

Such correlation should be much helped by dropping the assumption that all current sense of the term 'information content' stem from applied statistical mechanics.

Y. BAR-HILLEL In reply: I welcome Dr. MacKay's remarks to our paper. I deplore, together with him, the replacement of the unambiguous term 'statistical theory of communication' by 'theory of information', but this replacement is a historical fact, at least in the United States. How this came about, what are the confusions which originated, and what confusions were caused by it, is an interesting chapter in the sociology of science that should be treated some day. Let me only remark here that Fisher's work was little known at the time when the statistical theory of communication was developed at the Bell Telephone Laboratories and at M.I.T. And it is a fact, though a deplorable one, that many American scientists, mainly linguists and psychologists, are riding a wagon hitched to the wrong horse, though this is certainly not Shannon's fault.

With regard to the connection of our work with MacKay's own prior contributions, I am not able to make, at this moment, much more definite statements than he himself suggests. There seems to be a relationship between the primitive properties and the dimensions of the representational space, and there might be some connection between the properties forming a family of properties—a concept that does not appear in our present treatment but is a simple generalization of it-and the metron content. It is, however, my impression that the closest relation exists between our concepts and. what is called by MacKay 'selective information content'. His remarks there are too, concise to allow for more definite statements.