A New Way of Linking Information Theory with Cognitive Science

Chris Thornton
Informatics
University of Sussex
Brighton
BN1 9QH
UK
c.thornton@sussex.ac.uk

Abstract

The relationship between the notion of *information* in information theory, and the notion of *information processing* in cognitive science, has long been controversial. But as the present paper shows, part of the disagreement arises from conflating different formulations of measurement. Clarifying distinctions reveals it is the context-free nature of Shannon's information average that is particular problematic from the cognitive point of view. Context-sensitive evaluation is then shown to be a way of addressing the problems that arise.

Introduction

One of the longest standing puzzles of cognitive science is what to think about information theory. Set out in its standard formulation more than 60 years ago, this framework (Shannon and Weaver, 1949) is acknowledged to be a remarkably general and precise area of mathematics. So it is of great interest to discover whether the notion of *information* developed in information theory has anything to do with the notion of *information processing* at the heart of cognitive science.

In the original publication, Shannon notes that 'the semantic aspects of communication are irrelevant to the engineering aspects' (Shannon and Weaver, 1949, p. 31). On the assumption that information processing in cognitive science deals with semantic aspects in particular, a fundamental disconnect seems implied. But this is muddled somewhat by the qualification (in Weaver's contribution to the joint publication) that Shannon's assertion 'does not mean that the engineering aspects are necessarily irrelevant to the semantic aspects' (Shannon and Weaver, 1949, p. 8). Adding to the ambiguity, researchers such as Meyer (1957/1967), Miller (1953), Garner (1962), Mackay (1956) and Attneave (1959) note a range of ways in which issues of a semantic nature can be addressed in information-theoretic terms. Quinlan (1993) and others demonstrate algorithms that operate specifically on this basis. Recent decades have also seen increasing use of information-theoretic quantification in cognitive neuroscience (e.g. Tononi et al., 1996; Lungarella et al., 2005; Friston, 2010).

The range of positions adopted on this issue deepens the mystery. Haber (1983) argues that information-theoretic measures cannot address psychological questions due to being 'entirely independent of the recipient' (Haber, 1983, p. 71). Temperley (2007), on the other hand, takes the view that the difficulty with them is they are calculated purely from the perspective of the recipient. Luce takes the view that information theory cannot address questions about structural representation of content because the 'elements of choice in information theory are absolutely neutral and lack any internal structure' (Luce, 2003, p. 185). On the other hand, a community of researchers examines ways in which information-theoretic quantification can explain emergence of structural representation in sensory processing (e.g. Attneave, 1959; Barlow, 1961; Uttley, 1979; Srinivisan et al., 1982; Atick, 1992).

For Haber, it is beyond dispute that 'the demise of information theory in psychology' has already occurred (Haber, 1983, p. 71). But intermediate positions are also common. Barwise notes that while 'traditional information theory is not a semantic theory at all' it 'puts important constraints on cognitive theories' (Barwise, 1983, p. 65). Churchland and Churchland (1983) are more positive still, seeing information theory as having a significant 'role to play in an account of cognition' (Churchland and Churchland, 1983, p. 67), and arguing the connection can be made through something called 'calibrational content' specifically, where this is defined to be informationally quantifiable 'measurement or detection concerning the status of the objective world' (Churchland and Churchland, 1983, p. 67). Others are doubtful of there being any connection at all. Dretske, for example, argues that information theory does not even 'deal with information as it is ordinarily understood' (Dretske, 1983, p. 56).[1]

The present paper argues that one of the reasons the situation has become so confused is that the debate has conflated different formulations of measurement.[2] The notion of measurement at the heart of the framework is the logarithmic principle, originally proposed by Hartley

[1] Curiously, this view is part of an informational epistemology. However, Dretske's hard-line is consistent with the fact that his account has little in common with information theory (Sayre, 1983). In Kyburg's view 'Dretske seeks to clothe a relatively traditional approach to epistemology in new information-theoretic clothes' (Kyburg and Jr, 1983, p. 72).

[2] I ignore areas of the framework (concerned with noisy and/or non-discrete systems) that have not figured in the debate.

(1928). There is also the probabilistic formulation of the logarithmic principle: $-\log p$. Finally, there is the averaging formula

$$-\sum_x p(x) \log p(x)$$

This is called the entropy. These formulations build on each other. The averaging formula uses the probabilistic formulation, which is itself based on the logarithmic principle.[3] But the three formulations have different implications for the question of connectivity with cognitive science.

The position often taken is that there is one form of information-theoretic quantification, and it is the averaging formula. Information measurement is taken to involve calculations of entropy specifically (e.g. Dretske, 1983; Sayre, 1983; Luce, 2003). This may be a consequence of the extent to which the results of (Shannon and Weaver, 1949) are derived by means of this formulation. But these results involve objectives of telecommunications specifically. Regarding the objectives of cognitive science, the logarithmic principle and the probabilistic formulation are equally of interest.

The present paper reviews the steps that lead from the logarithmic principle to Shannon's averaging formula. Account is taken of the semantic implications of different stages of the argument. Some aspects of the connectivity debate are clarified along the way, and consideration is given to the problems that arise from the use of context-free forms of measurement. Derivation of context-sensitive quantities is shown to be a viable alternative, and some examples are set out that show how this approach connects to the representational concerns of cognitive science.

Context-sensitive information

Mathematical quantification of information begins with the logarithmic principle. Proposed originally by Hartley (1928), this has a number of foundations, as reviewed by Shannon (Shannon and Weaver, 1949, pp. 31-33). Where an outcome is within a known set, the informational value must relate to the number of outcomes in the set. A simple way of measuring the informational value of something that reveals a particular outcome is thus in terms of the number of possible outcomes that might have been revealed. This is a potential way to measure the informational value of a 'message' to a 'receiver' then, to use Shannon's own terminology. But as Hartley points out, it is much better to use a logarithmic function of the number of outcomes. This yields a measurement in which the quantity of information is also the number of digits needed to *identify* the outcome, provided the same base is used for logarithm and digits. The usual approach uses base 2. The quantify of information can then be stated in terms of 'bits' (short for BInary digiTS). The measure quantifies both the amount of information obtained, and the number of binary digits needed to encode the outcome.

On the logarithmic principle, then, the informational value of anything that reveals one of n outcomes is just $\log n$. To obtain a value measured in bits, we take the logarithm to base 2. (Use of this base is assumed henceforth.) The process can be illustrated using any set of mutually exclusive outcomes. Let's say a new regulation requires Wi-Fi hotspots to be classified according to level of service, with the possible classifications being W1, W2, W3 and W4. Given there are four possible outcomes, the informational value of anything that gives the classification of a hotspot is then $\log 4 = 2$ bits. This is also the number of base 2 (binary) digits required to identify a classification.

An advantageous property of the logarithmic principle is that it generalizes straightforwardly to the case where outcomes have different probabilities. Instead of defining the information obtained from a one-in-n outcome as $\log n$ bits, it can be defined more generally as $-\log p$ bits, where p is the probability of the outcome. This accommodates the simple case of equiprobable outcomes, since $-\log \frac{1}{n} = \log n$. But it also accommodates there being a mixture of probabilities.

Let's say Wi-Fi hotspots are classified as W1 with probability $\frac{1}{2}$, as W2 with probability $\frac{1}{4}$, and as W3 and W4 with probability $\frac{1}{8}$. The discovery that a hotspot has a W4 classification is more informative in the sense of being contrary to expectation, than observing it has a W1 classification. This is reflected in the information value obtained. The value of a W1 classification is just $-\log \frac{1}{2} = 1$ bit, whereas the informational value of a W4 classification is $-\log \frac{1}{8} = 3$ bits.

The probabilistic formulation of the logarithmic principle also provides the means of calculating averages. Given $p(x)$ is the probability of outcome x, the average informational value of an outcome is the weighted average

$$-\sum_x p(x) \log p(x) \tag{1}$$

This is the entropy formula, centrepiece of Shannon's development of the logarithmic approach. It can be used whenever there is a probability distribution over outcomes. The distribution for Wi-Fi hotspots yields an average information value of 1.75 bits, for example.

The average information has a number of appealing properties. It can be seen as measuring the uncertainty that exists with respect to the outcomes, in the sense of quantifying the 'choice' allowed by the distribution (Shannon and Weaver, 1949, p 48-53). As a weighted

[3] In practice, Shannon derives the entropy formula as the only acceptable way of measuring the 'choice' permitted by a distribution.

average, it can also be seen as defining the information that an outcome is *expected* to have. Given $-\log p$ is an encoding cost, we can also look at the formula as the average cost of encoding an outcome. (Shannon proves the average cost can be no less: Shannon and Weaver, 1949, p. 62-64).

It is important to notice, however, that this approach makes no distinction between subjective and objective perspectives. In order for probability $p(x)$ to be what fixes the amount of information an agent obtains from outcome x, this must be the probability the agent attributes to x. On this basis, $p(x)$ is subjective. But where it is used in the averaging formula, $p(x)$ becomes the objective probability of x. In fact, Shannon's framework makes no distinction between subjective and objective probabilities. In the telecommunications context that is the framework's main focus, this makes sense. A telecommunications device adopting a personal perspective would be worthless. In other contexts, however, subjective factors may be of more relevance. It is of interest, then, to consider ways in which context-sensitive information values can also be calculated.

Consider the case where there is a set of two outcomes, both of which have information values calculated in an objective way (i.e., by the logarithmic principle). A *context-sensitive* value can then be calculated for any distribution attributed, and any outcome arising. This is just the expected value of the distribution in regard to the outcome. On the principle that probability attributed to the given outcome must increase the distribution's value, while probability attributed to any other outcome must decrease it, the expected information is a weighted average in which outcome values are either positive or negative:

$$I(P_S) = \sum_{x \in S} P_x \begin{cases} I(x) & \text{if } x \text{ is given} \\ -I(x) & \text{otherwise} \end{cases} \quad (2)$$

Here, S is the set of outcomes, P_S denotes the distribution attributed, and $I(x)$ is the informational value of outcome x. (Calculated by the logarithmic principle, $I(x) = \log |S|$).) This formula is valid whenever there are just two outcomes. Where there are more than two, the number of outcomes *not* given is greater than 1, and thus greater than the number given. It is then necessary to ensure commensurability between additions and subtractions by normalizing the latter with respect to $|S| - 1$, the number of non-given outcomes. The general form of the context-sensitive evaluation is thus

$$I(P_S) = \sum_{x \in S} P_x \begin{cases} I(x) & \text{if } x \text{ is given} \\ -\frac{I(x)}{|S|-1} & \text{otherwise} \end{cases} \quad (3)$$

This is the expected information value of distribution P_S to the attributing agent, where a particular element of S is given, and all outcomes have known information values. It can also be seen as measuring the degree to which the distribution predicts the outcome in question.

Context-free evaluation of information (e.g., Eq. 1) is valid in most situations. Hence the generality of Shannon's framework. But where subjectivity is a possibility, context-sensitive evaluation (by Eq. 3) is entailed. The effects of evaluating information inappropriately can be illustrated using the hotspots example again. Let's say a particular agent expects every hotspot to be a W1. The agent attributes a probability of 1 to the W1 classification, and a probability of 0 to W2, W3 and W4. In objective reality, however, not all hotspots are W1: at least one is classified as W2. There is a subjective context, then, requiring amounts of information to be calculated in a context-sensitive way.

Should we choose to evaluation information in a context-free way regardless, the results are likely to be meaningless. The attributed distribution places all probability on one outcome. Its entropy is zero. On the strength of this, the average informational value of each outcome is deemed to be zero bits. This is appropriate in the case of a W1 classification, since the agent deems this to be the outcome in *all* cases. Unfortunately, it is also the value in the case of a W2 classification, which is a case the agent deems to be impossible.

This nonsensical result is a consequence of applying context-free evaluation to a situation in which there is a subjective context. On the context-free interpretation, there cannot *be* a W2 classification: its assumed probability is zero. Given the subjective perspective that is in force, context-sensitive evaluation using Eq. 3 is required. This produces a result that makes more sense. The context-sensitive value is found to be 2 bits for a W1 classification, and $-\frac{2}{3}$ bits for any other classification. Notice the potential for negative context-sensitive values, in contrast with context-free (entropy) values, which are always non-negative.

The general difficulty that arises for cognitive science will then be evident. Situations of interest for this discipline involve subjectivity by definition. The tendency to equate information-theoretic evaluation with context-free measurement is thus an obstacle. But there is another aspect to the problem. Both context-free and context-sensitive forms of evaluation are calculated with regard to a set of mutually-exclusive outcomes. The evaluations obtained depend solely on probabilities attributed, and the number of outcomes in the set. The difficulty is that each outcome has the potential to *signify* something completely different. Information values reflect the original outcomes, rather than any interpretations that may be forthcoming, however. Where an additional semantics is imposed on outcomes, both context-free and context-sensitive values may be meaningless in relation to the interpretations that apply.

Evaluations that are context-sensitive in the sense of being calculated by Eq. 3 may thus fail to be context-sensitive with regard to an imposed semantics. There are thus *two* ways in which information-theoretic evaluations can be inadequate from the cognitive point of view. The semantic disconnect that commentators such as Luce (2003), Haber (1983) and Dretske (1983) see as inherent in information theory originates in these two ways.

Illustrations

A useful context for illustrating context-sensitive evaluation is that of weather forecasting. Imagine we live in a world where the weather has just two outcomes: *rain* and *sun*. Let's say the forecast issued by the local met office for a particular day is *showery*, and that this signifies 60% chance of *rain*, and 40% chance of *sun*. Assume the outcome is *rain*. Eq. 3 can then be used to obtain a context-sensitive value for the attributed distribution given this particular outcome. With the forecast being *showery*, *rain* is predicted with probability 0.6. The outcome is in fact *rain*, and the information value of each outcome is assumed to be $\log 2 = 1$ bit. The context-sensitive value of the distribution is thus $(0.6 \times 1) - (0.4 \times 1) = 0.2$ bits. If the outcome is *sun*, on the other hand, the value is $(0.4 \times 1) - (0.6 \times 1) = -0.2$ bits.

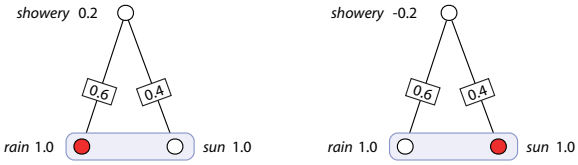


Figure 1: Context-sensitive evaluations.

The diagram of Figure 1 illustrates the two cases considered. In this and ensuing schematics, outcomes are represented by small circles, labeled with the outcome's name and informational value. Circles are filled where the outcome is given. Circles enclosed within the same bar are within the same choice of outcomes: the bar represents the choice. Where one outcome signifies a distribution over others—e.g., *showery* specifying *rain* with probability 0.6—the relationships are indicated using connecting lines. Annotations placed over these lines show the probabilities that are attributed.

The figure shows evaluations of the *showery* distribution for the two outcomes *rain* and *sun*. Notice how the values reflect the degree to which the distribution predicts the outcome given. The evaluation is negative where the implied distribution mispredicts the outcome, and positive otherwise. At the same time, its relatively indiscriminate nature ensures both values are small compared with those of the outcomes themselves.

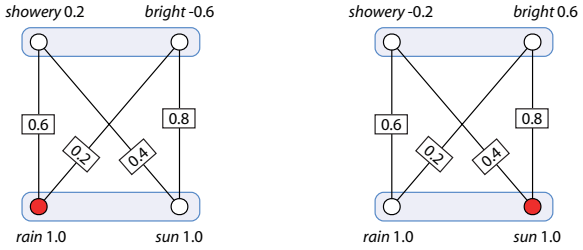


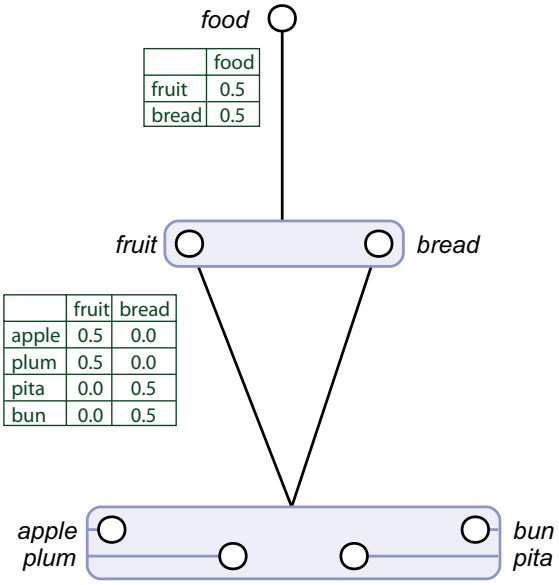
Figure 2: Derived evaluation of outcomes.

In the illustrated scenario, distributions are signified by entities (i.e., forecasts) that are themselves outcomes. By definition, these inherit the informational values of the distributions they designate. Any higher-level distribution must then be evaluated in terms of the derived values of predicted outcomes. To illustrate, let's say that in a certain season the forecast is always either *showery* or *bright*, with the latter meaning 20% chance of *rain* and 80% chance of *sun*. Context-sensitive values for these forecasts are then derived as in Figure 2. Potentially there can then be a second level of structure. A forecast of *unsettled* might mean 70% chance of *showery* and 30% chance of *bright*. The context-sensitive value of this forecast would then be calculated in terms of the derived values of *showery* and *bright*, rather than values obtained by the logarithmic principle.

Analysis of representation

Context-sensitive evaluations can be calculated wherever we have both a distribution and a given outcome. Where one outcome signifies such a distribution itself, the value obtained also belongs to the signifying outcome, as noted above. Context-sensitive evaluations can thus be a way of evaluating probabilistic representation. Such evaluations can be made at multiple levels. Where one outcome signifies a distribution over several others, one of which does the same thing, there are two levels of representation. The latter is embedded within the former. Context-sensitive measurement of information is a way of evaluating outcomes at multiple levels of representation.

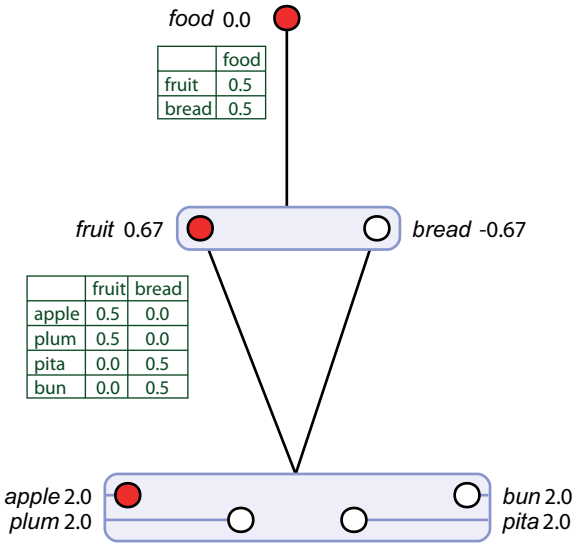
An assembly of signifying outcomes is a kind of representation structure, then. Such structures can take any form we like. For example, we might configure a representation structure in a way that expresses a category hierarchy. Let's say we have three categories as follows: a *fruit* category, in which the members are *apple* and *plum*; a *bread* category in which the members are *pita* and *bun*, and a *food* category in which the members are *fruit* and *bread*. To express this category hierarchy as a representation structure, categories must be treated as linked outcomes. Since category members are equiprobable instances of their category, member outcomes are always



	food
fruit	0.5
bread	0.5

	fruit	bread
apple	0.5	0.0
plum	0.5	0.0
pita	0.0	0.5
bun	0.0	0.5

Figure 3: Representation structure in the form of a category hierarchy.



	food
fruit	0.5
bread	0.5

	fruit	bread
apple	0.5	0.0
plum	0.5	0.0
pita	0.0	0.5
bun	0.0	0.5

Figure 4: Context-sensitive evaluation as classification.

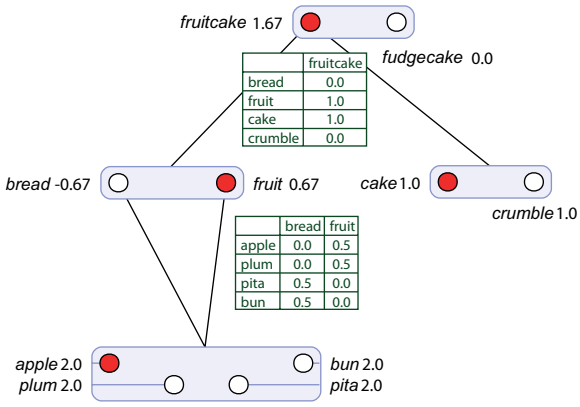
equiprobable attributions of the corresponding category outcome. The representation structure expressing the category hierarchy is thus the one of Figure 3. (Notice this diagram tabulates the probabilities involved, rather than displaying them on an individual basis.)

Regardless of what a representation structure expresses, it retains its capacity for informational evaluation. This can be a way of explaining the functionalities that are forthcoming. Where a representation structure is arranged as a category hierarchy, for instance, there is the possibility of explaining classifications mathematically. Classifying an outcome in a particular way can be seen to identify the category outcome with the highest context-sensitive evaluation.

Consider the values that are obtained in the representation structure of Figure 3, where *apple* is given. These are shown in Figure 4. The context-sensitive value of the correct classification (*fruit*) is 0.67 bits, whereas the value of the incorrect classification (*bread*) is -0.67 bits. The classification can be explained as identifying the most informative category outcome.

Representation structures can be arranged in a broad range of ways and can thus express any model constructed in terms of representational relationships. Their probabilistic foundation means they can represent conventional Bayesian models, for example. Being able to incorporate multiple levels of representation, they can express hierarchical Bayesian models. Another possibility is schematic models, involving representational relationships of a conjunctive nature.

Consider a schematic model in which a particular entity is considered to be a combination of other entities. Viewed as a representation structure, this is a case in which one outcome designates multiple distributions, each of which concentrates probability on a single outcome. Such cases can be analyzed using context-sensitive evaluation in the usual way. But in so doing it is necessary to take the possibility of multiple designations into account. This must be done in accordance with the principle that information can be summed only if is independent. Where distributions are not independent, the evaluation obtained is taken to be the maximum (i.e., greatest independent value) rather than the sum of values arising.



	fruitcake
bread	0.0
fruit	1.0
cake	1.0
crumble	0.0

	bread	fruit
apple	0.0	0.5
plum	0.0	0.5
pita	0.5	0.0
bun	0.5	0.0

Figure 5: Representation structure with disjunctive and conjunctive elements.

Figure 5 extends the *bread/fruit* scenario to illustrate what happens where representation structure includes conjunctive designations of this type. The *bread/fruit* structure from the previous diagram is seen here in the

lower-left corner. At the same level of representation, there is a *cake/crumble* choice. At the top level of representation, the outcome *fruitcake* is specified in a way that requires both *fruit* and *cake*. The effect is to reproduce the conjunctive character of a schema. As previously, the evaluations arising can explain classifications. If *apple* and *cake* are both given, the context-sensitive value of *fruitcake* is 1.67 bits. In the case of *fudgecake*, the value is 0 bits. Classifying a composite of apple and cake as *fruitcake* is then explained in terms of this category being most informative for the given context.

Conclusion

The traditional objection to use of information theory in cognitive science has been the assumption that it does not deal with semantic aspects of information. On close examination, this is found to be an over-simplification. Where information values are calculated by means of the entropy formula, they are context-free in the sense of ignoring any element of subjectivity. They may also be context-free in the trivial sense of ignoring a superimposed semantic interpretation. The latter problem can be resolved simply by outlawing such applications. The former can be resolved by pursuing evaluation in a way that takes subjective context into account. On this basis, information-theoretic evaluation can be of relevance to cognitive science. Specifically, it can be a way of mathematically explaining category representation.

References

Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing. *Network*, *3* (pp. 213-251).

Attneave, F. (1959). *Applications of Information Theory to Psychology.* New York: Henry Holt

Barlow, H. B. (1961). Possible principles underlying the transformation of sensory messages. In Rosenblith (Ed.), *Sensory communication* (pp. 217-234). Cambridge: MIT Press

Barwise, J. (1983). Information and semantics (commentary on Précis of *Knowledge and the Flow of Information*). *Behavioral and Brain Sciences*, *6* (pp. 65-66).

Churchland, P. M. and Churchland, P. S. (1983). Content: semantic and information-theoretic. *Behavioral and Brain Sciences*, *6* (pp. 67-68).

Dretske, F. I. (1983). Précis of *Knowledge and the Flow of Information. Behavioral and Brain Sciences*, *6* (pp. 55-90).

Friston, K. (2010). The free-energy principle: a unified brain theory? *Nat. Rev. Neurosci*, *11*, No. 2 (pp. 127-138).

Garner, W. R. (1962). *Uncertainty and Structure as Psychological Concepts.* New York:

Haber, R. N. (1983). Can information be objectivized? *Behavioral and Brain Sciences*, *6* (pp. 70-71).

Hartley, R. L. (1928). Transmission of information. *Bell System Technical Journal* (pp. 535).

Kyburg, H. E. and Jr, (1983). Knowledge and the absolute. *Behavioral and Brain Sciences*, *6* (pp. 72-73).

Luce, R. D. (2003). Whatever happened to information theory in psychology. *Review of General Psychology*, *7*, No. 2 (pp. 183-188).

Lungarella, M., Pegors, T., Bulwinkle, D. and Sporns, O. (2005). Methods for quantifying the information structure of sensory and motor data. *Neuroinformatics*, *3*, No. 3 (pp. 243-262).

Mackay, D. (1956). Towards an information-flow model of human behaviour. *Br. J. Psychol*, *43* (pp. 30-43).

Meyer, L. B. (1957/1967). Meaning in music and information theory. *Music, the Arts, and Ideas* (pp. 5-21). Chicago: University of Chicago Press

Miller, G. A. (1953). What is information measurement? *American Psychologist*, *8* (pp. 3-11).

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning.* San Mateo, California: Morgan Kaufmann

Sayre, K. M. (1983). Some untoward consequences of Dretske's "causal theory" of information. *Behavioral and Brain Sciences*, *6* (pp. 78-79).

Shannon, C. and Weaver, W. (1949). *The Mathematical Theory of Communication.* Urbana, Illinois: University of Illinois Press

Srinivisan, M. V., laughlin, S. B. and A, A. D. (1982). Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. Lond. B*, *216* (pp. 427-59).

Temperley, D. (2007). *Music and Probability.* Cambridge, Massachusetts: The MIT Press

Tononi, G., Sporns, O. and Edelman, G. (1996). A complexity measure for selective matching of signals by the brain. *Proceedings of the Nat. Academy of Science*, *93* (pp. 3422-3427).

Uttley, A. M. (1979). *Information Transmission in the Nervous System.* London: Academic