entropy

MDPI

Review

Photons, Bits and Entropy: From Planck to Shannon at the Roots of the Information Age

Mario Martinelli

Dipartimento di Elettronica Informazione e Bioingegneria, Politecnico di Milano, Milano 20133, Italy; mario.martinelli@polimi.it; Tel.: +39-02-2399-3676

Received: 26 May 2017; Accepted: 4 July 2017; Published: 8 July 2017

Abstract: The present age, which can be called the Information Age, has a core technology constituted by bits transported by photons. Both concepts, bit and photon, originated in the past century: the concept of photon was introduced by Planck in 1900 when he advanced the solution of the blackbody spectrum, and bit is a term first used by Shannon in 1948 when he introduced the theorems that founded information theory. The connection between Planck and Shannon is not immediately apparent; nor is it obvious that they derived their basic results from the concept of entropy. Examination of other important scientists can shed light on Planck's and Shannon's work in these respects. Darwin and Fowler, who in 1922 published a couple of papers where they reinterpreted Planck's results, pointed out the centrality of the partition function to statistical mechanics and thermodynamics. The same roots have been more recently reconsidered by Jaynes, who extended the considerations advanced by Darwin and Fowler to information theory. This paper investigates how the concept of entropy was propagated in the past century in order to show how a simple intuition, born in the 1824 during the first industrial revolution in the mind of the young French engineer Carnot, is literally still enlightening the fourth industrial revolution and probably will continue to do so in the coming century.

Keywords: entropy; photon; bit; second law; information theory; blackbody radiation; Planck; Shannon

1. Introduction

Internet traffic is entirely sustained by an optical backbone. It is the optical layer which acts as the "ultimate server" for any further development of the Internet. Therefore, the present age, which can be called the Information Age, depends on a core technology constituted by bits transported by photons. Without optical communication, there will not be enough bits, nor will they be economically available, to feed the continuous growing request for Internet packets. Both concepts, bit and photon, originated in the past century: "photon" is a term first used (as a neologism) by Lewis in 1926, but the concept was introduced by Planck in 1900 when he advanced the solution of the blackbody spectrum. "Bit" is a term first used by Shannon in 1948 when he developed information theory. Both concepts, bit and photon, share the same roots: nineteenth-century thermodynamics and the concept of entropy.

The link between Planck and Shannon may not be immediately apparent; neither is it obvious how they derived their basic results from the concept of entropy. Other important scientists shed light on Planck's and Shannon's results. In 1922, Darwin and Fowler published a couple of papers where they reinterpreted Planck's results and pointed out the centrality of the partition function to statistical mechanics and thermodynamics. The same roots have been more recently reconsidered by Edwin Jaynes, who, in a work written in 1957, extended the considerations raised by Darwin and Fowler to information theory.

This paper investigates how the concept of entropy has been propagated for the past century, from the works of Planck through Shannon, in order to show how a simple intuition, born in 1824 during the first industrial revolution in the mind of young French engineer Carnot, is literally still enlightening the fourth industrial revolution and probably will continue to do so in the coming century.

The work is organized in the following sections. In Section 2 (Planck and the Myth of Entropy), the Planck equation is derived from the reconstruction provided by leading historians of science (mainly Kuhn, Darrigol and Gearhart) with special attention to the role played by combinatorial analysis. Section 3 (Darwin, Fowler, Schrodinger and the Centrality of the Partition Function) describes how, in a pair of papers written in 1922, Darwin and Fowler proposed to derive Planck's equation according to a new formulation (still used today) based on the recovery of the "average value" of internal energy. They emphasized the centrality of the partition function in thermodynamics and in statistical mechanics. Together with Schrodinger, they shed light on the different concepts of entropy used by physicists.

Section 4 (Nyquist, Hartley and the Dawn of Information) describes Nyquist and Hartley's introduction, in the 1920s, of the seeds of the discipline eventually known as information theory. Like Darwin, Fowler, and Schrodinger, they built on the concept of entropy. Section 5 (Szilard, Brillouin and Beyond; Physicists Discover Information) discusses Szilard's work on Maxwell's demon paradox, through which the physics community became interested in information theory, an interest that continues in the present (with Brillouin, Landaurer and Bennet). Information theory is also discussed in Section 6 (Shannon and the Importance of the Channel Capacity), which discusses the founder of information theory. An analysis of his main theorems is given, highlighting continuity with the discussions in Sections 2–5 and the centrality of the concept of "channel capacity" in all his work.

Finally, in Section 7 (Jaynes or Synthesis), we describe Jaynes's effort to generalize the results reached by Shannon and reconcile these results with thermodynamics. His work continues that of Darwin and Fowler, and his synthesis is still seeking applications. Jaynes's work does not conclude the connection between thermodynamics and entropy. As mentioned above, the path has continued to be developed by introducing new themes that draw upon it: the reversibility of switching operations; the energy cost of the logic gates and memories; entropy and information of quantum systems. However, these themes are not discussed in this paper, which focuses on the common background shared by Planck and Shannon.

2. Planck and the Myth of Entropy

Planck's equation is so fundamental in the history of physics that many monographs and review papers have been written about its derivation. Among the monographs, the most important is the one written by Kuhn in 1978 [1], and the most recent is probably that published by Badino in 2015 [2], though an important contributiuon was also given by Agassi [3]. Milestone review papers have been recently given by Darrigol [4,5] and Gearhart [6].

Planck's equation establishes a universal relation between the temperature of a body and the spectrum of its radiative emission. The equation was the solution of the blackbody problem, a problem that interested the 19th century's physicists and the solution to which, the equation proposed by Planck in 1900, calls into question fundamental aspects of the structure of matter, thereby contributing to the transition from classic physics to quantum mechanics theory.

Modern approaches present the solution of the blackbody problem from a semi-classical [7] or quantum mechanics point of view [8]. Of course, this was not the case at the turn of the past century. Perhaps only few people know that Planck arrived at his equation by rejecting electrodynamic approaches (even if in the year 1900 Maxwell's electromagnetic theory had just been published) while moving towards the entropy concept (and thermodynamics concepts in general). Entropy was an idea originated in the 1824 by Carnot [9], reflecting on the growing technology of steam. Carnot published (at his own expenses) the "Reflexions sur la puissance motrice du feu" (Reflection on the motive power of fire), a text that the German scientist Clausius recognized [10] contained the fundamental concept

of entropy, a term first introduced by him in 1865 [11] together with the symbol and the equations connecting entropy with other thermodynamics parameters:

$$S_1 = S_0 + \int_l \frac{dQ}{T} \tag{1}$$

where S is entropy, Q heat, and T absolute temperature. The integral range is for a generic thermodynamics cycle l: when the cycle is reversible, the integral goes to zero (entropy is conserved, as guessed by Carnot). Otherwise, entropy unavoidably increases, leading to Planck's 1887 enunciation of the Second Law [1]: the principle of the increase of entropy. Since Planck considered the blackbody problem a case of thermal equilibrium, he looked for a solution corresponding to a maximum of entropy, as stated in one of his most popular books, the one he wrote in 1912 that was translated in 1914 by Masius under the title "The Theory of Heat Radiation" [12].

It is interesting to observe that Planck arrived at the solution to the blackbody problem after several frustrating attempts to obtain an electrodynamic solution. He in fact writes:

> Thus we see that the electrodynamical state is not by any means determined by the thermodynamic data and that in cases where, according to the laws of thermodynamics and according to all experience, an unambiguous result is to be expected, a purely electrodynamical theory fails entirely, since it admits not one definite result, but an infinite number of different results and before entering on a further discussion on this fact and of the difficulty to which it leads in the electrodynamical theory of heat radiation, it may be pointed out that exactly the same case and the same difficulty are met with in the mechanical theory of heat, especially in the kinetic theory of gases.

To solve the problem, Planck says that "there remains only one way out of the difficulty, namely, to supplement the initial and boundary conditions by special hypothesis of such a nature that the mechanical or electrodynamical equations will lead to an unambiguous result in agreement with experience" [12].

What is the "special hypothesis" introduced by Planck to solve the problem?

He says: "In mechanics, this is done by the hypothesis that the heat motion is a "molecular chaos"; in electrodynamics, the same thing is accomplished by the hypothesis of "natural radiation" which states that exist between the numerous different partial vibrations of a ray no other relations than those caused by the measurable mean values."

He continues: "If for brevity we denote any condition or process for which such a hypothesis holds as an 'elemental chaos', the principle, that in nature any state or any process containing numerous elements not in themselves measurable is an elemental chaos, furnished the necessary condition for a unique determination of the measurable processes in mechanics as well as in electrodynamics and also for the validity of the second principle of thermodynamics" [12].

Hence Planck assumes the second principle of thermodynamics as well as the existence of chaos as funding conditions for solving the blackbody problem. He also states, in the footnote of the same page (117), about the chaos hypothesis, that "wherever this hypothesis does not hold, the natural processes, if viewed from the thermodynamics (macroscopic) point of view, do not takes place unambiguously." After this clarification, Planck asks: "what mechanical or electrodynamical quantity represents the entropy of a state?"

Planck's answer is: "It is evident that this quantity depends in some way on the probability of the state. For since an elemental chaos and the absence of a record of any individual element forms an essential feature of entropy, the tendency to neutralize any existing temperature differences, which is connected with an increase of entropy, can mean nothing for the mechanical or electrodynamical observer but that uniform distribution of elements in a chaotic state is more probable than any other distribution" [12].

Hence Planck reaches the following fundamental proposition for any further discussions: "The entropy of a physical system in a definite state depends solely on the probability of this state" [12]; Thus, we have:

$$f(W) \tag{2}$$

where *f(W)* represents a universal function of the argument *W* which is "the probability of a physical system in a definite state" [12]. In order to better define what type of function connects *S* with *W*, Planck continues: "In whatever way *W* may be defined, it can be safely inferred from the mathematical concept of probability that the probability of a system which consists of two entirely independent systems is equal to the product of the probability of these two systems separately". Hence, by deploying the multiplicative properties of the probabilities, Planck arrives quickly at the relation

$$S = k \log W + const. \tag{3}$$

"an equation which determines the general way in which the entropy depends on the probability".

Hence Planck specifies that "the universal constant *k* is the same for a terrestrial as for a cosmic system the second additive constant of integration may", without any restriction as regards generality, be included as a constant multiplier in the quantity *W*, which here has not yet been completely defined, so that the equation reduces to

$$S = k \log W \tag{4}$$

Although Planck admits that "the logarithmic connection between entropy and probability was first stated by Boltzmann in his kinetic theory of gases" the above equation differs in its meaning from the corresponding one of Boltzmann in two essential points: "Firstly, Boltzmann equation lacks the factor *k*, which is due to the fact that Boltzmann always used gram molecules, not the molecules themselves, in his calculations. Secondly, and this is of greater consequence, Boltzmann leaves an additive constant undetermined in the entropy *S* as is done in the whole of classical thermodynamics, and accordingly there is a constant factor of proportionality, which remains undetermined in the value of the probability *W*".

Hence Planck concludes: "In contrast with this (Boltzmann theory) we assign a definite value to the entropy S. This is a step of fundamental importance which can be justified only by its consequences. As we shall see later, this step leads necessarily to the "hypothesis of quanta" and moreover it also leads, as regards radiant heat, to a definite law of distribution of energy of black body radiation". Again, "We shall designate the quantity *W* thus defined as the 'thermodynamics probability' in contrast to the 'mathematical probability' to which is proportional but not equal. For, while the mathematical probability is a proper function, the thermodynamics probability is, as we shall see, always an integer".

The above shows that for Planck the connection between entropy and probability (it is worth pointing out here, as will become more apparent a few pages later, that the term "probability" used here by Planck does not actually indicate probability as commonly understood but rather represents a weight, count, or multiplicity) of a state was not only fundamental but must be also definite in order to obtain an a posteriori justification of the introduction of energy quanta. Moreover, according to Gearhart, Planck introduced another fundamental passage because he "argued that the electromagnetic entropy could be identified with the thermodynamics entropy, the assumption of natural radiation leads to disorder, which in turn underlies irreversible processes" [6].

However, this was written in 1912. In the year 1900, according to several authors, Planck's vision was not so linear.

Let us follow Planck's reasoning according to the reconstructions made by Kuhn [1] and Darrigol [4] (where details can be found).

Planck began with the best-known representation of the experimental results, which had already been the object of Wien's theoretical speculation:

$$u = \left(\frac{8\pi\nu^2}{c^3}\right) \cdot \frac{b\nu}{e^{-a\nu/T}} \tag{5}$$

where u is the wanted energy density and the second factor is U, the average energy of the single oscillator. The frequency of the oscillator is represented by ν. However, low-frequency measurement made by Lummer and Pringsheim in March 1900 contradicted this function. According to Darrigol [4], in his communication to the German Academy of 19 October 1900, Planck acknowledged the mistake in his previous derivation of Wien's law and also the unquestionable experimental violation of the same law in the infrared regions of the spectrum. Then he proposed a new law. As suggested by his previous consideration about infinitesimal relaxation (of the oscillator) towards equilibrium, he focused again on the second order derivative of resonator entropy. For Wien's law (as in Equation (5)), this is

$$\frac{d^2S}{dU^2} = -\frac{1}{a\nu U} \tag{6}$$

Keeping in mind that this formula should remain a good approximation for small values of U, Planck imagined the simplest modification:

$$\frac{d^2S}{dU^2} = -\frac{1}{a\nu U(U + b\nu)} \tag{7}$$

The integration of this equation with the relation $dS/dU = 1/T$ gave the final formula

$$u = \left(\frac{8\pi\nu^2}{c^3}\right) \cdot \frac{b\nu}{e^{-a\nu/T} - 1} \tag{8}$$

where

$$U = \frac{b\nu}{e^{-a\nu/T} - 1} \tag{9}$$

is the wanted Planck's function, or average energy of the single resonator. Equation (8) reproduced the experiments with great accuracy (the constant b would later become h, Planck's constant, and the constant $a = h/k$, where k is the Boltzmann constant). Darrigol also suggested that "Planck made an interesting comment on the plausibility of his expression for d^2S/dU^2. It gave S as logarithmic function of U, "which is suggested by probability calculus". Two months later, on 14 December 1900, Planck's famous derivation of his new blackbody law starts with the words: "Entropie bedingt unordnung" (entropy conditions disorder). Then Planck recalls the manifestation of disorder in his resonator, the temporal irregularity of its phase and amplitude. Accordingly, the degree of disorder compatible with the energy U of the resonator is equal to the number of evolutions compatible with the time average energy U A semi-discrete version of this number is given by the number of distributions of energy $E = NU$ over a large set of N identical and independent resonators, since this set can be taken to represent a single resonator at N different times" [4].

In effect, as suggested by Kuhn [1], since Equation (9) "applies only to a single resonator it is therefore not yet suitable for interpretation in probabilistic terms. Planck's attempt to reformulate it may well have constituted an early stage of 'the most strenuous work' of his life". The introduction of many "numbers of evolution" solves the problem and allows the use of combinatorics to evaluate the "number of complexions"; thus, the number of microstates is compatible with a single macrostate, as already proposed by Boltzmann in dealing with the mechanical theory of the perfect gas.

Since Equation (9), completed with the Wien factor, fits exceptionally well with the experimental data, what Kuhn suggested is plausible: that Planck was "working backwards" [1] in order to recover

the combinatorial function expression W, which was inserted as the argument of Equation (4) in order to produce Equation (9). The steps passages can be summarized as follows [1].

Consider Equation (9): it can be obtained as a result of the relation $dS/dU = 1/T$ if S is given by the equation:

$$S = \frac{b}{a} \ln \left[\frac{\left(1 + \frac{U}{bv}\right)^{1+\frac{U}{bv}}}{\left(\frac{U}{bv}\right)^{\frac{U}{bv}}} \right] + const. \tag{10}$$

and for N identical and independent resonators we have:

$$S = \frac{b}{a} \ln \left[\frac{\left(N + \frac{P\varepsilon}{bv}\right)^{N+\frac{P\varepsilon}{bv}}}{N^N \left(\frac{P\varepsilon}{bv}\right)^{\frac{P\varepsilon}{bv}}} \right] + const. \tag{11}$$

where the total energy $E = NU$ has been subdivided in P elements of equal size e, as suggested directly by Planck, who says: "If E is taken to be an indefinitely divisible quantity the distribution is possible in an infinite number of ways. However, we regard E—and this is the essential point of the whole calculation—as made up of a completely determinate number of finite equal parts, and for this purpose this paper use the constant of the nature $b = 6.55 \times 10^{-27}$ erg sec. This constant, once multiplied by the common frequency of the resonators, gives the energy element e in ergs and by division of E by e gets the number P of energy elements to be distributed over the N resonator. When this quotient is not an integer, P is taken to be a neighboring integer" [4].

Hence, we obtain

$$S = \frac{b}{a} \ln \left[\frac{(N+P)^{N+P}}{N^N P^P} \right] + const. \tag{12}$$

This equation, apart from the constant and by putting $b/a = k$, is nothing more than the Stirling approximation of

$$S = k \ln \frac{(N+P-1)!}{(N-1)!P!} \tag{13}$$

where the argument of the log is

$$W = \frac{(N+P-1)!}{(N-1)!P!} \tag{14}$$

and where W is, according to Planck, "the combinations with repetitions of N elements taken P at a time" [12]. Equation (13) is the explicit expression of Equation (4) and allows Planck to demonstrate the direct derivation of his equation by a theory of thermodynamics and, significantly, by the expression of entropy.

Yet what sort of probability is Equation (14)? Planck says, "The number of all complexions which are possible with a given distribution in space we equate to the thermodynamic probability W of the space distribution" [12]. In effect, this was an easy use of the term probability, which has been continued from Planck's time until today. As also explained by Kuhn, Planck used the term "probability" with the meaning of "number of permutability" or, much more often, to indicate proportional to probability [1]. This has been pointed out by Gearhart, who says, in commenting on Equation (14): "Thus, in December 1900, Planck cited Boltzmann, especially his 1877 paper, and said that by introducing probability considerations, he could derive an expression for the entropy of a resonator. He argued that 'the entropy of a system of resonators with given energy is proportional to the logarithm of the total number of possible complexions'" [6].

Planck then broke this statement up into two separate theorems:

- Entropy is proportional to the logarithm of the probability of a state;
- The probability of any state is proportional to the number of corresponding complexions. Further, all complexions are equally probable.

The first theorem, Planck said, is no more than a definition. The second he described as the "core of the whole theory", but added that "in the last resort, its proof can only be given empirically".

What is the 1877 Boltzmann paper? Again according to Gearhart: in the 1877 Boltzmann consider a gas of molecules: each molecules could have only a discrete kinetic energy 0, e, $2e$,..., pe and the total kinetic energy of the system is fixed at some multiple of e (as example pe) Boltzmann described his method in the following terms: "I will seek out all combinations that are possible by the distribution of $p + 1$ kinetic energies among N molecules, and then show how many of these combinations correspond to each state distribution". Boltzmann called each combination a "complexion" (today it is usually called a microstate) and assumed that each complexion is equally probable. Boltzmann next showed that in general for a gas of N molecules, the number of complexions corresponding to the state characterized by w_0 molecules with energy 0, w_1 molecules with energy e, and so on is

$$W_p = \frac{N!}{w_0! w_1!, ..., w_p!} \tag{15}$$

(which is the regular multinomial formula) and noted without proof that the total number of complexions J is

$$J = \frac{(N + P - 1)!}{(N - 1)! P!} = \sum_p W_p \tag{16}$$

Boltzmann introduced J to define the probability of a state as the number of complexions for the state divided by the total number of complexions J [6].

As noted also by Darrigol [4], Planck himself fully explained the background of his combinatorics in a paper published in 1906: "Here we can proceed in a way quite analogous to the case of a gas molecules, if only we take into account the following difference: a given state of the system of resonators, instead of determining a unique distribution, allows a great number of distributions, since the number of resonators that carry a given amount of energy is not given in advance (thus the value of p). If we consider now every possible distribution of energy and calculate for each of these the corresponding number of complexions exactly as in the case of gas molecules, through addition of all resulting numbers of complexions we get the desired probability W of a given physical state".

So, in other terms, Planck confirms that the long-sought expression for probability, to insert as an argument of the logarithmic equation for the entropy (Equation (4)), is the combinatorial formula J already suggested by Boltzmann (but never used before) while dealing with the problem of gas molecules. In other words:

$$W = J = \sum_p W_p \tag{17}$$

Yet this expression is neither a probability nor proportional to probability. Rather, it is a total number of microstates (complexions) of the systems (a sort of multiplicity factor). In fact, historians agree that Equation (13) is rather obscure in its derivation. The situation was analyzed in detail a few years later by Darwin and Fowler, who not only clarified the exact meaning of Equation (13) but also (in doing so) introduced concepts that greatly extend the scope of the problem.

In summary, although Planck's main contribution was the courage to propose the quantization of electromagnetic radiation, he also dedicated effort to deriving the blackbody spectrum from the formula for entropy and, more generally, to inserting blackbody theory into the underpinnings of the thermodynamics.

Before moving on from Planck's theory, let us here face another matter that has always concerned Planck's combinatorial expression (Equation (14)) and the treatment of blackbody radiation: the theme of distinguishability and indistinguishability of particles. According to some authors, Planck's

equation differs from Boltzmann's equation because it refers to indistinguishable particles (photons) while Boltzmann refers to molecules and thus distinguishable particles. In 1983, Tersoff and Bayer [13] published a paper titled "Quantum Statistics for Distinguishable Particles" wherein they evaluated the probability $p[n_i]$ of a single configuration of N indistinguishable particles among P discrete states, using the expression:

$$p[n_i] = \frac{(N-1)!P!}{(N+P+1)!} \tag{18}$$

which correctly corresponds exactly to the inversion of Equation (14). They use the expression

$$p[n_i] = \frac{1}{P^N} \frac{N!}{w_0!w_1!..w_P!} \tag{19}$$

in evaluating the same probability for the distinguishable particles. Note that Equation (19) corresponds to Equation (15) normalized by all the combinations P^N.

At Planck's time, the question simply did not arise, because the appearance of three different statistics for photons (Bose–Einstein), electrons (Fermi–Dirac), and molecules (Maxwell–Boltzmann) was not on the scientific scene in 1900 and did not begin to appear until about 20 years later (in the 1930s Leon Brillouin summarized the differences between photons, electrons, and molecules in a book dedicated to the topic [14]). Darrigol has also demonstrated that combinatorial formulas are very often ambivalent, and that they can refer to situations in which there are both distinguishable elements and indistinguishable ones. In particular, they show that the combinatorial expression in Equation (14) can refer both to "P indistinguishable objects over N boxes" or to "N distinguishable objects over energy step e" [4].

On the other hand, Planck himself explicitly cites a model for his expression of a molecular gas. The literature on the subject is extremely vast, but the article by Tersoff and Bayer deserves special mention. In their paper, they write that "Two crucial assumptions are made in deriving [15,16]. First, each distinct configuration is taken to have equal probability. This assumption is (apparently) dictated by symmetry and simplicity. To obtain quantum statistics, one further assumes particle to be indistinguishable. Though contrary to familiar classical thinking, this assumption is accepted because it yields physically correct results for observable quantities". They continue: "Here we show that the latter assumption can be replaced with the more intuitive one of distinguishable particles, if we modify the assumption that all distinct configurations have fixed equal probability weighting". In effect, they then demonstrate that by using an arbitrary probability weighting applied to the Boltzmann formula (Equation (19)), they obtain exactly the Planck formula (Equation (18)).

In summary, we can say that the combinatorial expression used by Planck to obtain his equation was not derived from assumptions of indistinguishability. Furthermore, this expression is still carried forward to a classical multinomial probability expression, as in Equation (15), in which the different configurations (states) are properly "weighted". For a recent treatment of the subject see also Niven and Grendar [17].

3. Darwin, Fowler, and Schrodinger and the Centrality of the Partition Function

In 1922, Charles G. Darwin and Ralph H. Fowler published "On the Partition of Energy. Part I" and "On the Partition of Energy. Part II. Statistical Principles and Thermodynamics" [18,19], clarifying the aspects of combinatorial expressions used by Planck and introducing a fundamental connection between combinatorics and thermodynamics, thereby opening the route for the connection between thermodynamics and information theory. In [18], referring to the results recently obtained by Planck, they said that: "The object of the present paper is to show that these calculations (thus to obtain an expression for the probability of any state described statistically and then to make this probability a maximum) can all be much simplified by examining the average state of the assembly instead of its most probable state. The two methods are actually the same, but whereas the most probable state is only found by the use of the Stirling's formula, the average state can be found rigorously by the help

of the multinomial theorem, together with certain not very difficult theorems in the theory of complex variables. By this process it is possible to evaluate the average energy of any group in the assembly, and hence to deduce the relation of the partition to temperature, without the intermediary of entropy".

In Part II they summarized the advantage of this method in a more explicit way: "the power of our method on the statistical side invites a somewhat more general review of the fundamental connection between classical thermodynamics and statistical mechanics both of classical dynamics and the quantum theory" [19].

In the fourth paragraph of their second paper, Darwin and Fowler directly face Planck's use of entropy and its relation with the probability of a state [19]; thus Equation (4). Quoting Planck, they say, "Next, to evaluate W a definition is made of 'thermodynamics probability' as the number of complexions corresponding to the specific state: this is made a maximum subjected to the condition of constant energy, and the maximum of $k\log W$ is equated to the entropy S, which is then shown by examples to be the entropy of thermodynamics" [18]. They observe that this procedure, followed by Planck, involves two separate processes: "in the first the determination of the maximum fixes the most probable state of the assembly by itself. In the second the assembly is related to the outside world by determining its entropy and then the absolute temperature scale is introduced by the relation $dS/dE = 1/T$."

Further, they comment: "Now there is much to be criticized in this argument. In the first place, there is a good deal of vagueness as to what is happening". With explicit reference to Planck's summary of the extensive properties of entropy given in [12] they continue:

For the addition of entropies can only be realized by some form of thermal contact and is then only in general true when the temperatures are equal; and both these conditions require that the assemblies shall not be independent. So, it is only possible to give a meaning of

$$S_1 + S_2 = S_{12} \tag{20}$$

by making

$$W_1 W_2 = W_{12} \tag{21}$$

invalid. Again, without more definition the probability of a state is quite ambiguous: for example, we can speak of the probability of one particular system having, say, some definite amount of energy, and for independent assembly the (20) will be true of this type of probability, but it will have not relation whatever to entropy. This objection is supposed to be met by the definition of "thermodynamics probability"; but that is a large integer and not a fraction, as are all true probabilities, and so the (20) cannot be maintained simply as a theorem in probability.

Moreover, they also say: "Now it is established that actually the 'thermodynamics probability' does lead to the entropy, and so we must consider how it is to be interpreted in terms of true probability. It is clear that the 'thermodynamics probability' must be divided by the total number of admissible complexions, and that when we consider an assembly of a given energy this number is J".

Hence, Darwin and Fowler are not only concerned about Planck's easy use of probability but they also underline the fact that the word "probability", used in Equation (14) as argument to the logarithm, should be applied as a normalization constant to obtain the probability factor from the equation of complexions. How can this contradiction be overcome? According to Darwin and Fowler this may not be possible as long as we remain in an isolated system. In fact, they clearly state that "As long as we consider the whole assembly this is impossible, for J depends on $\exp(1/kT)$ and cannot be regarded as an ignorable constant when changes of temperature are contemplated" [19]. However, Darwin and Fowler suggest a way to overcome the contradiction. It is connected to the hypothesis that they consider at least two systems in thermal equilibrium with one of the two, which we name B, that can be viewed as a "heat reservoir" because it is larger than the one we are examining and that we call A.

In the first part of [18], Darwin and Fowler define how to evaluate the number of complexions generated by M molecules which occupy $a_1, a_2, \ldots, a_m$ states (characterized by different volumes in

their example). Assuming that each complexion is equally probable results the same expression used by Planck [12] and, earlier, by Boltzmann; thus, the multinomial expression given by Equation (15). Hence, they continue: "Take, for example, a group of M A's systems and suppose them immersed in a bath of a very much larger number of B's. We can now define the entropy of the A's when their specification is a_1, a_2, ... , a_m as k times the logarithm of the probability of that specification. In calculating the probability, we are indifferent about the distribution among the B's, so we sum the complexions involving all values of the b's consistent with the selected values of the a's".

Then

$$W(a_1, a_2, ...a_m) = \frac{\frac{M!}{a_1!a_2!,...,a_m!} \cdot \sum_b \frac{N!}{b_1!b_2!,...,b_n!}}{J} \tag{22}$$

where N are the systems of the reservoir B. Now, provided that N is much larger than M, the factor

$$\frac{\sum_b \frac{N!}{b_1!b_2!,...,b_n!}}{J} \tag{23}$$

will be practically independent of the a's and the energy of the group of A's, it may be taken as constant and omitted from the calculation, and we are left with the 'thermodynamic probability' as the only variable part.

It is only in this sense that a strict meaning can be assigned to Boltzmann hypothesis; and it is of the greatest interest that the conditions under which it has meaning correspond exactly to the conditions of the 'canonical ensemble' of Gibbs, as will be shown later. Even so, it is not a very convenient expression (Equation (21)) for we must always suppose that the assembly is a part of some much larger one, whereas the expression for the entropy is purely a function of the group and the temperature [19].

Hence, they conclude that for practical reasons: "It is more convenient to abandon the use of the principle of probability and to define the entropy as k times the logarithm of the number of complexions. We shall call this the kinetic entropy. This number of complexions has the multiplicative property (20) but now in virtue of its own combinatory formula and not of an appeal to an inapplicable probability theorem" [19].

Then Darwin and Fowler admit that the new definition "does not appear to have the same simplicity as the old, but that is only because in the old the necessity for a detailed definition of what is meant by probability was concealed" [19].

In conclusion, Darwin and Fowler revisit Planck's theory by specifying that it did not cover the concept of "probability" in the narrow sense and that Planck's entropy has a special meaning, which they define as "kinetic entropy".

However, even if we accept Darwin's and Fowler's observations, it is still not clear why Planck used Equation (17) as argument of his entropy function (thus the sum of all the multinomial expressions). Here, Darwin and Fowler are very subtle, because they state that when we define "the entropy as k times the logarithm of the number of complexions" it does not mean much if we consider "the total number of the complexions, or the average number or the maximum number". In fact, "Now if these quantities are calculated, it will be found that, to the approximation (in which the Stirling formula is applicable) they all have the same value. This value is easiest to find for the maximum number" [19].

In fact, for very large values of N (as for thermodynamic phenomena), the maximum of the multiplicity factor gives a number of complexions so large that it is does not differ much from the number of complexions corresponding to the sum of all possible multinomial distributions. Hence the term J used by Planck. This has already been noted by Gearhart, who says that Planck himself offers a clue [6]. In fact, on page 145 [12], Planck writes: "The total number of all possible complexions may be calculated much more readily and directly than the number of complexions referring to the state of equilibrium only" (thus invoking the maximum).

Therefore, it seems that only a mathematical convenience moved Planck to use the sum instead of the maximum of the multinomial distribution, in agreement with the observation made by Darwin and Fowler. In other words, it is possible to concluded that the Equation (13) can be now be better interpreted as:

$$S_{kinetic} = k \ln W_{MNMax} \tag{24}$$

where entropy is the "kinetic entropy" previously defined and W_{MNMax} is the maximum of the multinomial expression connected with the dedicated specification.

Although Equation (24) is a significant contribution on Darwin's and Fowler's part (one that allows us to view Planck's equation in its proper light and connect it with statistical mechanics), their contribution is even more significant because they stressed the importance of the partition function when defining entropy and for statistical mechanics in general. In other words, Darwin and Fowler distance themselves from Planck's method of using the formula for entropy as a means to obtain the wanted energy function; in doing so, he had formulated an unusual combinatory expression (the (14)).

In fact, Darwin and Fowler say, "we are led to a presentation of the entropy which is very closely related to that of classical thermodynamics, which frees it from the combinatory complications with which it is normally associated and brings it back to the direct dependence on the partition function which form the basis of our method [19]". Moreover, in the last part of the paper they introduced a thermodynamics function which is more appropriate than entropy for connecting thermodynamics to statistics, that is, the "characteristic function Y". They say: "In paragraph 8 the definition is considerably simplified mathematically by replacing the 'entropy' by the 'characteristic function' as the basal thermodynamic quantity".

The method proposed by Darwin and Fowler was implemented and considerably improved by Schrodinger in a series of seminars given at the School of Theoretical Physics, Dublin, in January–March 1944 and published in 1948 [15]. Since Schrodinger used more modern language, we will later combine quotes from Schrodinger and quotes from Darwin and Fowler in order to obtain their vision of the thermodynamics functions.

Schrodinger pointed out that in facing the essential problem of statistical thermodynamics, that is, "the distribution of a given amount of energy E over N identical systems", he adopts Gibbs's point of view. In fact, he says that "It has a particular beauty of its own, is applicable quite generally to every physical system, and has some advantages to be mentioned forthwith. Here, the N identical systems are mental copies of the one system under consideration, of the one macroscopic device that is actually erected on our laboratory table". Schrodinger points out that this point of view is preferable to those "older and more naïve" viewpoints associated with Maxwell, Boltzmann, and others, that posited "N actually existing physical systems are in actual physical interaction with each other" [15]. The advantages of Gibbs's point of view are summarized by Schrodinger as follows:

- N can be made arbitrarily large;
- No question about the individuality of the members of the assembly can ever arise, as it does, according to the new statistics, with particles.

In these two methodological details, we can see consistency with all developments held by Planck, who spoke of the "number of evolutions" of the same oscillator and did not have the problem of individuality (which for Schrodinger becomes "merely a question of enumerating correctly the states of the single system" [15]).

After this clarification, Schrodinger introduces the enumeration of a certain class of states of the assembly, 1, 2, . . . , l, each one characterized by a proper energy $e_1, e_2, \ldots, e_l$ with occupation number $a_1, a_2, \ldots, a_l$ (where occupation number means how many of the N systems are in the state 1, 2, . . . , l). Then Schrodinger introduces (without demonstration) this statement:

The number of single states, belonging to this class, is obviously

$$P = \frac{N!}{a_1!a_2!, ..., a_l!...} \tag{25}$$

The set of numbers all must, of course, comply with the conditions

$$\sum_l a_l = N \tag{26}$$

and

$$\sum_l \varepsilon_l a_l = E \tag{27}$$

The statement (of Equations (25)–(27)) really finish our counting. However, in this form the result is wholly unsurveyable [15].

It is easy to identify in Equation (25) the similarity with Equation (15), used by Boltzmann and Planck, in order to evaluate the number of complexions of N molecules with energy e_1, e_2, ... , and so on. This is the classical multinomial factor. Therefore, by definition, the number of ways to split N distinct objects into l distinct groups, of sizes a_1, a_2, ... , a_l, respectively. In other words, P represents a multiplicity factor or number of microstates belonging to the class of states whose total energy is E.

Then Schrodinger evaluates the maximum of the logarithm (natural) of function P with the help of Lagrange multiplier methods. Schrodinger does not explain why he considers the log of P, but this is a fairly common practice when you consider large N. Instead, he makes an important clarification: "The present method admits that, on account of the enormous largeness of the number N, the total number of distributions (i.e., the sum of all P's) is very nearly exhausted by the sum of those P's whose number sets all do not deviate appreciably from that set which gives P its maximum value (among those, of course, which comply with (Equations (26) and (27)). In other words, if we regard this set of occupation numbers as obtaining always, we disregard only a very small fraction of all possible distributions, and this is a vanishing likelihood of ever being realized" [15].

We recognize in this passage the same clarifications pointed out by Darwin and Fowler.

By introducing the two Lagrange multipliers λ and μ, Schrodinger looks for the unconditional maximum of

$$\log P - \lambda \sum_l a_l - \mu \sum_l \varepsilon_l a_l \tag{28}$$

hence

$$\log P = \log\left(\frac{N!}{\prod_l a_l!}\right) = \log N! - \log \prod_l a_l! = \log N! - \sum_l \log a_l! \tag{29}$$

At this point Schrodinger applies the Stirling approximation (that is, $\log(n!) \approx n(\log n - 1) \approx n \log n$ the use of which is also explicitly admitted by Darwin and Fowler [19]). Hence, we have

$$\log N! - \sum_l \log a_l! \Rightarrow N \log N - \sum_l a_l \log a_l \tag{30}$$

In order to evaluate the maximum of Equation (29) we consider Equations (26) and (27) and take the variation of the expression

$$N \log N - \sum_l a_l \log a_l - \lambda \sum_l a_l - \mu \sum_l \varepsilon_l a_l \tag{31}$$

and make it equal to zero. Since N is constant, we obtain

$$-\sum_l \delta a_l \log a_l - \lambda \sum_l \delta a_l - \mu \sum_l \varepsilon_l \delta a_l = 0 \tag{32}$$

The above, solved for each l, gives

$$\log a_l + \lambda + \mu \varepsilon_l = 0 \tag{33}$$

for those

$$a_l = \exp(-\lambda - \mu\varepsilon_l) \tag{34}$$

Hence, at the maximum of $\log P$, the average energy U of the system is

$$U = \frac{E}{N} = \frac{\sum_l \varepsilon_l \exp(-\lambda - \mu\varepsilon_l)}{\sum_l \exp(-\lambda - \mu\varepsilon_l)} = \frac{\sum_l \varepsilon_l \exp(-\mu\varepsilon_l)}{\sum_l \exp(-\mu\varepsilon_l)} = \frac{\partial}{\partial\mu} \log \sum_l \exp(-\mu\varepsilon_l) \tag{35}$$

and the occupation number a_l is

$$a_l = \frac{N}{N} \exp(-\lambda - \mu\varepsilon_l) = \frac{N \exp(-\lambda - \mu\varepsilon_l)}{\sum_l \exp(-\lambda - \mu\varepsilon_l)} = N \frac{\exp(-\mu\varepsilon_l)}{\sum_l \exp(-\mu\varepsilon_l)} \tag{36}$$

Hence the average energy of the system and the occupation number are both obtained according to the original program of Darwin and Fowler. It is interesting to observe that both quantities are functions of the sum

$$Z = \sum_l \exp(-\mu\varepsilon_l) \tag{37}$$

called the "partition function", the "sum over states", or "Zustandssumme", hence Z.

The result appears so significant that Schrodinger says, "The above equations (Equations (35) and (36)) it may be said to contain, in a nutshell, the whole of thermodynamics which hinges entirely on this basic distribution" [15]. The Lagrange multipliers are determined by using the virtual experiment of putting the system in contact with a reservoir, and it is possible to demonstrate that it results in

$$\mu = \frac{1}{kT} \tag{38}$$

while λ can be eliminated. Moreover, we observe that in Equations (35) and (36), function Z is effectively an argument of the logarithm. Hence it is convenient to introduce the "characteristic function Y" (which was previously used, even by Planck, and introduced by *M*. Massieu in the 1869 as a "function caracteristique du corps" [16], but without any connection to the partition function). We define Y as:

$$\Psi = k \log Z \tag{39}$$

By introducing the value of μ given by Equation (38) into Equation (35), we obtain:

$$U = T^2 \frac{d\Psi}{dT} \tag{40}$$

from which we obtain Planck's equation and other important physics equations [18].

In addition, by using the relation:

$$U \frac{dT}{T^2} = -d\left(\frac{U}{T}\right) + \frac{dU}{T} \tag{41}$$

we obtain from Equations (35) and (40) that

$$\frac{U dT}{T^2} = -d\left(\frac{U}{T}\right) + \frac{dU}{T} = d\Psi \tag{42}$$

or

$$dU = T d\left[\frac{U}{T} + \Psi\right] = T dS \tag{43}$$

which authorizes us to express entropy as

$$S = \Psi + \frac{U}{T} \tag{44}$$

This relationship was introduced by Massieu [16] in the context of fluid thermodynamics. Darwin and Fowler underline the centrality of Equation (44), emphasizing that "it agrees completely with the entropy of the thermodynamics in all cases where they (part of the assembly) can be compared: this agreement justifies our use of the (Stirling approximation) in these calculations. However, it is indifferent whether we define the entropy as the total, average, or maximum number of complexions and the (Stirling approximation) is always inexact" [19]. Moreover, Equation (44) gives precisely (italics in original) the thermodynamics expression in all comparable cases, and this suggests a direct definition in terms of partition functions. We may thus suppose that the combinatory processes are correctly looked after by the partition functions, and may define the entropy by either (Equation (44) or (35)). Pending its formal identification with the entropy of thermodynamics, we shall describe it as the "statistical entropy"

$$S_{stat} = \Psi + \frac{U}{T} \tag{45}$$

Hence, Darwin and Fowler describe how to complete the formal identification of "statistical entropy" with "thermodynamic entropy". First at all, they demonstrate that the claimed "increasing properties" of entropy are not enough to allow us to proceed with this identification because any other function of the type

$$\Sigma = S_{stat} + bU \tag{46}$$

(with b a universal constant) holds the same increasing properties; but if we attempt to define T, we obtain

$$\frac{1}{T} = \frac{d\Sigma}{dU} = \frac{dS_{stat}}{dT} + b \tag{47}$$

which can never absolutely determine the value of T. They comment: "This impasse is one aspect of the fact that in thermodynamics the absolute temperature and the entropy are introduced in the same chain of argument, the absolute temperature as integrating factor and the entropy as the resulting of the integral. Thus, and this is a point that has been overlooked by some writers, it is impossible to identify the entropy by using assemblies in which temperature is the only variable, for any (italics in original) function of the temperature is then a possible integrating factor" [19].

As a consequence, they underline the inconsistency of the arguments followed by both Boltzmann and Planck. On the other hand, they emphasize that "the only way of making the identification and that is to evaluate dQ, the element of heat, for an assembly of more than one variable from our statistical principles, and to show that a certain unique function of the temperature is an integrating factor for it" [19]. Hence, they proceed with this demonstration and conclude that statistical and thermodynamic entropy coincide.

Equation (45) is one of the most important contributions of Darwin and Fowler to the statistical foundation of thermodynamics, because it allows a rigorous connection between the statistical description of the assembly, given by the partition function Z, with thermodynamic parameters.

Before concluding, Darwin and Fowler define the importance of the "characteristic function Y" expressed by Equation (39). They give a formal definition of this function: "The characteristic function for any part of an assembly is k times the sum of the logarithms of the partition functions of all the component systems of the part when the argument of the partition function is $\exp(-1/kT)$." Moreover, they show that the characteristic function is the negative free energy F divided by the absolute temperature T

$$\Psi = -\frac{F}{T} = -S + \frac{U}{T} \tag{48}$$

about which they comment, "The characteristic function contains two arbitrary constants, which occur in the form $S_0 - E_0/T$. Of these, E_0 is seen to correspond to the arbitrary zero of energy of the system, which appears in each exponent of the partition function Z. The constant S_0 depends on the absolute values adopted for the weight factors (the weight factor of the exponential terms of the partition function, that was set equal to 1 in Schrodinger's notation). We have made the convention of taking this as a unity for simple quantized systems; but it is only a convention and quite without effect on the various average values, which are all that can ever be observed. Indeed, the only conditions attaching to the weight factor are precisely analogous to those attaching to the entropy in classical thermodynamics, a definite ration is required between the weights of states of systems which can pass from one to the other, but as long as two systems are mutually not convertible into one another, it makes absolutely no difference what choice is made for their relative weight".

Let us remember that while Schrodinger was in Dublin in 1943, he wrote "What is Life?" [20], in which he begun an important reflection around the theme of life. It is one of the first times that a physicist deals with such a matter, so far from his discipline, and Schrodinger observes the contradiction between the need to transmit "information" characteristic of all living beings and the observance of the second principle. Schrodinger asks, "How does the living organism avoid decay?" and answers: "What then is that precious something contained in our food which keeps us from death? That is easily answered. Every process, event, happening call it what you will; in a word, everything that is going on in Nature means an increase of the entropy of the part of the world where it is going on. Thus, a living organism continually increases its entropy or, as you may say, produces positive entropy and thus tends to approach the dangerous state of maximum entropy, which is death. It can only keep aloof from it, i.e., alive, by continually drawing from its environment negative entropy which is something very positive as we shall immediately see. What an organism feeds upon is negative entropy. Or, to put it less paradoxically, the essential thing in metabolism is that the organism succeeds in freeing itself from all the entropy it cannot help producing while alive".

Hence (for the first time, to my knowledge) Schrodinger introduces the term "negative entropy" or, as later used by Brillouin, negentropy. Moreover, he gives a precise definition of negentropy. In fact, he says: "If D is a measure of disorder, its reciprocal, $1/D$, can be regarded as a direct measure of order. Since the logarithm of $1/D$ is just minus the logarithm of D, we can write Boltzmann's equation thus: $-$(entropy) = $k$log$(1/D)$. Hence the awkward expression 'negative entropy' can be replaced by a better one: entropy, taken with the negative sign, is itself a measure of order. Thus, the device by which an organism maintains itself stationary at a fairly high level of orderliness (=fairly low level of entropy) really consists in continually sucking orderliness from its environment".

We will see in the following paragraphs that the same expression will be used by Brillouin and Shannon to define the amount of information.

4. Nyquist, Hartley, and the Dawn of Information

A couple of years after the publication of the paper by Darwin and Fowler, in 1924, Nyquist published a paper titled "Certain Factors Affecting Telegraph Speed" dedicated to exploring the "speed at which intelligence can be transmitted over a telegraph circuit with a given line speed, i.e., a given rate of sending of signal elements" [21]. Telecommunication by telegraph was in full expansion, and engineers had begun to explore means of increasing the efficiency of transmission. Nyquist points out that "by speed of transmission of intelligence is meant the number of characters, representing different letters, figures, etc. which can be transmitted in a given length of time assuming that the circuit transmits a given number of signal elements per unit of time". In Appendix B of his paper, Nyquist introduces the basic ingredient of any further development of information theory: that is, the assumption that information is transmitted by means of a definite number of characters (in the broadest sense) and that each character can be expressed by a combination of a definite set n of a limited number of "current values" m. The ensemble of the characters constitutes a code [21].

Nyquist says: "Let assume a code whose characters are all of the same duration. This is usually the case in printer codes. If n is the number of signal elements per character, then the total number of characters which can be construed equals m^n. In order that two such systems should be equivalent, the total number of characters that can be distinguished should be the same. In other words,

$$m^n = const. \tag{49}$$

This equation may also be written

$$n \log m = const. \tag{50}$$

Next, Nyquist gives the first definition of the capacity of the transmission of the system: "The speed at which intelligence can be transmitted over a circuit is directly proportional to the line speed and inversely proportional to the number of signal elements per character provided that the relations above are satisfied". Hence, we may write

$$\frac{line\,speed}{number\,of\,signal\,elements} = \frac{s}{n} \tag{51}$$

Substituting the value of n derived from the equation above, this equation becomes

$$\frac{s \log m}{const} \tag{52}$$

which may also be written

$$K \log m \tag{53}$$

where K incorporates both the line speed and the constant. Nyquist correctly emphasizes that "it will be noted that the formula has been deduced for codes having characters of uniform duration and that it should not be expected to be anything but an approximation for codes whose characters are of non-uniform duration."

This is the dawn of information theory. We note that this theory was originally based around the concept of transmission capacity rather than the concept of "quantity of information". This point of view also inspired Shannon.

Four years later (in 1928), Hartley focused on the theme of the transmission of intelligence by a telegraph line and summarized his reflections in a paper titled "Transmission of Information" [22]. Hartley substantially resumes Nyquist's argument with a slight change of terminology: he calls "number of selection n" the "signal elements" and instead of the "current values m" he introduces the term "primary symbols s". Through a series of steps equivalent to Nyquist's, he concludes that:

The amount of information H associated with n selection is

$$H = n \log s = \log s^n \tag{54}$$

what we have done then is to take as our practical measure of information the logarithm of the number of possible symbol sequences. The situation is similar to that involved in measuring the transmission loss due to the insertion of a piece of apparatus in a telephone system. The effect of the insertion is to alter in a certain ratio the power delivered to the receiver. This ratio might be taken as a measure of the loss. It is found more convenient, however, to take the logarithm of the power ratio as a measure of the transmission loss [22].

In short, the choice of logarithmic base seems dictated mainly by practicality. Note also that for the first time the symbol H is introduced as measurement of information. Moreover, Hartley continues with some important clarification: "The numerical value of the information will depend upon the system of logarithms used". Hartley uses the example of the Baudot service (where s = 2 and one

character involves the selection of five primary symbols) to conclude that the "information content of a Baudot character is 5log2". Moreover, he writes, "The same result is obtained if we regard a character as a secondary symbol and take the logarithm of the number of these symbols, that is log 2^5, or 5log2. The information associated with 100 characters will be 500log2."

Next, Hartley extended this consideration to "other forms of communication" like speech. In doing so "certain generalizations need to be made". He continues: "The actual physical embodiment of the word consists of an acoustic or electrical disturbance which may be expressed as a magnitude time function. We have then to examine the ability of such a continuous function to convey information. Obviously over any given time interval the magnitude may vary in accordance with an infinite number of such functions. This would mean an infinite number of possible secondary symbols, and hence an infinite amount of information". Here, he addresses for the first time the problem of the continuous function, and proposes solving it by introducing a certain amount of "discretization", realized by stepping the speech curve as illustrated by an apposite figure.

An imperfectly defined curve may then be thought of as one in which the interval between the steps is finite. The steps then represent primary selections. The number of selections in a finite time is finite. Also, the change made at each step is to be thought of as limited to one of a finite number of values. This means that the number of available symbols is kept finite [22].

A similar visualization and similar concepts were also considered later in a paper by Tuller, "Theoretical Limitations on the Rate of Transmission of Information" published in 1949 [23]. (Tuller's paper was sent to the publisher, however, in September 1948, just in between the appearance of the two famous papers by Shannon [24,25].) Tuller's drawing is plotted inside a regular frame and Tuller explicitly says that "The information function may thus be redrawn so as to follow only certain lines in a rectangular coordinate system. Such a function is called quantized, since it takes on values chosen from a discrete set. A plot of such a function quantized, and drawn in n,s space is also given". He continues, "The question now before us is: "What is the information content of a function in the n,s plane?" Hartley's answer is the "quantity of information" given by $H = kn \log s$ where k is a proportionality constant".

Then Tuller underlines the reasons behind Hartley's choice:

(a) Information must increase linearly with time. In other words, a two-minute message will in general contain twice as much information as a one-minute message;
(b) Information is independent of s and n if s^n is held constant.

Tuller concludes, "On the basis of these two requirements, it can be shown that the Hartley definition is the only possible definition of quantity of information."

It appears from the above that since the initial suggestion proposal by Nyquist, the concept of information has taken shape together with the concept of the capacity of transmitting it. Moreover, the introduction of the continuous function was overcome by Tuller by introducing some reasonable hypothesis of quantization. All the above occurred in an engineering context where the focus was the improvement of telecommunication systems (even then in continuous need of bandwidth).

5. Szilard, Brillouin, and Beyond; Physicists Discover Information

While in 1922 Darwin and Fowler's paper was published in parallel to the interest demonstrated by engineers like Nyquist and Hartley, Szilard, a Hungarian student, migrated to Berlin to take his doctoral thesis on the "Manifestation of Thermodynamics Fluctuations," in which he discussed the long standing puzzle known as the Maxwell demon (Maxwell's Demon has generated interest and response from its inception to now. Recent work on the concept includes two important volumes edited by Leff and Rex (*Maxwell's Demon Entropy, Information, Computing* (Princeton Univ. Press 1990), and *Maxwell's Demon 2: Entropy, Classical and Quantum Information Computing* (CRC Press 2003), collections of works by Landaurer, Bennet and others). In 1927, he summarized his studies on the same subject by preparing a paper (published only in 1929) titled "On the Decrease of Entropy in a Thermodynamic

System by the Intervention of Intelligent Beings" [26]. This paper represents a tremendous step in the growth of awareness of the importance of information in physics. Szilard himself would not continue with the subject matter (Szilard was a very eclectic scientist, as can be deduced from his biography), but his paper has paved the way for many reflections that reverberate throughout the branch of physics dealing with information, including the physics of computing and/or switching physics.

Szilard enters the subject by imposing two important approximations. First, he considers a thermodynamic system reduced to just to one (single) molecule. Second, he divides the experiment into two periods: the period of the measurement and the period of utilization of the measurement. At the beginning of his paper he says that "the objective of the investigation is to find the conditions which apparently allow the construction of a perpetual motion machine of the second kind, if one permits an intelligent being to intervene in a thermodynamics system." Again, "We show that it is a sort of memory faculty, manifested by a system where measurement occurs, that might cause a permanent decrease of entropy and thus a violation of the second law of thermodynamics, were it not for the fact that the measurements themselves are necessarily accompanied by a production of entropy" [26].

Hence, for the first time, an explicit declaration is given that a "measurement," that is, a sort of information processing, is connected to some change of entropy. Szilard considers a Maxwell-like experiment where the "demon" controls the movement of a piston which divides in two equal parts $V1 = V2 = V/2$ the total volume V. (Let us consider in the following a simplified Szilard's scheme, as proposed by Bennet [27], one much easier to follow that Szilard's original). The volume contains just one molecule. Szilard introduces a "memory device" (a sort of relay) which decouples the measurement of a fluctuating parameter x (like the position of the molecule) from the results of the measurement, y. The memory device makes the presence of the "demon" inessential and the experiment more feasible. Szilard says, "It appears that the ignorance of the biological phenomena need not prevent us from understanding that which seems to us to be the essential thing. We may be sure that intelligent living being can be replaced by non-living devices whose "biological phenomena" one could follow and determine whether in fact a compensation of the entropy decrease takes place as a result of the intervention by such a device in a system" [26].

Hence, Szilard considers a two-step process:

(1) The period of measurement when the piston has just been inserted in the middle of the cylinder and the molecule is trapped either in the upper or lower parts, so that we choose the origin of coordinate x appropriately (choice 1 or choice 2 depending on the position of the molecule) and associate x to the parameter of the piston, y.

(2) The period of utilization of the measurement, that is, "the period of decrease of entropy" (of the reservoir) during which the piston is moving up or down "according to the value of y" (up if the molecules are on the lower part, down if the molecules are on the upper part). During this period "the molecule must bounce on the piston and transmits energy to it" [26].

In the scheme proposed by Bennet, the molecule is present in a vessel of volume V divided by a mobile partition into two equal volumes $V1$ and $V2$. The opposite walls of the vessel are the faces of the corresponding pistons 1 and 2. The mobile partition can be inserted and extracted without work. When the demon sees the molecule, he notices where it is located (that is, whether in $V1$ or $V2$), registers the position (choice 1 or 2), and utilizes the measurement by acting on a switch which enables the opposite piston (2 or 1) to reach the partition. This is done without work because the volume not occupied by the molecule is empty and the piston expands without requiring work. Then the partition is removed, allowing the molecule to expand to the final volume V. Finally, the original position is reached, the register is erased, and the cycle begins again [27].

Since the molecule is in thermal equilibrium at the average energy kT, the pressure p exercised on the piston will be

$$p = \frac{\langle E \rangle}{V} = \frac{kT}{V} \tag{55}$$

and it will produce work per unit of volume of the order of

$$W = -\int_{1/2}^{1} pdV = -\int_{1/2}^{1} \frac{kT}{V} dV = -[kT \ln V]_{1/2}^{1} = kT \ln 2 \tag{56}$$

This work is produced at the expense of a net entropy increase for the machine (or decrease for the reservoir) of the order of

$$\begin{aligned} \Delta S1 &= k \ln \tfrac{V1+V2}{V1} \quad for\, the\, choice\, 1 \\ \Delta S2 &= k \ln \tfrac{V1+V2}{V2} \quad for\, the\, choice\, 2 \end{aligned} \tag{57}$$

Since the probability of the choices is

$$\begin{aligned} choice\, 1 = w_1 &= \tfrac{V1}{V1+V2} \\ choice\, 2 = w_2 &= \tfrac{V2}{V1+V2} \end{aligned} \tag{58}$$

and by remembering that $V1 = V2 = V/2$, the total production of entropy results in

$$\Delta S_{total} = 2\left(\frac{1}{2} k \ln 2\right) = k \ln 2 \tag{59}$$

Therefore, a perpetual motor of the second kind was invented (in effect, the demon's intervention produces an unlimited expansion in the Carnot cycle, without the compression stage). To avoid this contradiction, Szilard's "Ansatz" (the original term used in his paper) was intended to consider "the period of utilization of the measurement the period of decrease the entropy" a quantity expressed as:

$$\Delta S_I = \frac{dQ}{T} = \frac{-dW}{T} = -\frac{kT \ln 2}{T} = -k \ln 2 \quad information\, introduced\, in\, the\, system \tag{60}$$

So, the entropy produced by the measurement should be, at least equal to $-k\log 2$ and the corresponding energy introduced by the measurement into the system should be at least equal to $-kT\log 2$ [27].

Szilard established the minimum cost for obtaining a bit of information; gave the expression of the information introduced in the system (with the same meaning used later on by Shannon); and pointed out that the entropy introduced by the measurement process must have the opposite sign of the entropy of the process.

Later, in 1953, Brillouin, in one of many papers dedicated to the argument, "The Negentropy Principle of Information," summarized these results in what he called the "generalization of the second principle of thermodynamics" [28]. He proposed considering the total entropy of a system equal to the sum of the entropy of the process plus an additive term, due to the introduction of the information in the process. This additive term would have a negative sign and so be equivalent to negentropy (a term he recognized was borrowed by Schrodinger [19]: $\Delta S = (\Delta S_{tot} + \Delta S_I) \geq 0$

In the same paper, Brillouin lists some points which summarized his long investigation of the subject:

(a) Information can be changed in negentropy and vice versa;

(b) Any experiment by which information is obtained about a physical system corresponds, on average, to an increase of entropy in the system or in its surroundings. This average increase is always larger than (or equal to) the amount of information obtained. In other words, information must always be paid for in negentropy, the price paid being larger than (or equal to) the amount of information received. Correspondingly, when the information is lost, the entropy of the system is increased;

(c) The smallest possible amount of negentropy required in an observation is of the order of k. A more detailed discussion gives the value $k$ln2 (0.69 k, the minimum of information) as the exact limit, a result which concurs with our previous discussion of Szilard. In binary digits, this minimum represents just one bit;

(d) These remarks lead to an explanation of the problem of Maxwell's demon, which represents a device changing negentropy into information and back into negentropy;

(e) When a communication channel is considered, the information formula used by Shannon is obtained. The average information per signal is $I = -k\sum_i p_i \ln p_i$.

In effect, Brillouin's critique of Maxwell's demon contained an observation that goes right into the analysis we are doing. In his 1951 paper "Maxwell's Demon cannot Operate: Information and Entropy," after mentioning Szilard's work and the above comments, he says: "In order to select the fast molecules, the demon should be able to see them; but he is an enclosure equilibrium at constant temperature, where the radiation must be that of black body, and it is impossible to see anything in the interior of a black body", and "No wonder Maxwell did not think of including radiation in the system in equilibrium at temperature T. Blackbody radiation was hardly known in 1871, and it took 30 more years before the thermodynamics of radiation was clearly understood and Planck's theory developed" [29]. This was just one criticism of the "Maxwell's demon" literature, the "edifice soundness" of which was considered "very questionable" even by Popper (see a complete review in the quoted volumes by Leff and Rex referenced in endnote 2).

Brillouin's contribution to the foundation of information theory, and its relation to physics and the measurement processes, was collected in an important book published in 1956 titled "Science and Information Theory". In it, he detailed many aspects of the subject such as the relationship between "observation and information", the themes of "information theory, the uncertainty principle and physical limits of observation" and applications to "writing, printing and reading", "information and computing", and "information and organization" [30]. It should be noted here explicitly that Brillouin's main contribution to the physical understanding of information takes place after Shannon's theory (which we explore in the next paragraphs). Brillouin adds nothing to Shannon's theory, a fact which he fully recognizes on more than one occasion; but he tries to reconcile that theory to a more physical and thermodynamic view of the problem (which Shannon will never do).

The explanation of Szilard's experiment and the consequent hypothesis of numeric equivalence between the content of information and negentropy made by Brillouin certainly helps us better understand the physics of information. However, it imposes an important premium (little discussed after Brillouin) on the weight that information plays in physical processes (and even more in organizational and "human" processes in the broad sense).

In fact, in his 1954 paper "Negentropy and Information in Telecommunications, Writing and Reading," in his final paragraph, "Information and Organization," Brillouin tried to evaluate the content of information held by a "blueprint," by evaluating the entropy related to the number of possible "connections" contained in the "blueprint" itself. Since the Boltzmann constant is very small, even if these connections are numerous, the negentropy content of the blueprintis is "completely negligible when compared with the total entropy of the machine. This example exhibits some characteristic features of the whole theory". He concludes: "The negentropy principle of information requires that every piece of information be connected with a corresponding negative term in the physical entropy of a system. This connection is absolutely needed for consistency and was proved by the discussion of the preceding paper (reference to the paper is given). Practically, these negative terms can be ignored in most problems because of the small coefficient (10^{-23} in SI system) introduced by the change from binary digit (bits) to thermodynamical units. The smallness of these terms is the fundamental reason why transmission of information by any practical method: writing, printing, telecommunications, is so inexpensive in entropy units, which also means inexpensive in dollar units.

Modern life is based on these facts and would be completely different in a world where the negentropy of information would have a larger value" [31].

If all this was true in 1954, how much more true must it seem today, in the age of the Internet, where every human process is carried out with the help of amounts of information not even imaginable in Brillouin's times. However, there are other, more subtle connections between information theory and thermodynamics, which is what we will explore in the next paragraphs.

The papers and analysis of Brillouin do not completely exorcise Maxwell's demon. A few years later, in 1961, Landaurer published a paper, "Irreversibility and Heat Generation in the Computing Process" [32] which begins with Brillouin's conclusions. Information theory began to produce first results and the computing machines, the "computers", began to become a reality. The demand for the cost of a switching operation was no longer just an abstraction but could have important technological implications.

Landaurer frames this question into a more general problem of the irreversibility of the computing machines. At the beginning of his article, he poses the central question: "The search for faster and more compact computing circuits leads directly to the questions: What are the ultimate physical limitations on the progress in this direction?" He continues: "The simplest way of anticipating our conclusion is to note that a binary device must have at least one degree of freedom associated with the information. Classically a degree of freedom is associated with kT of thermal energy. Any switching signals passing between devices must therefore have this much energy to override the noise. This argument does not make it clear that the signal energy must actually be dissipated. An alternative way of anticipating our conclusions is to refer to the arguments by Brillouin and earlier authors, as summarized by Brillouin in his book [30]. To the effect that the measurement process requires a dissipation of the order of kT".

Hence, Landaurer seeks to make "refinement of the line of thought" and introduces a machine with N binary elements. After various considerations, he notes that "the entropy can increase by $kNln2$ as the initial information becomes thermalized". As pointed out years later by Bennet, another great protagonist of the theory of computing and switching: "The measurement itself (again referring to the Szilard experiment) can be performed reversibly, but an unavoidable entropy increase, which prevents the demon form the violation of the second Law, occurs when the demon erases the result of one measurement to make room for the next" [33]. In other terms, the modern interpretation of the Szilard experiment focuses not on the measurement or storage operations (as Szilard did) but on the erasure operation. With these words of Bennet, we end our excursion into the contemporary frontiers of information thermodynamics and return to the original journey.

6. Shannon and the Importance of Channel Capacity

At the beginning of the first of his two fundamental papers of 1949 [24,25], "A Mathematical Theory of Communication," Shannon explicitly discusses the "logarithmic choice" in measuring information and tries to generalize the concept. Shannon says that "The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point." Hence, "if the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed by Hartley, the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will have in all cases an essential logarithmic measure" [24].

Shannon also says, the logarithm measure is more convenient for various reasons:

1. It is practically more useful. Parameters of engineering importance such as time, bandwidth, number of relay tend to vary linearly with the logarithm of the number of possibilities: adding one relay doubles the number of possible states of the relays, doubling the time roughly squares the number of possible messages;

2. It is nearer to our intuitive feeling as the proper measure ... one feels, for example, that two punched cards should have twice the capacity of one for information storage and two identical channels twice the capacity of one for transmitting information;
3. It is mathematically more suitable. Many limiting operations are simple in terms of the logarithm.

The choice of a logarithmic base corresponds to the choice of a unit for measuring the information. If the base 2 is used, the resulting units may be called binary digits or more briefly bits, a word suggested by Tukey. A device with two stable positions, such as a relay or a flip-flop circuit, can store one bit of information [24].

Neither Nyquist, Hartley, nor Shannon appear to arrive to the logarithmic expression starting from thermodynamics or from the concept of entropy. It seems more likely that they reached the logarithm measure for practical reasons related to the fact that two Baudot punched tapes or two relays carry double the information of one. Shannon, however, is the only one who does not consider the choice of the logarithm notation obvious. Moreover, Shannon uses, in a more abstract way, the terminology employed by Nyquist and Hartley (who had picked it up directly from telegraph technology). Where Nyquist would speak of "intelligence transmission" Shannon refers to "sent messages". Where Hartley would write about "selection of the event", Shannon speaks of "choice and uncertainty": "Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome?" [24].

After clarifying the use of the log function, Shannon deals with the problem of capacity: "The question we now consider is how one can measure the capacity of such a channel to transmit information" [25]. Shannon had previously offered the famous diagram of a general communication system where the main components are drawn: the transmitter "which operates on the message in some way to produce a signal suitable for transmission over the channel"; the channel, which "is merely the medium used to transmit the signal from transmitter to receiver. It may be a pair of wires, a coaxial cable, a band of radio frequencies, a beam of light, etc."; and the receiver, which "ordinarily performs the inverse operation of that done by the transmitter, reconstructing the message form the signal" [24]. The answer to the problem of capacity led Shannon, three pages later, to the formulation of his First Theorem (thereby confirming the centrality of the capacity concept in building Shannon's information theory). He says: "In the teletype case where the symbols are of the same duration, and any sequence of the 32 symbols is allowed, the answer is easy. Each symbol represents five bits of information. If the system transmits n symbols per second it is natural to say that the channel has a capacity of $5n$ bits per second." Here, he reaches the same conclusion previously reached by Hartley and Nyquist.

However, Shannon now adds: "In the more general case with different lengths of symbols and constraints on the allowed sequences", we make the following definition:

Definition: The capacity C of a discrete channel is given by

$$C = \lim_{T \to \infty} \frac{\log N(T)}{T} \tag{61}$$

where $N(T)$ is the number of allowed signals of duration T [24].

Before proceeding any further, it is useful to note that Shannon introduced two specifications of the allowed signals: lengths and constraints. On the previous page, he had mentioned telegraphy as an example of symbols with different lengths and constraints. In a modern way, we say that the alphabet (the collection of different symbols) can be standard or non-standard: standard means all symbols have the same duration, like the teletype example. Non-standard means that different symbols have different durations, like telegraphy, where we have an alphabet composed of four characters: dot d, Dash D, space s and word S. The duration of each letter is respectively: 2, 4, 3, and 6 units of time t). Moreover, the language of the allowed signals can be free or restricted: in telegraphy, the restrictive

rule is that no spaces follow each other. Hence, after a space, the state of the system changes and only dot or dash can be transmitted. Shannon illustrated with a dedicated sketch.

So, we note that the two specifications introduced by Shannon significantly complicate the situation illustrated by Nyquist and Hartley, while considerably extending the application of the concept of capacity.

In the case of a standard alphabet and free language, $N(t)$ (the number of allowed signals) will be the combination of all q symbols constituting the alphabet in time t, or

$$N(t) = q^t \tag{62}$$

Hence

$$C = \lim_{t\to\infty} \frac{\log q^t}{t} = \log q \tag{63}$$

and we obtain the same results advanced by Nyquist, Equation (53). Before proceeding with Shannon's First Theorem (which treats the general case of a non-standard alphabet and non-free language and results, and is, according to many authors, for example, Khandekar, Mc Elice, and Rodemich [34] "brief and in places quite cryptic"), let us consider the intermediate case of a non-standard alphabet and free language.

A similar problem was implicitly faced by Hartley in 1928, when he dealt with limitations on the transmission rate due to intersymbol interference. Hartley said that "the form of the transmitted wave may be altered due to the storage of energy in reactive elements such as inductances and capacities, and in subsequent release". This phenomenon causes "a disturbance which is the resultant of the effects of all the other symbols as prolonged by the storage of energy in the system" [22]. In other words, Hartley tried to face, for the first time, the physical problem caused by an imperfect channel which still presents some characteristic time that depends on its impedance. This limits the transmission of "impulsive signals" even with standard alphabet, which Hartley noticed. He continued: "Obviously the magnitude of the intersymbol interference which affects any symbol depends on the particular sequence of symbols which precedes it". Hartley illustrated the situation in a figure, and emphasized that "no attempt will be made to obtain a complete or rigorous solution of the problem". He suggested the following method (the original notation is maintained):

If the current that disturbs the symbol s is

$$i = (s-1)\frac{E}{R}\exp(-\alpha\tau) \tag{64}$$

where E is the electromotive force, R is the resistance of the circuit, α is the damping constant $(1/RC)$, t is the symbol's interval (inverse of the symbol rate) and q is the number of selections that precede the disturbed symbol, we have that the interference current will be

$$i_{\text{int}} = (s-1)\frac{E}{R}\sum_{q=1}^{\infty}\exp(-q\alpha\tau) = (s-1)\frac{E}{R}\frac{1}{\exp(-\alpha\tau)-1} \tag{65}$$

This interference current becomes harmful when it equals the primary symbol current, thus E/R

$$\frac{E}{R} = (s-1)\frac{E}{R}\frac{1}{e^{\alpha\tau}-1} \tag{66}$$

hence, by simplifying by E/R we find the minimum allowable value of the symbol rate $\tau_{permissible}$

$$e^{\alpha\tau} - 1 = s - 1 \Rightarrow \tau_{permissible} = \frac{\log s}{\alpha} \tag{67}$$

so the maximum number n of allowed signals transmitted in the time T will be

$$n = \frac{T}{\tau_{permissible}} = \frac{\alpha T}{\log s} \tag{68}$$

That is, this quantity depends on the characteristic of the channel, here represented by the circuit constant a

$$\alpha = \frac{n \log s}{T} \tag{69}$$

Hartley concludes: "Here the numerator is, in accordance with our measure of information, the amount of information contained in n selection, so the left-hand member is the information per unit time or the rate of communication". In the example: "This is equal to the damping constant of the circuit" [22]. The results obtained by Hartley represent a physical limit of the channel capacity. Yet if we generalize the above by considering a very long time and the log base 2, we obtain the logical limit given by Shannon, thus $C = \alpha$. It is interesting to observe that this result is obtained when the sum of the interference currents become equal to the symbol current, thus for:

$$\sum_{q=1}^{\infty} \exp(-q\alpha\tau) = 1 \tag{70}$$

The left-hand member of this expression is no different from the previously mentioned "partition function Z", which leads us to identify capacity as the real root of the equation:

$$Z(\alpha) = 1 \tag{71}$$

By applying Equation (71) to the previously mentioned telegraph alphabet, we can express the partition function Z as

$$\begin{aligned} Z(\alpha) = \sum_{q} \exp(-q\alpha\tau) = \exp(-2\alpha\tau) + \exp(-3\alpha\tau) + \\ + \exp(-4\alpha\tau) + \exp(-6\alpha\tau) \end{aligned} \tag{72}$$

and by solving for the variable α in Equation (72), we obtain the capacity α = 0.41 nat/unit of time [34] (where nat means natural unit of information and refers to the use of a natural logarithm). This nat/unit of time is equivalent to 0.59 bit/unit of time if we use the base 2 logarithm.

In brief, we can now enunciate Shannon's First Theorem for a free language (standard and non-standard alphabet). The combinatorial capacity of a free language is given by

$$C = \alpha_0 \tag{73}$$

where α_0 is the unique solution of the equation $Z(\alpha)$ = 1, where Z is the partition function evaluated on the alphabet of the language.

We emphasize that Shannon's definition of capacity refers not to the physical channel, as does Hartley's, but to the "language", returning to its original meaning as given in Equation (61). This represents the radical innovation introduced by Shannon. It is also emphasized at the end of the first part of Shannon's paper, where he writes: "If a source can produce only a particular message its entropy is zero, and no channel is required. For example, a computing machine set up to calculate successive digits of p produces a definite sequence with no chance element. No channel is required to "transmit" this to another point. One could construct a second machine to compute the same sequence at the point. However, this may be impractical. In such a case, we can choose to ignore some or all of the statistical knowledge we have of the source" [20]. "If the telegraph alphabet were standard, with all the characters having the same length, the capacity would be given by Equation (63). We also emphasize that the above capacity was defined [34] as "combinatorial" to underline its derivation

from Equation (72). A further reduction of capacity is observed by considering a non-free language (a language which presents some syntactic rules). In the example given by Shannon for telegraph language, there are "not spaces (which) follow each other". Evaluating capacity in this case requires the introduction of word rules for the beginning (state i) and the ending (state j)". Hence, we write the partition function Z_{ij} for the words allowable between the two states:

$$\begin{aligned} Z_{11} &= 0 \\ Z_{12} &= e^{-2\alpha\tau} + e^{-4\alpha\tau} \\ Z_{21} &= e^{-3\alpha\tau} + e^{-6\alpha\tau} \\ Z_{22} &= e^{-2\alpha\tau} + e^{-4\alpha\tau} \end{aligned} \tag{74}$$

It is now convenient to apply the generalization of Equation (74) in the form of a "partition matrix" to obtain:

$$Z(\alpha) = 1 \Rightarrow \begin{bmatrix} 0 & e^{-2\alpha\tau} + e^{-4\alpha\tau} \\ e^{-3\alpha\tau} + e^{-6\alpha\tau} & e^{-2\alpha\tau} + e^{-4\alpha\tau} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{75}$$

which finally leads to Shannon's original formulation of the First Theorem for any kind of language (Shannon 's original formulation of the theorem nonetheless differs, because he used a definition based on the function logarithm and a generic function W instead of the exponential. When the base of the logarithm and the exponential are the same, the definition is Equation (75)). If Equation (75) is numerically evaluated, a capacity for the "constraint" language results of about 0.538 bits/unit of time [34], less than that of the free language.

After his theorem on the capacity of a channel, Shannon introduced a theorem about the "transmitter", which he regards as the means of measuring the amount of information produced by a source. He asks, "Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process (an ergodic process) or better, at what rate information is produced?" [24]. He considers a set of possible events whose probability of occurrence are $p_1, p_2, \ldots, p_n$ and asks: "Can we find a measure of how much 'choice' is involved in the selection of the event or how uncertain we are of the outcome?" If there is such measure, say $H(p_1, p_2, \ldots, p_n)$ it is reasonable to require of it the following properties:

1. H should be continuous in p_i.
2. If all values of p_i are equal, $p_i = 1/n$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.
3. If a choice is broken down into two successive choices, the original H should be the weighted sum of the individual values of H.

After these forewords, he introduces the Second Theorem: The only H satisfying the three above assumptions is of the form

$$H = -K\sum_{i=1}^{n} p_i \log p_i \tag{76}$$

Shannon points out that the derivation of this theorem was given in Appendix II of his paper, and he comments: "This theorem, and assumptions required for its proof, are in no way necessary for the present theory. It is given chiefly to lend a certain plausibility to some of our later definitions. The real justification of these definitions, however, will reside in their applications." Moreover, Shannon recognizes that the function H, as in Equation (77), has already been introduced in statistical mechanics. He says: "Quantities of the form $H = -K\sum_{i=1}^{n} p_i \log p_i$ (the constant K merely amounts to a choice of a unit of measure) play a central role in information theory as measure of information, choice and uncertainty. The form H will be recognized as that of entropy as defined in certain formulation of statistical mechanics (here, Shannon quotes in the footnote the Tolman's book "Principle of Statistical Mechanics" [35], published in 1938) where p_i is the probability of a system being in the cell I of its phase

space. H is then, for example, the H in Boltzmann's famous theorem. We shall call $H = -K\sum_{i=1}^{n} p_i \log p_i$ the entropy of the set of probabilities $(p_1, p_2, \ldots, p_n)$."

As Shannon said, Equation (76) was well established in statistical mechanics and its standard derivation as well. It is obtained by searching for the maximum of the combinatorial expression of the entropy which presents as argument of the log the multinomial combinatorial formula as in Equation (15) or Equation (25) (eventually normalized, as in Equation (19)). This was also given in Tolman's book on thermodynamics [35], which Shannon itself quoted. Nevertheless, Shannon's Appendix II does not use this approach to obtain Equation (76). Instead, he accepts the original definition of information given by Nyquist and Hartley in Equation (54):

$$H = K \log m \tag{77}$$

Next, he systematically applies the extensive property of H by considering "log$m$" as an average value, that is:

$$H = K \log m = K\langle \log m \rangle = -K\sum_i p_i \log p_i \tag{78}$$

where p_i is the probability of occurrence of the "primary symbols" m (the same expression is also reported by Grandy [36] and Hamming [37]).

We do not know why Shannon did not derive Equation (77) by following classical thermodynamics methods. It seems that he preferred to draw on Szilard [26], who presents a similar equation with the same probabilistic derivation.

Once channel information capacity and the quantity of information produced by a source are defined, it is reasonable to face the most relevant engineering problem, that is: how much information is possible to send through a channel with reasonable reliability? The answer to this question is not obvious and represents one of Shannon's most relevant contributions. Shannon's Theorem 9 has been called "the fundamental theorem for a noiseless channel."

We will now justify our interpretation of H as the rate of generating information by proving that H determines the channel capacity required with most efficient coding. Theorem 9: Let a source have entropy H (bits per symbol) and a channel have a capacity C (bits per second). Then it is possible to encode the output of the source in such a way as to transmit at the average rate $(C/H - e)$ symbols per second over the channel where e is arbitrarily small. It is not possible to transmit at an average rate greater than C/H [24].

In this theorem, Shannon used the concepts of "coding" and "efficient coding". Coding is a widely used concept in telecommunication operations, defined as a rule which assigns to each source symbol exactly one word (thus a finite sequence of symbols) made by the code symbols. Binary code is an example, and was the reference code for Shannon. A particular kind of coding is said to be unique and decodable when two arbitrary and distinct source messages have distinct word codes [26]. Unique decodable code is rare, because often some code words are parts of others. A code is defined as instantaneous provided that no code word is a prefix of another code word (an instantaneous code is also a unique decodable one). Suppose now that the alphabet of the source (the alphabet is the collection of $a_1, a_2, \ldots a_n$ symbols originated by the source) has symbols that occur with different probabilities $p(a_1), p(a_2), \ldots, p(a_n)$. We define the average length of the code words according to the parameter:

$$\langle L \rangle = \sum_{i=1}^{n} p(a_i) d_i \tag{79}$$

where d_i is the length (number of symbols) of the code word (in binary code, the number of zeros or ones). An efficient code is one which has the smallest average length. For a binary instantaneous code,

it is possible to demonstrate (see Reference [37]) that a relation between the average length of the code words and entropy H of the alphabet source exists. It is:

$$\langle L \rangle \geq H_{source} \tag{80}$$

It can happen that the symbols originated by the source present a different probability. In this case, the coding process helps to match the symbol rate of the source with the bit rate of the channel. The coding consists of setting a codeword to a group of source symbols with the intention of allocating the shortest code word to the most probable group of source symbols. This process is also called "data compression" or the "extension" of the source. So we have: "second order extension" when two source symbols are taken, "third order extension" when three source symbols are taken, and so on.

Coding must be instantaneous and with the shortest possible length (so-called "Huffman coding"). The question is: how much further can we compress an original message? The answer is: we will never compress data below the entropy of the source. Moreover, data compression is ruled by the following important property. Let us call S^k the k-th extension of the information source S:

$$H\left(S^k\right) = kH_{source} \tag{81}$$

We can now better understand Shannon's noiseless coding theorem. In fact, because of Equations (81) and (82), we know that for the average-length word of the k-th extension the following property will hold:

$$\lim_{k\to\infty} \frac{\langle L(S^k) \rangle}{k} = H_{source} \tag{82}$$

Thus, the compressed language can be arbitrarily chosen as close as possible to the entropy of the source language.

Since the compressed language can be interpreted as the "channel", (as defined in Equation (56)) the relation above satisfies Shannon's Theorem 9. It is interesting to read Shannon's comments about his theorem: "In order to obtain the maximum power transfer from a generator to a load a transformer must in general be introduced so that the generator is seen from the load as the load resistance. The situation here is roughly analogous. The transducer (pages before Shannon defined "transducer" the operations performed by the transmitter and receiver in encoding and decoding the information) which does the encoding should match the source to the channel in a statistical sense. The source as seen from the channel through the transducer should have the same statistical structure as the source which maximizes the entropy in the channel. The content of Theorem 9 is that, although an exact match is not in general possible, we can approximate it as closely as desired. The ratio of the actual rate of transmission to the capacity C may be called the efficiency of the coding system. This is of course equal to the ratio of the actual entropy of the channel symbols to the maximum possible entropy" [24].

7. Jaynes or Synthesis

We have seen in the previous paragraphs that the concept of entropy was part of the origin of Planck's equation. Moreover, as admitted (but not adopted) by Shannon, the combinatorial expression of entropy was also at the basis of his Second Theorem. Moreover, we have also seen that the concept of entropy was subject to different interpretations and we have encountered at least three definitions of entropy: Planck's "kinetic entropy" (Equation (4)); Darwin's and Fowler's "statistical entropy" (Equation (45)); and the "probabilistic entropy" implicitly introduced by Shannon in deriving Equation (78).

After Shannon's work, many scientists contributed important reflections on the connection between entropy and information theory, as mentioned in the first part of this paper. However, the author that made the largest generalization of the entropy concept was probably Jaynes. Among his publications on the subject, we particularly note his 1957 paper "Information Theory and Statistical

Mechanics" [38] and a second paper published in 1959, "Note on Unique Decipherability" [39]. Important contributions in the same line of thought were also given by Grandy, as already discussed [36].

In Reference [38], Jaynes faces the following problem: The quantity x is capable of assuming the discrete values x_i ($i = 1, 2, \ldots, n$). We are not given the corresponding probabilities p_i; all we know is the expectation value of the function $f(x)$:

$$\langle f(x) \rangle = \sum_{i=1}^{n} p_i f(x_i) \tag{83}$$

On the basis of this information, and, of course, of the normalization condition

$$\sum_i p_i = 1 \tag{84}$$

what is the expectation value of the function $g(x)$? Jaynes comments: The great advantage provided by information theory lies in the discovery that there is a unique, unambiguous criterion for the "amount of uncertainty" represented by a discrete probability distribution, which agrees with our intuitive notions that a broad distribution represents more uncertainty than does a sharply peaked one, and satisfies all other conditions which make it reasonable the quantity which is positive, which increases with increasing uncertainty, is

$$H(p_1, ..., p_n) = -K \sum_i p_i \log p_i \tag{85}$$

Further on he continues: "It is now evident how to solve our problem; in making inferences on the basis of the partial information we must use that probability distribution which has the maximum entropy subjected to whatever is known. This is the only unbiased assignment we can make; to use any other would amount to arbitrary assumption of information which by hypothesis we do not have."

Equations (83) and (84), introduced by Jaynes, are not different from Equations (26) and (27), given by Darwin and Fowler (and later on by Schrodinger), when solving the problem of "the distribution of a given amount of energy E over N identical systems." Where the function $f(x_i)$ assumes the values of the quantized energy e_l, the expectation value of $f(x)$ is the total energy E and the probability p_i is the normalized occupation number a_l/N. Moreover, we know that quantity x_i belongs to the multinomial formula suggested in Equation (25). Hence, a complete generalization of the methods employed by Darwin and Fowler is possible. Jaynes has shown that this generalization found application in the domain of information theory in order to recover the principal noiseless Shannon theorems. In Reference [32] he considers a "partition function" "base 2" (instead of "base e"), and recovers the partition function as

$$Z(\lambda) = \sum_{i=1}^{a} 2^{-\lambda l_i} \tag{86}$$

where a is the dimension of the alphabet, l_i is the length of the code word (in binary code) and λ is a generic Lagrange multiplier. Shannon's First Theorem (in effect, since Shannon's capacity is expressed in bit/s, equality is verified when (86) is written according to the duration of the symbols), expressed by Equation (72), becomes then

$$Z(\lambda) = 1 \tag{87}$$

Jaynes comments: "It is interesting that such a fundamental notion as channel capacity has no thermodynamics analog. In thermodynamics, the absolute value of the entropy has no meaning; only entropy differences can be measured in experiments. Consequently, the condition that H is maximized, equivalent to the statement that the Helmholtz free energy function vanishes ($A = E - TS = 0$), corresponds to no condition which could be detected experimentally".

In fact, we remember from Equation (48) that Helmholtz free energy (which is often indicated by A instead of F) is equivalent to $-\log Z/T$ and goes to zero when Equation (88) is verified.

Moreover, even the fundamental thermodynamics relation in Equation (45), underlined by Darwin and Fowler (which expresses entropy as the sum of the characteristic function Y and the internal energy of the system divided by its temperature U/T), can be replaced by:

$$S = \log Z(\lambda) + \lambda \langle l \rangle \tag{88}$$

where Jaynes made a one-to-one correspondence between the thermodynamic entropy of Darwin and Fowler and the probabilistic entropy of Shannon which, expressed in bit/symbol, becomes [39]:

$$H = \frac{S}{\langle l \rangle} = \frac{\log Z(\lambda)}{\langle l \rangle} + \lambda \tag{89}$$

The maximum of the transmission rate will consequently be obtained by setting the derivative of Equation (89) at zero. Thus for

$$\frac{dH}{d\langle l \rangle} = -\frac{\log Z(\lambda)}{\langle l \rangle^2} = 0 \Rightarrow \log Z(\lambda) = 0 \tag{90}$$

which is verified for Equation (87) or when the transmission rate equals the capacity, as expressed by Shannon's Theorem 9 (Equation (82)). Of course, Shannon's Second Theorem is obtained by Stirling's approximation of the multinomial formula.

Hence, we have seen that the generalization of the use of the partition function allow us to achieve, relatively quickly, significant results in information theory.

Together with Jaynes we can state that: "the basic mathematical identity of the two fields (thermodynamics and information theory) has had, thus far, very little influence on the development of either. There is an inevitable difference in detail, because the applications are so different; but we should at least develop a certain area of common language we suggest that one way of doing this is to recognize that the partition function, for many decades the standard avenue through which calculations in statistical mechanics are 'channelled' is equally fundamental to communication theory. Even within communication theory, there are advantages to be had by adopting this terminology" [39].

After Jaynes, the centrality of the partition function was recognized by Tribus, who proposed a route opposed to that of Jaynes. Tribus considers information theory the foundation theory for the development of the whole of thermodynamics. In his 1961 paper, "Information Theory as the Basis for Thermostatics and Thermodynamics", Tribus observed that: "The problem of statistical inference (as Hume pointed out) is to find a method for assigning probabilities that is minimally biased; that is, which uses the available information and leaves the mind unbiased with respect to what is not known" [40]. Hence, he continues: Of course, if in the past we have observed the frequency of events and found them to be of magnitude f_i, then for the future we should assign, as a rational procedure, $p_i = f_i$ (thus frequency equals probability). However, what should we do if the frequencies of occurrence are not measurable, as is the case with atoms and molecules, as well as many other problems of statistical inference? The answer, according Jaynes, is this (Jaynes' "Principle of Minimum Prejudice"): Assign that set of values to pi which is consistent with the given information and which maximizes the uncertainty [40].

After these assumptions, Tribus develops a complete theory of thermodynamics (which consistently reflects those developed by Darwin and Fowler) and introduces an interesting interpretation of the expression of the expectation value of variable energy $\langle U \rangle$ (which in the generalization given by Jaynes corresponds to Equation (83)). He shows that the complete differential of expected energy:

$$d\langle U \rangle = d\sum_i p_i \varepsilon_i = \sum_i dp_i \varepsilon_i + \sum_i p_i d\varepsilon_i \tag{91}$$

presents two terms. The first term modifies the statistic of the occupation number (Equation (26), or the statistics of the partition function), and can be interpreted as a change in heat:

$$dQ = \sum_i dp_i \varepsilon_i \tag{92}$$

This change causes another change in the uncertainty of the system and consequently a related change of entropy, $dS = dQ/T$.

The second term modifies the allowed eigenvalues of the energy (or the eigenvalues of the partition function) and can be interpreted as a change in the given work:

$$-\langle dW \rangle = \sum_i p_i d\varepsilon_i \tag{93}$$

This change does not cause a change in the uncertainty of the system and therefore entropy remains constant. Hence, the change can be interpreted as a reversible work.

After the above interpretations, Equation (91) becomes

$$d\langle U \rangle = dQ - \langle dW \rangle \tag{94}$$

which represents the First Law of Thermodynamics. Tribus comments: "In the classical development of thermodynamics the concepts of heat and work are usually taken as primitive. In this treatment, they are derived (perhaps delineated is a better word)" [40].

Grandy generalizes the above concept. He expresses Equation (94) following Jaynes' formalism, and it becomes:

$$d\langle f \rangle = dQ + \langle df \rangle \tag{95}$$

Grandy explains that Equation (95) "can be interpreted as a generalized First Law of Thermodynamics which is now seen as a special case of a more general rule in probability theory: a small change in any expectation value consists of a small change in the physical quantity ("generalized work") and a small change in the probability distribution ("generalized heat"). Just as with ordinary applications of the First Law, we see that the three ways to generate changes in any scenario are interconnected, and specifying any two determines the third" [36].

The fundamental relation in below equation, pointed out by Darwin and Fowles,

$$S = k \log Z + \frac{U}{T}$$

can be directly derived from the "partition function" of a particular ensemble, called micro-canonical (see for example Wannier [41]), an ensemble where the volume and the number of molecules are constant and the average energy/molecule is quasi-constant. In fact, in this case the partition function results in:

$$Z = \sum_l e^{-\varepsilon_l / kT} = M e^{-\langle U \rangle / kT} \tag{96}$$

where M represents the spectral multiplicity or degeneracy factor and $\langle U \rangle$ is the average energy possessed by all the states. The term M gives an idea of "how many states of energy around U are present in the system" [24]. As pointed out by Wannier, "a real Z is admitted not the simple structure but the complications are in many cases not as great as one would guess form the appearances". As the range of contributing energy is usually small, the Boltzmann exponential cannot depart much from (the above). Moreover, "It must be understood of course that a certain amount of temperature dependence is now inherent in M, although this dependence is usually weaker than the dependence

on $1/kT$ directly visible in the formula" [24]. By taking the logarithm of Equation (97) and multiplying it by k we obtain

$$k \log M = k \log Z + \frac{U}{T} = S \tag{97}$$

Grandy underlined that entropy, Equation (45), represents an absolute maximum never overcome by any other probability distribution. Hence, it is possible to write, following Jaynes's formalism, that given a generic probability distribution $\{p_i\}$ it is:

$$-k \sum_i p_i \log p_i \leq k \log Z + k\lambda \langle f \rangle \tag{98}$$

where the right side remains fixed and the equality is valid for the micro-canonical distribution.

8. Conclusions

Darwin's and Fowler's intuition in 1920 has permitted us to identify in the partition function a route to the same equation discovered by Planck, through an evaluation of the average state of energy rather than a maximum of entropy. Moreover, Darwin and Fowler recognized in the logarithm of the partition function the parameter which connects statistical mechanics with thermodynamics. Further reflections by Jaynes in the 1950s extended this connection to probability, offering a unifying tool which enables us to recover the main noiseless information theory theorems.

The partition function acts as a sort of energy spectrum whose eigenstates are those allowed by the systems (classical or quantum). Jaynes had the merit to extend this concept to a more general vision wherein the eigenstates are substituted by n_i distinguishable groups allowing the accumulation of indistinguishable results originated by N independent draws. Each draw is associated to an elective property of which only the expectation value is known. This mechanism originates a standard multinomial distribution from which Shannon's entropy is derived.

This generalization allowed the recovery of the main principles of thermodynamics, the recovery of the Planck equation (from whose form the photon concept was originated), and the main principles of information theory (which uses the bit as unit of measure). Hence, photons and bits, these two ingredients which sustain the 21st-century "information age", share the same roots in the thermodynamics of the 19th century.

Acknowledgments: This paper summarizes the content of a course I did at Milano Politecnico for PhD students whose comments helped me to improve it. I thank the members of PoliCom Labs and in particular Pierpaolo Boffi, Paolo Martelli and Paola Parolari who have been interested to discuss with me in the long journey that led me to elaborate the text. I also thank my daughter in law Valeria Vitelli who was been interested in commenting with me many parts of the paper.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Tolman, R.C. *The Principles of Statistical Mechanics*, 1st ed.; Oxford University Press: New York, NY, USA, 1938.
2. Kuhn, T.S. *Black-Body Theory and the Quantum Discontinuity 1894–1912*; Oxford University Press: New York, NY, USA, 1978.
3. Badino, M. *The Bumpy Road, Max Planck from Radiation Theory to the Quantum 1896–1906*; Springer: Berlin, Germany, 2015.
4. Agassi, J. *Radiation Theory and the Quantum Revolution*; Birkhauser: Basel, Switzerland, 1993.
5. Darrigol, O. Statistical and Combinatorics in Early Quantum Theory. In *Historical Studies in the Physical and Biological Sciences*; University of California Press: Berkeley, CA, USA, 1988.
6. Darrigol, O. *Statistical and Combinatorics in Early Quantum Theory, II: Early Symptoma of Indistinguishability and Holism*; University of California Press: Berkeley, CA, USA, 1991.
7. Gearhart, C.A. Planck, the Quantum and the Historians. *Phys. Perspect.* **2002**, *4*, 170–215. [CrossRef]
8. Loudon, R. *The Quantum Theory of the Light*; Oxford University Press: New York, NY, USA, 1983.

9. Scully, M.O.; Zubairy, M.S. *Quantum Optics*; Cambridge University Press: Cambridge, UK, 1997.
10. Carnot, S. *Reflections on the Motive Power of Fire*; Dover: Mineola, NY, USA, 1960.
11. Clausius, R. *The Mechanical Theory of Heat*; Hardpress: Sligo, Ireland, 2015.
12. Clausius, R. Under verschieden fur die Anwendung bequeme Formen der Hauptgleichungen der mechanischen Warmetheorie. *Ann. Phys.* **1865**, *125*, 353–400. [CrossRef]
13. Planck, M. *The Theory of Heat Radiation*; The Maple Press: York, PA, USA, 1948.
14. Tersoff, J.; Bayer, D. Quantum Statistics for Distinguishable Particles. *Phys. Rev. Lett.* **1983**, *50*, 553–554. [CrossRef]
15. Niven, R.H.; Grendar, M. Generalized classical, quantum and intermediate statistics and the Polya urn model. *Phys. Lett. A* **2009**, *373*, 621–626. [CrossRef]
16. Darwin, C.G.; Fowler, R.H. On the partition of energy. *Philos. Mag.* **1922**, *6*, 450–479. [CrossRef]
17. Brillouin, L. *Les Statistiques Quantiques et Leurs Applications*; Les Presses Universitaire de France: Paris, France, 1930.
18. Schrodinger, E. *Statistical Thermodynamics*, 1st ed.; Cambridge University Press: Cambridge, UK, 1948.
19. Massieu, M. Thermodinamique—Sur les functions caracteristiques des divers fluids. *Compte Rendus* **1869**, *69*, 858–862. (In French)
20. Darwin, C.G.; Fowler, R.H. On the partition of energy, Part II, Statistical principles and thermodynamics. *Philos. Mag.* **1922**, *6*, 823–842. [CrossRef]
21. Schrodinger, E. *What is the life*; Cambridge University Press: Cambridge, UK, 1945.
22. Nyquist, H. Certain factors affecting telegraph speed. *Bell Syst. Tech. J.* **1924**, *3*, 324–346. [CrossRef]
23. Hartley, R. Transmission of Information. *Bell Syst. Tech. J.* **1928**, *7*, 535–563. [CrossRef]
24. Tuller, W.G. Theoretical Limitations on the Rate of Transmission of Information. *Proc. IRE* **1949**, *37*, 468–478. [CrossRef]
25. Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [CrossRef]
26. Shannon, C.E. A mathematical theory of communication, Part 2. *Bell Syst. Tech. J.* **1948**, *27*, 623–656. [CrossRef]
27. Szilard, L. On the decrease of Entropy in a thermodynamic system by the intervention of intelligent beings. *Z. Phys.* **1929**, *53*, 840–856. [CrossRef]
28. Bennet, C.H. Demons, engines and the second law. *Sci. Am.* **1987**, *257*, 108–116. [CrossRef]
29. Brillouin, L. The negentropy principle of Information. *J. Appl. Phys.* **1953**, *24*, 1152–1163. [CrossRef]
30. Brillouin, L. Maxwell's Demon Cannot Operate: Information and Entropy. I. *J. Appl. Phys.* **1951**, *22*, 334–337. [CrossRef]
31. Brillouin, L. *Science and Information Theory*; Academic Press: Waltham, MA, USA, 1956.
32. Brillouin, L. Negentropy and Information in Telecommunications, Writing, and Reading. *J. Appl. Phys.* **1954**, *25*, 595–599. [CrossRef]
33. Landaurer, R. Irreversibility and heat generation in the computing process. *IBM J. Res. Dev.* **1961**, *5*, 183–191. [CrossRef]
34. Bennet, C.H. Notes on the history of reversible computation. *IBM J. Res. Dev.* **2000**, *44*, 270–277. [CrossRef]
35. Khandekar, A.; McEliece, R.; Rodemich, E. The discrete noiseless channel revisited. *Proc. ISCTA* **1999**, *99*, 115–137.
36. Grandy, W.T. *Entropy and the Time Evolution of Macroscopic Systems*; Oxford University Press: New York, NY, USA, 2008.
37. Hamming, R.W. *Coding and Information Theory*; Prentice-Hall: Upper Saddle River, NJ, USA, 1980.
38. Jaynes, E.T. Information theory and Statistical Mechanics. *Phys. Rev.* **1957**, *106*, 620–630. [CrossRef]
39. Jaynes, E.T. Note on Unique Decipherability. *IRE Trans. Inf. Theory* **1959**, *5*, 98–102. [CrossRef]
40. Tribus, M. Information theory as the basis for thermostatics and thermodynamics. *J. App. Mech.* **1961**, *28*, 1–8. [CrossRef]
41. Wannier, G.H. *Statistical Physics*, 1st ed.; Wiley & Sons: New York, NY, USA, 1966.