


# Entropy: From Thermodynamics to Information Processing

Jordão Natal <sup>1,\*</sup>, Ivonete Ávila <sup>2</sup> , Victor Batista Tsukahara <sup>1</sup>, Marcelo Pinheiro <sup>3</sup> and Carlos Dias Maciel <sup>1,\*</sup> 

<sup>1</sup> Signal Processing Laboratory, Department of Electrical and Computing Engineering, University of São Paulo (USP), São Carlos 3566-590, Brazil; vhbtsukahara@usp.br

<sup>2</sup> Laboratory of Combustion and Carbon Captur, Department of Energy, School of Engineering, State University of São Paulo (Unesp), São Carlos 3566-590, Brazil; iavila@feg.unesp.br

<sup>3</sup> Versatus Studio, São Carlos 4011-002, Brazil; marcelo@versatus.studio

\* Correspondence: jordao.oliveira@usp.com (J.N.); maciel@sc.usp.br (C.D.M.)

**Abstract:** Entropy is a concept that emerged in the 19th century. It used to be associated with heat harnessed by a thermal machine to perform work during the Industrial Revolution. However, there was an unprecedented scientific revolution in the 20th century due to one of its most essential innovations, i.e., the information theory, which also encompasses the concept of entropy. Therefore, the following question is naturally raised: “what is the difference, if any, between concepts of entropy in each field of knowledge?” There are misconceptions, as there have been multiple attempts to conciliate the entropy of thermodynamics with that of information theory. Entropy is most commonly defined as “disorder”, although it is not a good analogy since “order” is a subjective human concept, and “disorder” cannot always be obtained from entropy. Therefore, this paper presents a historical background on the evolution of the term “entropy”, and provides mathematical evidence and logical arguments regarding its interconnection in various scientific areas, with the objective of providing a theoretical review and reference material for a broad audience.



**Citation:** Natal, J.; Ávila, I.; Tsukahara, V.B.; Pinheiro, M.; Maciel, C.D. Entropy: From Thermodynamics to Information Processing. *Entropy* **2021**, *23*, 1340. <https://doi.org/10.3390/e23101340>

Academic Editors: José A. Tenreiro Machado and Leonid M. Martyushev

Received: 20 July 2021

Accepted: 24 September 2021

Published: 14 October 2021

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

**Keywords:** entropy; thermodynamics; information theory

## 1. Introduction

Entropy is a measure largely used in science and engineering [1]. It was initially introduced in thermodynamics by Clausius [2], developed by Boltzmann and Gibbs through the 19th century [3] and generalized by Shannon in the 20th century [4] to the point that it can be applied in a broad range of areas. It has been applied to biology [5–9], economics [10–12], engineering [13–15], linguistics [16–18] and cosmology, at the center of one of the greatest open problems in science [10–12]. Given this general use in different fields of knowledge, it is important to think about what the measure of entropy actually represents in each different context and the possible equivalence between them.

A misunderstanding about the meaning of entropy is reported in several papers when applied to areas other than physics and information theory [19–23]—sometimes even in these areas—[24,25], and is also reported among students [26]. It is not uncommon to signify entropy as “disorder” [27,28]; although we can often assume for didactic appeal, moderation is necessary so that we do not use a subjective human concept. This is not a rigorous statement since “disorder” is a subjective construction and cannot be measured by entropy [29].

This work intends to contribute with a review of the historical evolution of the concept of entropy, demonstrating the current level of understanding regarding the connection between thermodynamics and information theory. Our target audience is mainly readers outside the fields of physics and engineering, who will have no trouble following two demonstrations about the equivalence between Boltzmann–Gibbs entropy and information theory entropy.

This paper is structured as follows: Section 2 presents the historical evolution of the concept of entropy in physics and information theory; Section 3 explores conceptual rela-

tionships between apparitions in these two fields; and finally, we conclude by summarizing the discussed points in Section 4.

## 2. Historical Background

The Industrial Revolution and the development of steam engines was a period of reflection on the physical properties of matter, energy, heat, work and temperature. These phenomena needed to be well understood in order to create efficient engines. It is in this context that the empirical laws that describe the thermal behavior of macroscopic matter were systematized in what it is know today as classical thermodynamics.

In 1825, Sadi Carnot, expanding on his father's reflections, who had already inferred that perpetual motion was impossible, noted the impossibility of an ideal thermodynamic cycle—Carnot cycle—being reversible [30]. These were actually the first, perhaps rudimentary, formulations of the second law of thermodynamics.

### 2.1. Clausius Entropy

A few decades later, Clausius developed the concept of an extensive quantity, which cannot be measured directly, called entropy. This was associated with an asymmetry in the flow of heat; in nature, heat always flows from a hotter body to a colder one, but the reversal process does not happen spontaneously [31].

The concept of entropy (from the Greek word meaning “change”) was developed to explain the tendency of heat, pressure and density to gradually disappear with time, or similarly, the inevitable generation of heat when work is done on a system by changing temperature. The definition of the state function  $S$ , in honor of Sadi, called entropy, is as follows:

$$dS = \frac{\delta Q}{T} + \delta S_{gen} \quad (1)$$

with unit J/K.  $\delta Q$  is conventionally used to indicate an inexact differential [32] in which integration depends not only on the starting and ending states, but on the process path in between. On the other hand, entropy is a thermodynamic property; therefore,  $dS$  is an exact differential, and integration does not depend on the process path between the starting and ending states. The amount of entropy generation,  $\delta S_{gen}$ , is null in reversible processes and greater than zero when an irreversible phenomena occurs within the system. However, there is a modification in the system's entropy due to a change in state,  $dS$ , which can be either positive or negative depending on the direction of heat transfer (to or from the system).

For an adiabatic process,  $\delta Q = 0$ , and when the entropy differential,  $dS$ , is not null, its value is  $\delta S_{gen}$  and is always associated with irreversible paths. Contrary to energy, the entropy of an isolated system increases when the process occurs irreversibly, and thus, is not conserved. A reversible process is ideal, but it never really occurs in nature. Therefore, an amount of irreversibility is always present in the system, i.e., the isolated system's entropy keeps increasing and never reduces.

This concept refers to the increase in entropy principle [33]: the entropy variation of an isolated system (a) never decreases and (b) tends to increase, due to the process' irreversibility.

### 2.2. Boltzmann–Gibbs Entropy

In the late 1800s, cutting-edge physics was trying to model the ideal gas problem. In this context, Maxwell—and shortly afterwards, Boltzmann—developed the Boltzmann equation as a new model for some problems in classical mechanics, such as that of ideal gas.

The entropy,  $S$ , of an ideal gas is a state function of a possible number of microstates,  $W$ , for molecules in a macrostate (defined by temperature, volume and pressure). Considering a system comprising an ideal gas and dividing it into two parts, it is hypothesized according to [3] that  $S = S_1 + S_2$  and  $W = W_1 \times W_2$ , given the Boltzmann equation,  $S = k \log W + c$ , as shown in Figure 1. Hoffmann [34] considered that an ideal gas at 0 K has null entropy and only one microstate,  $k \log 1 + c = 0 \rightarrow c = 0$ , and  $S = k \log W$  is the entropy of an ideal

gas, where  $k$  is the Boltzmann constant. Gibbs [35] enhanced the concept of Boltzmann entropy in cases where microstates are not evenly likely:

$$S = -k \sum_{i=1}^n p_i \log p_i \quad (2)$$

where  $p_i$  is the probability of the  $i$ -nth microstate, given that all  $W$  microstates are evenly likely, and  $p_i = (1, 2, 3, \dots, n) = 1/n$  and Equation (2) are the same Boltzmann equation. This model led to the notion of entropy with statistical meaning and the conciliation of microscopic reversibility with macroscopic irreversibility.

$$S(W_1) + S(W_2) = S(W_1 W_2)$$

Deriving both sides with respect to  $W_1$  and keeping  $W_2$  constant results in the following:

$$S'(W_1) = W_2 S'(W_1 W_2)$$

Deriving in  $W_2$  by keeping  $W_1$  constant and applying the chain rule, we obtain the following:

$$S'(W_1 W_2) + W_1 W_2 S''(W_1 W_2) = 0$$

$$S'(W) + W S''(W) = 0$$

Replacing  $S'(W) = f(W)$ , we obtain the following:

$$f(W) + W \frac{df(W)}{dW} = 0$$

$$f(W)dW + Wdf(W) = 0$$

$$(fW)' = 0$$

By integrating both sides, it returns to the following:

$$fW = k$$

which is the same as the following:

$$W \frac{dS}{dW} = k$$

$$\int dS = k \int \frac{dW}{W}$$

$$S = k \log W + c$$

**Figure 1.** Boltzmann's entropy formula derivation: since it is known that total entropy  $S$  is the sum of its parts and the total number of microstates  $W$  is the product of its parts, the only function  $S(W)$  relating these variables is a logarithm.

### 2.3. Shannon Entropy

In 1948, Shannon [4] published the foundational concept of information theory with the concept of entropy of the information of a discrete probability distribution related to the maximum possible data compression.

Following an axiomatic approach, with one enunciate and two desirable properties, it is possible to define the Shannon entropy. Considering an event with  $p$  probability, and the corresponding function  $I(p)$ , the two desirable properties are as follows: (i)  $I(p) \geq 0$  is a decreasing function of  $p$ ; (ii) for any two independent events with probabilities  $p_1$  and  $p_2$ ,  $I(p_1 p_2) = I(p_1) + I(p_2)$ . The  $I(p)$  interpretation is a measure of “surprise” or “uncertainty” depending on the occurrence of the event. From here, it is possible to determine that the logarithmic function,  $-\log p$ , satisfies the requested conditions for  $I(p)$ . Now, let  $X$  be a random variable. The random variable  $I(p(X)) = -\log p(X)$  is called *self-information* or *information content* of  $X$  [36].

In the case of a discrete random variable  $X$  with probability distribution  $p(x)$ , the average information content about  $X$  is given by the expected value or Shannon entropy:

$$H(X) = - \sum_{i=1}^n p_i(x) \log p_i(x) \quad (3)$$

The above entropy is dimensionless, although it is common to use the base 2 logarithm and measuring the entropy itself in bits. Apparently, Shannon obtained the name “entropy” from von Neumann himself, as he related [37]:

*“My greatest concern was what to call it. I thought of calling it ‘information’, but the word was overly used, so I decided to call it ‘uncertainty’. When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, ‘You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one really knows what entropy really is, so in a debate you will always have the advantage’.”*

Shannon’s original motivation was to create a measure useful in quantifying the channel capacity needed to send a binary message (encoded in a given electrical signal) through telephones lines. One of the uses for the entropy in information theory lies in the measurement of ultimate data compression. For example [1], let us suppose that eight letters, whose frequencies are  $1/2, 1/4, 1/8, 1/16, 1/64, 1/64, 1/64$  and  $1/64$ , respectively, must be sent. By initially using binary coding, one could assume that 3 bits are needed (000, 001, 010, 011, 100, 101, 110, 111). However, since their frequencies are different, it is possible to encode them as 0, 10, 110, 1110, 111100, 111101, 111110 and 111111, making the average number of bits 2. A fundamental extension of this concept is the derivation of the mutual information between variables  $X$  and  $Y$ , given by  $I(X; Y) = H(X) - H(X|Y)$ , which measures, on average, how much knowing  $Y$  decreases the uncertainty over  $X$ .

It is important to emphasize that Shannon entropy by itself does not provide any means to estimate the probability distribution; therefore, it relies on statistics or the observer’s knowledge. In information theory, it is not uncommon to assume uniform distribution, which makes entropy become a trivial function measuring the multiplicity of the different symbols, just like its counterpart measure of Boltzmann–Gibbs entropy that counts the number of possible micro-states of particles in a given volume of space.

#### 2.4. Partial Information Decomposition

Recent advances in information theory resulted in the methodology called partial information decomposition [38]. Given a set of variables  $R_1, R_2, \dots, R_n$  defined as inputs of a system, and an output  $Y$ , the objective of this method is to decompose the information on  $\mathbf{R}$  (be it on the independent  $\mathbf{R}$  components or joint distributions of these elements). This proposal has the objective of providing information theory with the necessary tools for characterizing the structure of multivariate interactions. Let  $A_1, A_2, \dots, A_k$  be nonempty and overlapping sets of  $\mathbf{R}$  called sources. Since the mutual information for each  $I(S; A_i)$  is an average value over the distributions as mentioned before, two sources might provide the same average amount of information, while also providing information about

different outcomes of  $S$  [38]. Formally, the information about  $S$  provided by  $\mathbf{A}$  is given by the following:

$$I(S; \mathbf{A}) = \sum_s p(s) I(S = s; \mathbf{A}) \quad (4)$$

in which the specific information  $I(S = s; \mathbf{A})$  is given by the following:

$$I(S = s; \mathbf{A}) = \sum_{\mathbf{a}} p(\mathbf{a}|s) \left( \log \frac{1}{p(s)} - \log \frac{1}{p(s|\mathbf{a})} \right) \quad (5)$$

and defining,

$$I_{\min}(S; \mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_k) = \sum_s p(s) \min_{A_i} I(S = s; \mathbf{A}_i) \quad (6)$$

the partition information function PI can be defined as follows:

$$I_{\min}(S; \alpha) = \sum_{\beta \preceq \alpha} PI(S; \beta) \quad (7)$$

In Equation (7),  $\alpha$  belongs to the set of all nonempty subsets of  $\mathbf{R}$ , and the ordering relationship is given by  $\alpha \preceq \beta \iff \forall \mathbf{B} \in \beta, \exists \mathbf{A} \in \alpha, \mathbf{A} \subseteq \mathbf{B}$ . The partial information function quantifies the (redundant) information coming from  $\alpha$  that does not come from any simpler collection  $\beta \preceq \alpha$ .

### 2.5. Algorithmic Information Theory

Algorithmic information theory is the application of elements of Shannon's theory to algorithms. The most famous of these applications is the Kolmogorov complexity (KC) in a universal Turing machine (a finite state machine that has an input of symbols of a finite alphabet and processes them, returning a new set of symbols) [39]. The KC  $K(s)$  of the string  $s$  is the number of units of information (bits, for example) of the smallest algorithm in a language that can reproduce the object. This measure of complexity has, in its core, an interrogation about randomness. If a string is deterministic, then its KC is low since the code that generates it is simple. For example, the string "001001001001001" and the string "011001101111011" both have 15 bits, but the first one can be coded as "repeat (001) 5 times", and the second one seems to be random, so the code to generate it will have to contain the entire string.

Shannon's entropy and KC hold a remarkable relationship. Using the Kraft inequality, it can be shown that the following holds [1]:

$$E \left[ \frac{1}{n} K(X^n | n) \right] \rightarrow H(X) \quad (8)$$

and therefore, the compressibility of KC in the universal computer goes to the entropy limit. Moreover, [40] showed that, even though Kolmogorov complexity and Shannon entropy are conceptually different measures, their values are equivalent when dealing with both recursive probability distributions (those which are computable by a Turing machine) or in the case of a time-bounded relationship; this is not always the case in such generalizations as Tsallis and Rényi entropies. However, it is important to notice that these theoretical equivalences suppose that there is perfect knowledge about the distributions originating the data, which is hardly the case [41]; since the KC is distribution independent, which is not the case of the statistical approaches from Shannon's entropy, one can almost certainly expect a different measurement from these two tools.

New developments in this area resulted in the so-called algorithmic thermodynamics, in which an analogue to the fundamental thermodynamic equation  $dE = TdS - PdV + \mu dN$  and the partition function  $Z$  are defined in order to study cycles on algorithms analogous to those in heat engines [42], or how problems such as recursion and networks can be dealt with, using information theory tools [43,44].

## 2.6. Algorithmic Information Dynamics

This is a new field focused on the connections between information theory and causality [45]. Algorithmic information dynamics (AID) deals with dynamic systems such as its mathematical model, and is computable, combining perturbation theory and algorithmic information theory, using Bayes' theorem.

One of the tools used by AID is the coding theorem method (CTM), which deals with compressing without relying on statistical frameworks [43]. It is based on a fundamental identity, given a fundamental prior probability  $m(s)$  describing a string and the Kolmogorov complexity  $K(s)$ :  $m(s) = 2^{-K(s)} + c$ .

Another tool introduced by AID is the block decomposition method (BDM). One of the motivations justifying both of these methods is the Champernowne constant ( $x = 0.1234567891011\dots$ ) information content since the sequence generating its digital expansion has no statistical pattern; therefore, it would have maximum entropy on statistical approaches, such as Shannon's entropy [45].

BDM therefore extends the power of CTM in the field of algorithmic randomness and should be useful in understanding the computation aspects of cognitive processes in the brain [45–47].

## 3. Equivalence of Entropy in Thermodynamics and Information Theory

### 3.1. Unity Analysis

The Boltzmann constant linking the thermodynamic macroscopic quantity  $S$  and the microscopic sum over all the possible micro-states of a system—a dimensionless quantity—clearly has the dimensions of energy divided by temperature (J/K). Since Shannon lacks any proportionality constant, such as the Boltzmann constant, it has no dimension.

Considering purely dimensional units, Shannon's formulation of entropy seems to have no connection with the formulation of Clausius or Boltzmann–Gibbs entropies. Although being a concept that is purely probabilistic, it shares its randomness nature with the latter. It was demonstrated that the unit is historically associated with the definition of the Kelvin temperature system: the Lagrangian temperature has units of energy in statistical mechanics [48]. In plasma physics, it is common to express temperature in eV [49,50]. In a more generic approach, thermodynamical entropy is dimensionless, and the difference between Shannon and Gibbs's entropies lies in Boltzmann's constant.

### 3.2. Underlying Probability

In statistical thermodynamics, the probability of a particular microstate as a function of its energy is given by the so-called Boltzmann distribution,  $p_i \propto e^{-E_i/kT}$ , a sufficient and necessary condition for the compatibility of statistical mechanics (with microscopic reversibility) and thermodynamics (with macroscopic irreversibility) formulations and, therefore, the equivalence between the Clausius entropy and Boltzmann–Gibbs entropy.

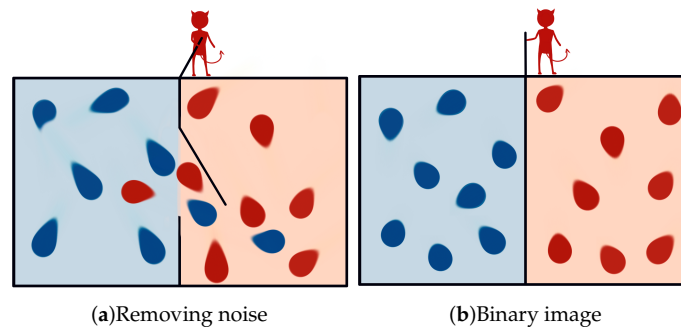
As we saw earlier, however, in information theory, it is not possible to derive any underlying probability distribution, *which makes Shannon's entropy a mere combinatorial measure of diversity*. This limitation, so to speak, of Shannon entropy is one of the main attractions of the formulation since it can only quantify meaning when one knows the type of information being treated. Thus, it can be used for a large range of problems involving information.

### 3.3. Shannon Entropy and Thermodynamics

Years before Shannon's information theory, a thought experiment known as Maxwell's demon (Figure 2) challenged the second law of thermodynamics. In his own words, it is described as follows [51]:

*“... if we conceive of a being whose faculties are so sharpened that he can follow every molecule in its course, such a being, whose attributes are as essentially finite as our own, would be able to do what is impossible to us. For we have seen that molecules in a vessel full of air at uniform temperature are moving with velocities by no means*

*uniform, though the mean velocity of any great number of them, arbitrarily selected, is almost exactly uniform. Now let us suppose that such a vessel is divided into two portions, A and B, by a division in which there is a small hole, and that a being, who can see the individual molecules, opens and closes this hole, so as to allow only the swifter molecules to pass from A to B, and only the slower molecules to pass from B to A. He will thus, without expenditure of work, raise the temperature of B and lower that of A, in contradiction to the second law of thermodynamics."*

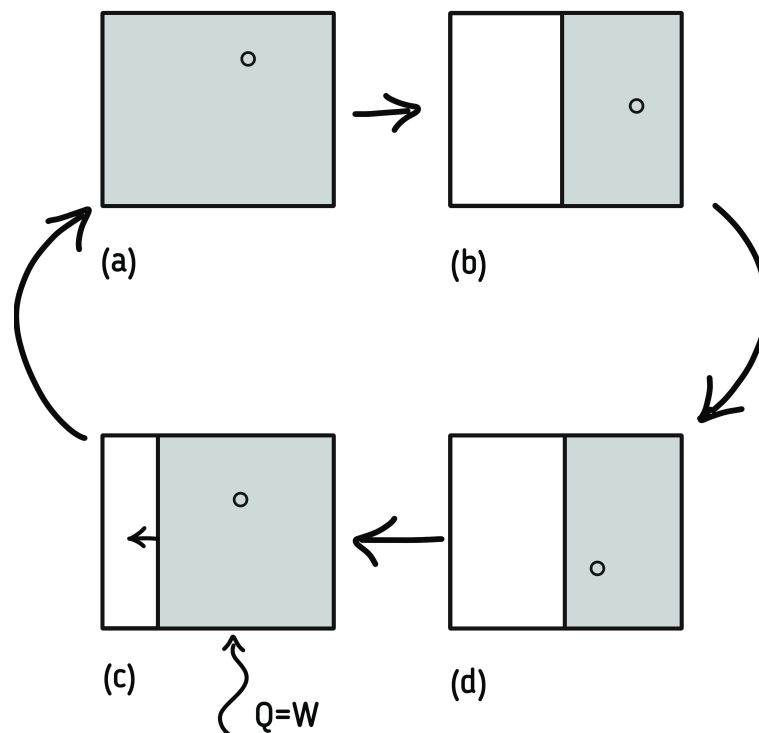


**Figure 2.** Maxwell's demon: a being who knows the velocity of every particle in the box and can select their passage, using a opening in the wall that divides it, which could separate those with high energy from those with low energy without performing work, thus violating the second law of thermodynamics. The demon has to forget the past states of the system but, according to Landauer's principle, this process generates heat (at least  $kT \log 2$  J per bit erased) and entropy.

The demon, capable of measuring the kinetic energy of the molecules, can separate fast and slow particles. In this way, the overall entropy of the system will be decreased in a clear violation of the second law of thermodynamics. In addition to that, even in Maxwell's time, there were already proposals for measurement apparatuses that clearly would not introduce an increase in entropy capable of compensating for the overall decrease proposed in the original setup.

The first important step to clarify the discussion was suggested in 1929 by Szilard [52], introducing a variation of Maxwell setup known as a Szilard engine. The idea was to focus on the measurement itself performed by the demon rather than the work he would have done.

The new thought experiment consists of a single molecule of gas inside a box with thermal walls (connected to a reservoir); the demon, in addition to measuring the kinetic energy of the single particle, also inserts and removes a piston in order to divide the vessel in two parts. After its introduction, the gas can isothermally expand to its equilibrium position, doing work that is the equivalent of  $kT \log 2$  J (Figure 3). Considering that the demon needs to acquire and store information, even if for a small fraction of time, about the kinetic energy of the particles, it has to be part of the macrostate of the system. Therefore, the information in the demon's brain can be part of one of the possible configurations, so the second law is not violated.



**Figure 3.** The process of extracting work from a system, thought of by Szilard: in (a), there is a single molecule of a fluid inside a box with energy  $Q$ . If one knows in which half of the box the molecule is (i.e., a single bit of information about its position), a piston can be inserted by halving the box (b) and from the fluid expansion, work ((c,d))  $W = Q$  can be extracted from the system while it returns to its initial state.

Although the Szilard engine was the first ever link relating information with thermodynamics, it is still unable to explain the reverse cycle, where the demon forgets what he knew, consequently decreasing the entropy of the system. In fact, the explanation of the reverse cycle came only in 1982 with Bennett, using Landauer's principle. The principle states that to erase information (logical bit), at least an increment of  $kT \log 2$  J of heat is needed [53]. Moreover, the principle can be used to solve the Maxwell's demon paradox, allowing the demon brain to be updated (forgetting some information to acquire and store others), constituting an irreversible process that generates heat and increases entropy. Rescuing the second law of thermodynamics with the use of information theory also connects Shannon entropy with the already connected entropies of Clausius and Boltzmann–Gibbs. Moreover, the mutual information between the partitions is often null in thermodynamical system since the subsystems are often uncorrelated, which makes the entropy additive in conventional systems; however, in the case of Maxwell's demon, there is a correlation between the demon and the system, and the solution proposed by Landauer is in accordance with the fluctuation theorem [54].

### 3.4. Information Theoretical Proof that Boltzmann–Gibbs Entropy is the Same as Clausius's

With the development of information theory in the twentieth century and the concept of maximum entropy for statistical mechanics [55], which states that a system in global and stable thermodynamic equilibrium has reached its maximum entropy by the second law of thermodynamics (being, therefore, in the macrostate that has the most microstates, corresponding to gas velocities), it is possible to derive Clausius' entropy from Boltzmann–Gibbs formulation of statistical mechanics.



Using Equation (2), and the unitarity principle,  $\sum_i p_i = 1$ , in which  $i$  is the  $i$ -nth state, we can write the ensemble average energy as follows:

$$\langle E \rangle = \sum_i p_i E_i = U \quad (9)$$

Applying Lagrange multipliers, we have the following:

$$\mathcal{L} = -k \sum_i p_i \log p_i - \lambda_1 \left( 1 - \sum_i p_i \right) - \lambda_2 \left( U - \sum_i p_i E_i \right) \quad (10)$$

Differentiating and equaling zero, we have the following:

$$-k \log p_i - k + \lambda_1 + \lambda_2 E_i = 0 \quad (11)$$

Isolating  $p_i$ , we have the following:

$$p_i = \exp\left(\frac{-k + \lambda_1 + \lambda_2 E_i}{k}\right) \quad (12)$$

Using unitarity with Equation (12), energy can be isolated as follows:

$$\sum_i p_i = \exp\left(\frac{-k + \lambda_1}{k}\right) Z \quad (13)$$

in which  $Z$  is called the *partition function* and therefore, the following holds:

$$Z = \sum_i \exp\left(\frac{\lambda_2 E_i}{k}\right) \quad (14)$$

The partition function combines state functions, such as temperature and energy for the microstates, and has a central role in statistical mechanics [56]. Therefore, using unitarity once more, Equation (13) can be used to isolate  $\lambda_1$  as follows:

$$\lambda_1 = k - k \log Z \quad (15)$$

Thus, Equation (12) can be expressed as the following:

$$p_i = \frac{1}{Z} \exp\left(\frac{\lambda_2 E_i}{k}\right) \quad (16)$$

Using unitarity again, Equation (12) can be written as follows:

$$\exp\left(\frac{-k + \lambda_1}{k}\right) Z = 1 \quad (17)$$

Therefore, the following holds:

$$\log Z = 1 - \frac{\lambda_1}{k} \quad (18)$$

Rewriting Equation (2) in terms of  $Z$  results in the following:

$$S = -k \sum_i p_i \left( \frac{\lambda_2 E_i}{k} - \log Z \right) \quad (19)$$

$$\begin{aligned} S &= -\lambda_2 \sum_i p_i E_i + k \log Z \sum_i p_i \\ &= -\lambda_2 U + k \log Z \end{aligned} \quad (20)$$

Using the definition of thermodynamics temperature, we have the following [57]:

$$\frac{1}{T} = \frac{\partial S}{\partial U} \quad (21)$$

Since  $\frac{\partial S}{\partial U} = -\lambda_2$ , Equation (2) can be written as follows:

$$S = \frac{U}{T} + k \log Z \quad (22)$$

Now, let us change the system energy by an inexact differential  $\delta Q$ . Each microstate increases its energy by  $q_i$ . A calculation of the change in entropy results in the following:

$$dS = \frac{\delta U}{T} + k \delta \log Z \quad (23)$$

Calculating the second term, we have the following:

$$\delta \log Z = \frac{d \log Z}{dZ} \delta Z = \frac{\delta Z}{Z} \quad (24)$$

Considering that  $Z = \sum_i \exp(-E_i/kT)$ , the new partition function can be written as follows:

$$Z = \sum_i \exp\left(-\frac{E_i + q_i}{kT}\right) \quad (25)$$

Applying Taylor expansion in  $e^{-q_i/kT}$ , since  $q_i$  is infinitesimal, a good approximation is the following:

$$\exp\left(-\frac{q_i}{kT}\right) = 1 - \frac{q_i}{kT} \quad (26)$$

Therefore, this new partition function can be written as follows:

$$Z = \sum_i \left(1 - \frac{q_i}{kT}\right) \exp\left(-\frac{E_i}{kT}\right) = Z_0 + \delta Z \quad (27)$$

Therefore, the partition function variation is given by the following:

$$\delta Z = -\frac{1}{kT} \sum_i q_i \exp\left(-\frac{E_i}{kT}\right) \quad (28)$$

According to the first law of thermodynamics, the change in  $U$  can be expressed as follows:

$$\delta U = \sum_i \delta E_i p_i + \sum_i q_i p_i = \delta Q + \delta W \quad (29)$$

Calculating  $\delta \log Z$ , replacing (28) in (24), we have the following:

$$\delta \log Z = \frac{-\frac{1}{kT} \sum_i q_i \exp\left(-\frac{E_i}{kT}\right)}{Z} \quad (30)$$

However, through Equations (16) and (21), it is known that the following holds:

$$p_i = \frac{1}{Z} \exp\left(-\frac{E_i}{kT}\right) \quad (31)$$

and therefore, we have the following:

$$\delta \log Z = -\frac{1}{kT} \sum_i p_i q_i \quad (32)$$

This value is exactly  $-\delta W/kT$ . By replacing this relation in Equation (23), we obtain the following:

$$dS = \frac{\delta Q}{T} \quad (33)$$

which is the Clausius first definition of entropy.

### 3.5. Using Kullback–Leibler Divergence to Obtain an Analogous of the Second Law of Thermodynamics

Today, modern supervised machine learning techniques use extensively a measure formulated using the Kullback–Leibler divergence as a cost function when training classifiers, the cross-entropy. It is important to show the connection between this important measure of information theory with the second law of thermodynamics.

The relative entropy or Kullback–Leibler divergence between two probability distributions over  $X$ ,  $p(x)$  and  $q(x)$  is defined as follows:

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)} \quad (34)$$

It should be noticed that  $D(p||q) = 0$  if  $p = q$  (considering  $0 \log 0/0 = 0$ ) in Equation (34). However, it is not a distance in a formal sense since  $D(p||q) \neq D(q||p)$ . Relative entropy measures how similar the two distributions are.

Let us assume that  $\alpha_n$  and  $\alpha'_n$  are distributions of states in the Markov chain state space describing a physical thermal system. Once  $\alpha_{n+1}$  and  $\alpha'_{n+1}$  are their evolution in time, and  $p$  and  $q$  are their corresponding joint distribution, and given that they are in the Markov chain space, we can write the following:

$$p(x_n, x_{n+1}) = p(x_n)\pi(x_{n+1}|x_n) \quad (35)$$

$$q(x_n, x_{n+1}) = q(x_n)\pi(x_{n+1}|x_n) \quad (36)$$

in which  $r$  is the probability transition in the Markov chain. Two relations can be obtained for these equations:

$$D(p(x_n, x_{n+1})||q(x_n, x_{n+1})) = D(p(x_n)||q(x_n)) + D(p(x_{n+1}|x_n)||q(x_{n+1}|x_n)) \quad (37)$$

$$D(p(x_n, x_{n+1})||q(x_n, x_{n+1})) = D(p(x_{n+1})||q(x_{n+1})) + D(p(x_n|x_{n+1})||q(x_n|x_{n+1})) \quad (38)$$

Due to the fact that both  $p$  and  $q$  come from the Markov chain, we have  $p(x_{n+1}|x_n) = q(x_{n+1}|x_n) = \pi(x_{n+1}|x_n)$ ,  $D(p(x_{n+1}|x_n)||q(x_{n+1}|x_n)) = 0$ . Since relative entropy is always non-negative, we have the following:

$$D(p(x_n)||q(x_n)) \geq D(p(x_{n+1})||q(x_{n+1})) \quad (39)$$

$$D(\alpha_n||\alpha'_n) = D(\alpha_{n+1}||\alpha'_{n+1}) \quad (40)$$

This means that, as time passes, the probability distributions in the Markov chain (and therefore, in the system being described) becomes increasingly similar.  $D(\alpha_n||\mu)$  generates a monotonically decreasing sequence and has a limit. Assuming that  $\alpha'_n = \mu$  is a stationary distribution over time,  $\alpha'_{n+1} = \mu$ . Hence, we have the following:

$$D(\alpha_n||\mu) \geq D(\alpha_{n+1}||\mu), \quad (41)$$

which means that each distribution becomes closer to stationary as time passes. In thermodynamics, a stationary distribution is considered uniform with  $W$  different states. By applying Equation (34) in Equation (41), we have the following:

$$D(\alpha_n||\mu) = \log W - H(\alpha_n) = \log W - H(X_n) \quad (42)$$

Since  $D(\alpha_n || \mu)$  decreases,  $H(X_n)$  must increase as time passes.

#### 4. Conclusions

The concept of entropy started as an abstract mathematical property in thermodynamics at the center of the first Industrial Revolution. It developed with the advent of statistical mechanics in an important measure with a mathematical formulation that later would become ubiquitous. Further development came from information theory with Shannon entropy, which is just a combinatorial diversity, being compatible with Boltzmann–Gibbs entropy under certain conditions. Even more recent developments clarified that information is not something amorphous; instead, a medium is needed in order to be acquired and stored. Hence, the medium is the connection between temperature and the bit of information—the connection between thermodynamics and information theory, at least on a macroscopic scale, in which the thermodynamics entropy is additive since the correlation between parts of a system is null; otherwise, a more precise description, involving the fluctuation theorem is necessary. Finally, important concepts related to Shannon entropy seem to be at the center of the fourth industrial revolution [41].

It is worth noting that in the context of Shannon entropy, which applies to any probability distribution, the Boltzmann distribution is only a special case. The possibility of choosing different distributions makes this formulation applicable to several domains, but it is imperative to keep the application context in mind in order to understand the meaning of the measures.

**Author Contributions:** Conceptualization, J.N. and C.D.M.; Formal analysis, J.N. and M.P.; Project administration, C.D.M.; Supervision, C.D.M.; Writing—original draft, J.N.; Writing—review & editing, I.Á., V.B.T., M.P. and C.D.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research was funded by Fundação de Amparo à Pesquisa do Estado de São Paulo grant number 2018/19150-6, Conselho Nacional de Desenvolvimento Científico e Tecnológico grant number 465755/2014-3 and Fundação de Amparo à Pesquisa do Estado de São Paulo grant number 2014/50851-0.

**Data Availability Statement:** Since this is a theoretical paper there is no data to be available.

**Acknowledgments:** This work was partially supported by the following agencies: FAPESP 2014/50851-0, CNPq 465755/2014-3 and BPE Fapesp 2018/19150-6.

**Conflicts of Interest:** The authors declare no conflict of interest.

#### References

- Cover, T.M.; Thomas, J.A. *Elements of Information Theory*; John Wiley & Sons: Hoboken, NJ, USA, 2012.
- Graven, A.; Keller, G.; Warnecke, G. *Entropy*; Princeton University Press: Princeton, NJ, USA, 2014; Volume 47.
- Wehrl, A. General properties of entropy. *Rev. Mod. Phys.* **1978**, *50*, 221. [[CrossRef](#)]
- Shannon, C.E. A mathematical theory of communication. *Bell Syst. Tech. J.* **1948**, *27*, 379–423. [[CrossRef](#)]
- Demirel, Y.; Gerbaud, V. *Nonequilibrium Thermodynamics: Transport and Rate Processes in Physical, Chemical and Biological Systems*; Elsevier: Amsterdam, The Netherlands, 2019.
- De Martino, A.; De Martino, D. An introduction to the maximum entropy approach and its application to inference problems in biology. *Heliyon* **2018**, *4*, e00596. [[CrossRef](#)] [[PubMed](#)]
- Caro, J.A.; Valentine, K.G.; Wand, A.J. Role of Conformational Entropy in Extremely High Affinity Protein Interactions. *Biophys. J.* **2018**, *114*, 67a. [[CrossRef](#)]
- Demirel, Y. Information in biological systems and the fluctuation theorem. *Entropy* **2014**, *16*, 1931–1948. [[CrossRef](#)]
- Brooks, D.R.; Wiley, E.O.; Brooks, D. *Evolution as Entropy*; University of Chicago Press: Chicago, IL, USA, 1988.
- Maldacena, J. Black hole entropy and quantum mechanics. *arXiv* **2018**, arXiv:1810.11492.
- Xiao, M.; Du, P.; Horne, K.; Hu, C.; Li, Y.R.; Huang, Y.K.; Lu, K.X.; Qiu, J.; Wang, F.; Bai, J.M.; et al. Supermassive Black Holes with High Accretion Rates in Active Galactic Nuclei. VII. Reconstruction of Velocity-delay Maps by the Maximum Entropy Method. *Astrophys. J.* **2018**, *864*, 109. [[CrossRef](#)]
- Bousso, R. Black hole entropy and the Bekenstein bound. *arXiv* **2018**, arXiv:1810.01880.
- Zeeshan, A.; Hassan, M.; Ellahi, R.; Nawaz, M. Shape effect of nanosize particles in unsteady mixed convection flow of nanofluid over disk with entropy generation. *Proc. Inst. Mech. Eng. Part E J. Process Mech. Eng.* **2017**, *231*, 871–879. [[CrossRef](#)]

14. Rostaghi, M.; Azami, H. Dispersion entropy: A measure for time-series analysis. *IEEE Signal Process. Lett.* **2016**, *23*, 610–614. [[CrossRef](#)]
15. He, D.; Wang, X.; Li, S.; Lin, J.; Zhao, M. Identification of multiple faults in rotating machinery based on minimum entropy deconvolution combined with spectral kurtosis. *Mech. Syst. Signal Process.* **2016**, *81*, 235–249. [[CrossRef](#)]
16. Degaetano-Ortlieb, S.; Teich, E. Modeling intra-textual variation with entropy and surprisal: Topical vs. stylistic patterns. In Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Vancouver, BC, Canada, 10 August 2017; pp. 68–77.
17. Reynar, J.C.; Ratnaparkhi, A. A maximum entropy approach to identifying sentence boundaries. In *Proceedings of the Fifth Conference on Applied Natural Language Processing*; Association for Computational Linguistics: Stroudsburg, PA, USA, 1997; pp. 16–19.
18. Campbell, J. *Grammatical Man: Information, Entropy, Language, and Life*; Simon and Schuster: New York, NY, USA, 1982.
19. Tame, J.R. On Entropy as Mixed-Up-Ness. In *Approaches to Entropy*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 153–170.
20. Adami, C. What is information? *Philos. Trans. R. Soc. A Math. Phys. Eng. Sci.* **2016**, *374*, 20150230. [[CrossRef](#)] [[PubMed](#)]
21. Kovalev, A.V. Misuse of thermodynamic entropy in economics. *Energy* **2016**, *100*, 129–136. [[CrossRef](#)]
22. Hayflick, L. Entropy explains aging, genetic determinism explains longevity, and undefined terminology explains misunderstanding both. *PLoS Genet.* **2007**, *3*, e220. [[CrossRef](#)] [[PubMed](#)]
23. Morowitz, H. Entropy and nonsense. *Biol. Philos.* **1986**, *1*, 473–476. [[CrossRef](#)]
24. Martyushev, L. Entropy and entropy production: Old misconceptions and new breakthroughs. *Entropy* **2013**, *15*, 1152–1170. [[CrossRef](#)]
25. Henderson, L. The von Neumann entropy: A reply to Shenker. *Br. J. Philos. Sci.* **2003**, *54*, 291–296. [[CrossRef](#)]
26. Sozbilir, M. What students' understand from entropy?: A review of selected literature. *J. Balt. Sci. Educ.* **2003**, *2*, 21–27.
27. Wright, P. Entropy and disorder. *Contemp. Phys.* **1970**, *11*, 581–588. [[CrossRef](#)]
28. Schrodinger, E. Order, disorder and entropy. In *Modern Systems Research for the Behavioral Scientist*; Aldine: Chicago, IL, USA, 1968; pp. 143–146.
29. Soubane, D.; El Garah, M.; Bouhassoune, M.; Tirbiyine, A.; Ramzi, A.; Laasri, S. Hidden Information, Energy Dispersion and Disorder: Does Entropy Really Measure Disorder? *World* **2018**, *8*, 197–202. [[CrossRef](#)]
30. Erlichson, H. Sadi Carnot, Founder of the Second Law of Thermodynamics'. *Eur. J. Phys.* **1999**, *20*, 183. [[CrossRef](#)]
31. Clausius, R. *On the Motive Power of Heat, and on the Laws Which Can Be Deduced from It for the Theory of Heat*; Annalen der Physik: Dover, NY, USA, 1960.
32. Blinder, S. Mathematical methods in elementary thermodynamics. *J. Chem. Educ.* **1966**, *43*, 85. [[CrossRef](#)]
33. Boltzmann, L. The second law of thermodynamics. In *Theoretical Physics and Philosophical Problems*; Springer: Berlin/Heidelberg, Germany, 1974; pp. 13–32.
34. Hoffmann, H.J. Energy and entropy of crystals, melts and glasses or what is wrong in Kauzmann's paradox? *Mater. Werkst.* **2012**, *43*, 528–533. [[CrossRef](#)]
35. Jaynes, E.T. Gibbs vs. Boltzmann entropies. *Am. J. Phys.* **1965**, *33*, 391–398. [[CrossRef](#)]
36. Jones, D.S. *Elementary Information Theory*; Oxford University Press: New York, NY, USA, 1979.
37. Tribus, M.; McIrvine, E.C. Energy and information. *Sci. Am.* **1971**, *225*, 179–188. [[CrossRef](#)]
38. Williams, P.L.; Beer, R.D. Nonnegative decomposition of multivariate information. *arXiv* **2010**, arXiv:1004.2515.
39. Kolmogorov, A.N. On tables of random numbers. *Sankhyā Indian J. Stat. Ser. A* **1963**, *25*, 369–376. [[CrossRef](#)]
40. Teixeira, A.; Matos, A.; Souto, A.; Antunes, L. Entropy measures vs. Kolmogorov complexity. *Entropy* **2011**, *13*, 595–611. [[CrossRef](#)]
41. Zenil, H. Towards Demystifying Shannon Entropy, Lossless Compression and Approaches to Statistical Machine Learning. *Proceedings* **2020**, *47*, 24. [[CrossRef](#)]
42. Baez, J.; Stay, M. Algorithmic thermodynamics. *Math. Struct. Comput. Sci.* **2012**, *22*, 771–787. [[CrossRef](#)]
43. Zenil, H.; Kiani, N.A.; Tegnér, J. The thermodynamics of network coding, and an algorithmic refinement of the principle of maximum entropy. *Entropy* **2019**, *21*, 560. [[CrossRef](#)]
44. Zenil, H.; Kiani, N.A.; Tegnér, J. Low-algorithmic-complexity entropy-deceiving graphs. *Phys. Rev. E* **2017**, *96*, 012308. [[CrossRef](#)] [[PubMed](#)]
45. Zenil, H.; Kiani, N.; Abrahão, F.; Tegner, J. Algorithmic Information Dynamics. *Scholarpedia* **2020**, *15*, 53143. [[CrossRef](#)]
46. Zenil, H.; Kiani, N.A.; Marabita, F.; Deng, Y.; Elias, S.; Schmidt, A.; Ball, G.; Tegnér, J. An algorithmic information calculus for causal discovery and reprogramming systems. *Isience* **2019**, *19*, 1160–1172. [[CrossRef](#)] [[PubMed](#)]
47. Zenil, H.; Hernández-Orozco, S.; Kiani, N.A.; Soler-Toscano, F.; Rueda-Toicen, A.; Tegnér, J. A decomposition method for global evaluation of shannon entropy and local estimations of algorithmic complexity. *Entropy* **2018**, *20*, 605. [[CrossRef](#)] [[PubMed](#)]
48. Leff, H.S. What if entropy were dimensionless? *Am. J. Phys.* **1999**, *67*, 1114–1122. [[CrossRef](#)]
49. Bernard, T.N.; Shi, E.L.; Gentle, K.; Hakim, A.; Hammett, G.W.; Stoltzfus-Dueck, T.; Taylor, E.I. Gyrokinetic continuum simulations of plasma turbulence in the Texas Helimak. *arXiv* **2018**, arXiv:1812.05703.
50. Bagryansky, P.; Shalashov, A.; Gospodchikov, E.; Lizunov, A.; Maximov, V.; Prikhodko, V.; Soldatkina, E.; Solomakhin, A.; Yakovlev, D. Threefold increase of the bulk electron temperature of plasma discharges in a magnetic mirror device. *Phys. Rev. Lett.* **2015**, *114*, 205001. [[CrossRef](#)]

51. Maxwell, J.C.; Pesic, P. *Theory of Heat*; Courier Corporation: North Chelmsford, MA, USA, 2001.
52. Szilard, L. Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen. *Z. Für Phys.* **1929**, *53*, 840–856. [[CrossRef](#)]
53. Landauer, R. Irreversibility and heat generation in the computing process. *IBM J. Res. Dev.* **1961**, *5*, 183–191. [[CrossRef](#)]
54. Sagawa, T.; Ueda, M. Fluctuation theorem with information exchange: Role of correlations in stochastic thermodynamics. *Phys. Rev. Lett.* **2012**, *109*, 180602. [[CrossRef](#)] [[PubMed](#)]
55. Jaynes, E.T. Information theory and statistical mechanics. *Phys. Rev.* **1957**, *106*, 620. [[CrossRef](#)]
56. Klauder, J.R.; Skagerstam, B.S. *Coherent States: Applications in Physics and Mathematical Physics*; World Scientific: Singapore, 1985.
57. Callen, H.B. Thermodynamics and an Introduction to Thermostatistics. *Am. J. Phys.* **1998**, *66*, 164. [[CrossRef](#)]