Entropy and Information

James Wills

1 Introduction

Entropy is a remarkably multi-faceted physical quantity. From its beginnings in thermodynamics, related concepts have since been imported into many different fields such as statistical mechanics, information theory, dynamical systems theory, computation theory and quantum theory. Academic interest in information has also been growing over the last few decades and is widely seen as playing a crucial role in our understanding of the world and our relation to it. With the development of these two concepts in parallel, their interconnections have purported to reveal interesting and surprising things about the world. This article will explore some major topics in entropy and information and the various natures of their connection.

This article adopts a quasi-historical approach to the subject, tracing the beginnings, development and intersection of the two concepts across time. Therefore, we begin with entropy in thermodynamics, its original incarnation, before moving on to discussing entropy in statistical mechanics (Boltzmann's and Gibbs'). Attempts to reduce or explain macroscopic thermodynamic behaviour in terms of the underlying microscopic mechanics of molecules led to various proposed definitions of entropy in statistical mechanics. It is here where hints of entropy's connection with information first make themselves visible. We then move on to discuss *Shannon information*, a precisely defined mathematical quantity in the theory of communication, which bears great formal and conceptual similarities with entropy in statistical mechanics. Up until the development of a precise mathematical characterisation of information given to us by communication theory, the concept of information employed has been the rough, ordinary language sense of information as something we learn or the thing by which we increase our knowledge. Therefore, it is with the Shannon information measure that we are able to really assess precise formal and conceptual links between entropy and information. A major contribution to this project was put forward by Edwin Jaynes, who proposed a new way of looking at classical statistical mechanics which puts Shannon information at its foundations. Further developments in this direction were put forward by Rolf Landauer in the 1960s in the context of the theory of computation. He proposed that there is an unavoidable entropy production in the processing of information by computers. The article concludes with more modern and current research topics which explore entropy and information in quantum theory and quantum computation.

2 Entropy in thermodynamics

Entropy was first introduced into physics in the 19th century by Clausius in the context of thermodynamics. In order to explore thermodynamic entropy's connection with information, we need to run through a brief development of the concept, following Clausius' reasoning. It begins with the following observation, which he called the *Second Main Principle of the Mechanical Theory of Heat*:

> "A passage of heat from a colder to a hotter body cannot take place without compensation" (Clausius 1879, p. 78)

This sentence expresses the fact that heat transfer from a hotter to colder body takes place spontaneously; one must perform work in order to transfer heat from a colder to a hotter body[1]. This observation may be seen as a restriction on the kinds of processes we may use to move heat around and extract work from heat sources. To develop the mathematical and physical implications of this observation, we consider a process which can be used to extract work from the transfer of heat from hot to cold and which can consume work to move heat from a colder body to a hotter one. This process is called a *Carnot cycle*, named after Sadi Carnot who first introduced the idea in his work *Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance* (Carnot 1890). For a classic and more modern reference, see Clausius (1879, pp. 69–89) and Callen (1960, pp. 77-79).

By considering the heat transferred between the hot and cold bodies and the heat converted into work along the cycle, we arrive at the following result:

$$\oint \frac{dQ}{T} = 0. \tag{1}$$

This says that the integral of the function dQ/T along a reversible[2] cycle is zero and is true if and only if the Second Main Principle is true. This means that dQ/T is a *state function*: a function dependent only on the thermodynamic state of the system. We can therefore define the state function *entropy*[3]:

$$dS = \frac{dQ}{T}. \tag{2}$$

This concludes our brief conceptual and mathematical development of the foundations of entropy in thermodynamics. From this, one can go on to show many interesting consequences, including the Second Law of thermodynamics, but that is not our focus here

[1] This fact is known as the *Clausius statement* of the Second Law of thermodynamics, to distinguish it from the *Kelvin statement* which states that it is impossible to completely convert heat into work. The two statements can be shown to be equivalent (see Blundell and Blundell (2010, p. 131)). Here, I leave Kelvin's statement aside and focus only on Clausius reasoning.

[2] A process from an initial state to a final state is *reversible* just in case there is a process which can recover the initial state of the system and the environment. See Uffink (2001) which contains a detailed discussion of the historical and conceptual development of reversibility in thermodynamics.

[3] Clausius coined the term after the Greek word for 'transformation' $\tau\rho o\pi\tau\eta$ since, roughly speaking, it assigns a value to the transformation of heat into work and vice versa. Clausius' arguments to this effect can be found in Clausius (1879, pp. 91–107).

(for a thorough and rigorous conceptual and mathematical development, see Clausius (1879)). We have seen how it is defined mathematically and the principle on which its existence depends. This was sufficient to see that the thermodynamic entropy has no connection with information at all; the thermodynamic entropy came into being solely off the back of observations about the relations between heat and work and way heat moves between hot and cold bodies.

The link with information comes from the definitions of entropy in statistical mechanics which claim to explain the behaviour of the thermodynamic entropy in terms of the dynamics of the microconstituents of matter. We therefore move on to explain and discuss definitions of entropy in statistical mechanics and their link to information.

3 Entropy in Statistical Mechanics

The project of statistical mechanics is to account for macroscopic thermal phenomena in terms of the dynamics of the microscopic constituents of matter. In particular, the search for the statistical analogue of the thermodynamic entropy has guided and continues to guide much research in this area, although the Second Law is definitely not the only focus. See Uffink (2007) and Frigg (2008) for recent overviews of the research happening in the foundations of statistical mechanics. The purpose of this section is to examine the purported link between information and Boltzmann's and Gibbs' statistical mechanical definitions of entropy.

3.1 Boltzmann entropy

Boltzmann's statistical mechanical definition of entropy was motivated by attempts to account for the Second Law of thermodynamics in terms of the mechanics of molecules. We do not need to directly assess how successful this endeavour was in order to examine his entropy's link with information. In order to introduce Boltzmann's definition and his argument for it, we need to introduce some formalism. Boltzmann's argument takes place in μ-space: the 6-dimensional single particle phase space whose points are of the form $(\mathbf{x}, \mathbf{p})$ where $\mathbf{x} = (x, y, z)$ is the position of the particle and $\mathbf{p} = (p_x, p_y, p_z)$ is its momentum. A particle with a particular momentum at a particular position is represented as a point in the μ-space. A system consisting of N particles will then be represented by N points distributed in the μ-space. We can partition the μ-space into discrete cells indexed with $i \in \mathbb{N}$. The number of particles in cell i is denoted by $n_i = f(\mathbf{x}, \mathbf{p}) d^3\mathbf{x} d^3\mathbf{p}$ where f is the *distribution function*, denoting the number of particles per unit volume of μ-space. The system satisfies $\sum_i n_i = N$.

A *distribution* specifies the number of molecules in each cell. For example, if there are n_i particles in cell i, we may denote this by a tuple $D = (n_1, n_2, ..., n_i, ...n_k)$. A *complexion* specifies which particles are in which cells. For example, in a system of two cells and three particles, a distribution might be $D = (2, 1)$ while one possible complexion for this distribution might be particles A and B in cell 1 and particle C in cell 2 and another different complexion might be particles B and C in cell 1 and particle A in cell

2. It follows that the number of complexions corresponding to a distribution is given by the following expression:

$$P_D = \frac{N!}{n_1!n_2!...n_k!}. \tag{3}$$

where '!' denotes factorials, i.e. $x! := x(x-1)...1$, for any natural number x and $0! := 1$. We shall drop the subscript D in most of what follows unless it is important to make it explicit.

The permutability measures how many ways a particular distribution can be achieved. For example, if the distribution is $(3, 0)$ then $P = 3!/3!0! = 1$; this tells us that there is only one way of scattering the three particles over the two cells in μ-space such that they all land in cell 1. If the distribution is $(2, 1)$, then $W = 3!/2!1! = 3$; this tells us that there are three ways of scattering the three particles over the two cells in μ-space such that two of them land in cell 1 and one of them lands in cell 2.

This forms the basis for Boltzmann's 1877 combinatoric argument that entropy is related to the quantity P. One can find written in numerous classic and modern sources[4] the following definition of the Boltzmann entropy:

$$S = k \ln P \tag{4}$$

where k is the Boltzmann constant. Despite this formula being on his tombstone, Boltzmann did not write it down. In fact, he substituted in the definition of P, dropped an additive constant[5] and finally arrived at:

$$\Omega = -\int f \ln f \; \mathrm{d}\mu \tag{5}$$

where $d\mu = d^3\mathbf{x}d^3\mathbf{p}$ denotes an element of μ-space. It is this quantity, called the *permutability*, which he explicitly identified as the statistical mechanical analogue of the thermodynamic entropy up to a multiplicative constant. To what extent is successful is beyond the scope of this article. The formula for Ω was given the symbol H in Boltzmann's 1872 work which is famous for containing the H-theorem and the Boltzmann equation. We do not discuss this work here as it is tangential to our main goal but its legacy is greatly felt and discussion of it forms a large and important part of the literature in the foundations of statistical mechanics (see Uffink (2017) for a discussion).

The Boltzmann entropy has taken on a different form under more modern developments[6] of the concept. The idea is now to set the problem, not in the 6-dimensional single particle phase space μ-space but in the $6N$-dimensional N-particle phase space, Γ-space whose points take the form $(\mathbf{x}_1, \ldots, \mathbf{x}_N, \mathbf{p}_1, \ldots, \mathbf{p}_N)$. A coordinate in Γ-space therefore encodes the position and momentum of all N particles. These points represent *microstates*: the mechanical state (as specified by the position and momentum) of the N particle system. In the Boltzmann picture, Γ-space is equipped with a *Lebesgue measure*,

[4] For example, Rushbrooke (1949) and Blundell and Blundell (2010).

[5] The details of this are beyond the scope of this article. They can be found in a translation of Boltzmann's original paper: Sharp and Matschinsky (2015).

[6] See for example Goldstein (2001) and Lebowitz (1999).

μ (not to be confused with the μ in μ-space) which assigns sizes to sets of points of the phase space, in much the same way as we assign a set of points inside a box a volume measure. $\mu(A)$ denotes the measure of region A of the phase space. For some partitions of Γ-space, the elements of the partition will correspond to *macrostates*: $M_1, ..., M_n$ (where n is the number of elements in the partition) which are states of the system specified, not by the mechanical state of the N particles, but by macrovariables such as temperature, pressure or volume. In general, many different microstates can correspond to the same macrostate; this is expressed more precisely by saying that macrostates *supervene* on microstates; there cannot be a change in the former without a change in the latter. The regions $M_1, ..., M_n$ have measures $\mu(M_1), ..., \mu(M_n)$. We then define the entropy of a system in macrostate M_i as follows:

$$S(M_i) = k \ln[\mu(M_i)]. \tag{6}$$

So, we have two conceptually different notions of the Boltzmann entropy: Ω and $S(M_i)$. They are related, however: we can recover the expression for Ω from Equation 6. To see this, note that Γ-space is the Cartesian product of N copies of μ-space. Therefore, a partition of μ-space induces a partition on Γ-space, i.e. Γ-space will also be divided up into cells which do not overlap and which cover the whole space. If the volume of the cells in μ-space is $\delta\omega$, and denoting the set of points in Γ-space which correspond to the distribution D_i by Γ_{D_i}, then the measure of that set is given by:

$$\mu(\Gamma_{D_i}) = P_{D_i}(\delta\omega)^N \tag{7}$$

We may view the Γ_{D_i} as a concrete example of how to partition Γ-space. We can substitute Equation 7 into 6 to obtain:

$$S(\Gamma_{D_i}) = k \ln[P_{D_i}(\delta\omega)^N] = k \ln[P_{D_i}] + kN \ln[\delta\omega]$$

We can see that this is Equation 4 up to an additive constant. If we now substitute in the expression for P_{D_i} and use Stirling's approximation[7], we obtain:

$$S = -k \sum_{i=1}^{k} n_i \ln n_i$$

up to an additive constant depending on N and $\delta\omega$. Recalling that $n_i = f(\mathbf{x}, \mathbf{p}) d^3\mathbf{x} d^3\mathbf{p}$ from earlier, we can see that this is equal to Ω up to an additive constant. What this has shown is that $S(M_i) \approxeq \Omega$ (up to an additive constant) if the measure μ is given by Equation 7. The expression in Equation 6 is therefore a more general version of Equation 5 since it allows for an arbitrary measure. While the two definitions of the Boltzmann entropy are related for a specific choice of measure, they are in general conceptually distinct, since Ω is defined on μ-space and $S(M_i)$ on Γ-space. Given this conceptual difference, we consider each of their connections with information.

[7] For the details, see Frigg and Werndl (2011, p. 127).

We begin with Equation 5. First, observe that f decreases when the particles are more spread out over μ-space. At first, this does not seem to readily lend itself to an interpretation in terms of information but it can be done. Suppose the particles are maximally spread out over a volume V. f takes on a certain form. Then suppose that we compress the gas to $V/2$. Then the particles have less of the μ-space available to them and f will increase in some parts of μ-space. We can interpret what has happened in terms of information. To start with an analogy: suppose you are asked to guess the month of your friend's birthday. You have twelve choices. Then they give you a clue that the month starts with the letter 'J'. The number of choices you have for their birthday has decreased and so, intuitively at least, the information you have about when the birthday is has increased. In a similar way, when the volume is halved, the number of choices for where the particles are in the μ-space has decreased and so your information for where they are has increased[8]. We can also express this thought in terms of uncertainty. We are more uncertain about where a particle is if they are more spread out over μ-space than if they are confined to a smaller region of the space. The purported (qualitative) link between entropy and information, then, is this: when the f is defined over a larger region of μ-space, the information about where a particular particle of the system is in the μ-space decreases.

We now consider Equation 6's link with information. Similar reasoning applies to the previous case. Suppose we have a container of gas in macrostate $M = (V, T)$. This macrostate is compatible with many different microstates. The Lebesgue measure of this macrostate determines the entropy of the macrostate via the formula $S(M) = k \ln[\mu(M)]$. The larger the measure, the larger the entropy. This idea can be interpreted in terms of information. If we have two macrostates M_1 and M_2 and $\mu(M_1) > \mu(M_2)$, then we may say that we have more information about the microstate of M_2 than we do about M_1. Put another way, we are more uncertain about the microstate of M_1 than we are about that of M_2 precisely because M_2 has the larger measure.

While this does seem to make the link between Boltzmann entropy and information clearer and more precise, it is important to emphasise the limitations of this link. Information is an epistemic notion, typically associated with knowledge. This characterisation is inevitably rough and vague because information does not get a precise characterisation when discussing its links with the Boltzmann entropy. The most that can be said is that information is some kind of epistemic notion. But Ω is not a function of anything that can be interpreted as epistemic; it is a function of the actual distribution of the position and momentum of the particles over μ-space. Things get more interesting when we consider the more general definition of the entropy $S(M_i)$, since the entropy depends upon the choice of measure. The question then seems to be open about how the measure is determined: will it be determined on the basis of epistemic considerations or objective ones? That is, does the measure μ reflect our ignorance about the state of the system or does it reflect some objective feature of the state of the system? [**Roman: is there a classic reference I can use here?**] If we are to interpret the measure epistemically, then a case could be made that $S(M_i)$ measures the lack of information we have about

[8] Arguments similar to this can be found in Zemansky and Dittman (1997).

the state of the system.

3.2 Gibbs entropy

In his 1902 *Elementary Principles in Statistical Mechanics*, Gibbs set up a clear and general mathematical framework to think about macroscopic thermal phenomena from the point of view of the mechanics of the microscopic atoms making up the system. He then invites us to

> "imagine a great number of systems of the same nature, but differing in the configurations and velocities which they have at a given instant," (Gibbs 1902, p. iii)

This collection of systems is known as the *ensemble*; an enormous collection of copies of the system under study which contains all of the possible ways the actual system could be in[9]. It is the ensemble, not the actual system, that is the object of study in Gibbs' statistical mechanics:

> "And here we may set the problem, not to follow a particular system through its succession of configurations, but to determine how the whole number of systems will be distributed among the various conceivable configurations and velocities at any required time, when the distribution has been given for some one time. The fundamental equation for this inquiry is that which gives the rate of change of the number of systems which fall within any infinitesimal limits of configuration and velocity." (p. iii)

We are invited to imagine that this ensemble is distributed in a particular way among the possible mechanical states in Γ-space. This distribution tells us how many systems lie in a given region of the phase space. If we divide this number by the total number of systems in the ensemble, we get a probability distribution denoted by ρ. It is this probability distribution which will provide the link with information which we will explore later in this section. We then study how the probability distribution over the phase space evolves under the system's dynamics, in Gibbs' case, Hamiltonian dynamics. This is in contrast to Boltzmann's statistical mechanics, in which we study how a single system evolves under the dynamics. In general, ρ will change with time. There are however, special probability distributions which do not change with time under Hamiltonian evolution, those in *statistical equilibrium*, and Gibbs singles these out as worthy of special attention since they can be shown to give rise to equations which are analogous to equations from thermodynamics.

The first probability distribution he considers[10] in the so-called *canonical ensemble* which we denote here by ρ_c. By considering small changes in the constant values of the

[9] For other characterisations of ensembles, see for example Schrödinger (1989, p. 3) and Tolman (1979, p. 43).

[10] Gibbs studies two other probability distributions: the *microcanonical* and the *grand canonical* ensembles. We do not explore these in great detail here because it is the canonical distribution which is most often connected with information.

canonical distribution (see Gibbs (1902, pp. 43–44)), he derives an equation which bears great formal similarity to the fundamental equation of thermodynamics. The extent to which this reduces thermodynamics to statistical mechanics and the strength of the analogy between the statistical equation and the thermodynamic equation is of great interest in the debate concerning to what extent thermodynamics is reduced to Gibbs' statistical mechanics, but this is beyond the scope of this article. What is interesting for our purposes is what corresponds to the thermodynamic entropy. By formal analogy with the thermodynamic equation, the following expression, which Gibbs gives the symbol $\bar{\eta}$, is analogous to the thermodynamic entropy:

$$\bar{\eta} = -\int \rho_c \log \rho_c \, \mathrm{d}\Gamma \tag{8}$$

It should be read as the expectation value of $\eta = \log \rho_c$ in the canonical ensemble.

When considering the link of $\bar{\eta}$ with information, we run up against the same problem as when we considered the link of $S(M_i)$ with information. If we interpret the probability distribution ρ_c, and hence the entropy, epistemically[11], then the entropy plausibly reflects our lack of knowledge or information about the exact microstate of the system. Such interpretations are controversial since it is unclear how exactly to square such an epistemic interpretation with the objective physics Gibbs' statistical mechanics apparently gives us. On the other hand, if we interpret the probability distribution ontically so as to remove that problem, then it becomes unclear what objective feature of the system the probability distribution is meant to capture. This is just one aspect of the general problem of trying to make sense of Gibbs' statistical mechanics. See Frigg and Werndl (2018) for further discussion on this point.

4 Entropy in Information Theory

What we have heard about information so far has been qualitative and only imprecisely characterised; information is something which a source can have more or less of, may be conveyed from one source to another and may or may not impart knowledge. In what we have so far seen, certain mathematical objects may be seen as containing or imparting information in some sense: the distribution function f may be seen as encoding information about how many particles there are in a region of μ-space and the measure μ from Equation 6 and the Gibbsian probability distribution ρ_c may be seen as encoding information about which mechanical state our system is in. But if we are to make the link between entropy and information clear and precise we must make mathematically precise the concept of information.

This was done in 1948 by Claude Shannon in his article *A Mathematical Theory of Communication*. Shannon was concerned with what he called the 'fundamental problem

[11] There are broadly two ways of interpreting probabilities: epistemically or ontically. The latter understands probabilities as reflecting the state of an agent's knowledge while the former takes probabilities as reflecting some feature of the world. For a review of the various interpretations of probability see Hájek (2019) and references therein.

of communication': basically, to send a message from one point to another. In his paper, he made contributions concerning noisy channels and how to make savings based on the statistical nature of the message. What is important for our purposes is his introduction of a precise mathematical definition of information. It is important to emphasise that information in this context is meant in quite a technical sense but it does capture some features of the ordinary language sense of information. The feature it captures is the surprise when we get a message we were not expecting; the more surprising the message, the more informative it is. For example, we would say in an intuitive non-mathematical sense, that the proposition "It is raining or it is not raining" has no information content whatsoever because the proposition is a tautology, it is not surprising. Similarly, if you tell someone something they already know, then you are not imparting any information to them; what you tell them is not surprising. On the other hand, when we learn that Boltzmann never wrote down the formula that he is most famous for, we are surprised and the message is informative because of it.

To further illustrate the link between information and surprise, consider the following example. Suppose it is the night of a nation's democratic elections and the polls predict that Party A are going to beat Party B. As the votes are being counted, we do not actually know who is to win, but, going by the polls, we think it is more likely that Party A beats Party B. Suppose the results come in and Party B wins. This was not impossible, just less likely. Intuitively, we might say that we have gained more information learning that Party B wins than we would have done had Party A won because it was less likely that Party B wins and thus more surprising that Party B won. Thus, intuitively, the more surprising something is, the less expected it is, the more information we gain. The mathematical sense of information introduced by Shannon captures this intuition and makes it more precise.

The setup is as follows. We introduce the concept of an *information source* which is the probability distribution $\{p(m_1), \ldots, p(m_n)\}$ over a finite set of messages $\{m_1, \ldots, m_n\}$. Shannon then asks:

> "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?" (Shannon 1948, p. 392)

Shannon's key idea is that this measure is a function only of the probability of the message. We write this function $H(p_1, p_2, ..., p_n)$. He argues that it is reasonable for the information function to satisfy certain properties (see pp. 392–393):

1. *Continuity.* H should be continuous in the p_i.

2. *Monotonicity.* If all the 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. *Branching.* If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H.

Shannon shows that the only function[12] which satisfies these properties is:

$$H = -\sum_{i=1}^{i=n} p_i \log_2 p_i \tag{9}$$

For more discussion on these three properties and the proof which establishes this expression, see Shannon (1948, pp. 392–393 and 419–420). To understand how best to interpret H, it is helpful to consider an example of how this formula works using the simple example of a coin flip. If the probability distribution over the set of messages $\{H, T\}$ is $\{p(T) = 0, p(H) = 1\}$, then $H = 0$. Interpreted, this is saying that no information is produced by the source; the outcome is certain and not surprising to the receiver. On the other hand, if the probability distribution is $p(H) = p(T) = 0.5$ then $H = -0.5 \log 0.5 - 0.5 \log 0.5 = 1$. Interpreted, this is saying that the information source is producing the maximal amount of information. The outcome is maximally uncertain and surprising to the receiver.

We can interpret H as a measure of the receiver's average uncertainty about the message produced by the source[13]. If the message outcome is certain, we are not surprised by the outcome, we learn nothing and thus gain no information. If the message is uncertain then we do learn something and so gain some information. This interpretation as a measure of uncertainty is important when it comes to discussing entropy in dynamical systems theory in Section 4.3.

It is very important to emphasise that H is *not* a measure of the information contained in a particular message. In this respect, it does depart from ordinary usage somewhat since we often associate information with meanings of individual messages. In the context of information theory, H is the measure of information of a source, understood as a probability distribution, that can output a variety of possible messages. The reason for this departure from typical usage has to do with Shannon's overall aims in his 1948 paper; we want to consider together all the possible outputs of the source in order to work out the capacity of the communication channel required to transmit a sequence of messages. When given the probabilities of the possible messages, it is possible to reduce the capacity of the channel in order to transmit the required messages. This leads us on to our next point: the interpretation of H as the amount of information in a source.

The amount of information is measured in *bits*. This is a term found in Shannon (1948, p. 380) and is short for *binary digits* when the logarithm is base 2. A device which can be in two physical states (which we may denote 0 or 1), such as a switch, is said to store one bit of information. For an ensemble of N devices, the total number of possible states of the ensemble of N switches is 2^N so the number of bits that can be stored is given by $\log_2 2^N = N$.

[12] Here, we are only considering the discrete Shannon entropy. There is a continuous version: see Shannon and Weaver (1963, p. 87).

[13] Although we can interpret H as a measure of uncertainty, care has to be taken in this regard; Timpson (2013, §2.2.3) points out that the link between information and uncertainty is not so clean because H is not a unique measure of uncertainty while it is a unique measure of information.

Given an information source H tells us, roughly speaking, the number of bits required to transmit a given message. To see this, consider this example[14]. Suppose we have a long sequence of N messages from a source. The number of possible sequences of length N made up from n messages, where p_i is the probability (assumed to match the frequency of the message in the long sequence) of the source outputting message m_i, is:

$$W = \frac{N!}{n_1!n_2!\dots n_n!}$$

Now consider $\log_2 W$. Using Stirling's approximation ($\log x! \approxeq x \log x$ for large x) we find that $\log_2 W = NH$ where H is the Shannon entropy[15]. Thus the total number of possible (very long) sequences is given by $W = 2^{NH}$. Expressing the possible sequences in this way allows us to see that we only need NH bits to encode a message of length N messages. Since $0 \leq H \leq 1$ (where equalities with 0 and 1 hold in maximal certainty and uncertainty respectively (see the example with the coin flip earlier in this section)) one can use fewer bits (on average) than there are messages in the sequence to communicate the message[16]. In this sense, the Shannon entropy is a measure of the amount of information in a source or, put another way, by how much the sequence can be compressed and still recovered at the receiving end of the communication channel. This interpretation of H as a measure of the information of a source is important when we come to discuss quantum entropy and information in Section 7.

Having introduced this technical conception of information and seen the ways in which it can be interpreted, we are now in a position to discuss information theory's link with physics.

4.1 Link with entropy in statistical mechanics: Shannon

Equation 9 looks formally very much like the two expressions for statistical mechanical entropy (Equations 5 and 8) we have already seen. Shannon notes this formal similarity:

> "The form of H will be recognized as that of entropy as defined in certain formulations of statistical mechanics where p_i is the probability of being in cell i of its phase space. H is then, for example, the H in Boltzmann's famous H theorem." (Shannon 1948, p. 393)

Here, Shannon notes the formal similarity to Boltzmann's definition of entropy[17], Ω. In order to see the similarity more clearly, we need to manipulate the expression for Ω a bit in order to recast it in terms of probabilities. Boltzmann writes Ω in terms of f, the distribution function, where $f d\mu$ denotes the number of particles in a cell in μ-space.

[14] This is a rough sketch of Shannon's *noiseless coding theorem* (Shannon 1948, §9) found in Timpson (2013, pp. 21–22).

[15] Here is how we arrive at this result. $\log_2 W \approxeq N \log_2 N - \sum_i n_i \log_2 n_i$. Substituting in $n_i = Np_i$, recalling that $\sum_i p_i = 1$ and cancelling the terms in $N \log_2 N$, we achieve the desired result.

[16] For a more concrete example, see Shannon (1948, §11).

[17] This same formula was given the symbol H in Boltzmann's 1872 work on the H theorem which is what Shannon is referring to here. See Section 3.1.

Recalling that the number of particles in cell i is $n_i = fd\mu$, the probability p_i that a particle is found in cell i of μ-space is then n_i/N where N is the total number of particles. Writing $p = f/N$, we can recast Ω as a function of p:

$$\Omega(p) = -N \int p \log[p] d\mu.$$

which is equal to Equation 5 up to an additive constant depending on N. We may denote the entropy in Equation 5 by $\Omega(f)$ to make the contrast more explicit. In order to consider more precisely the link between $\Omega(p)$ and Shannon's H, we write $\Omega(p)$ in terms of a discrete probability distribution[18]:

$$\Omega(p) = -N \sum_i p_i \log p_i. \tag{10}$$

where p_i is the probability of finding a particle in cell i of the μ-space. Having done this, we can now interpret $\Omega(p)$ as the entropy of an information source. The set of messages contains elements of the form "there is a particle in cell i of μ-space" and the probability distribution is $p_i = n_i/N$ where n_i is the number of particles in cell i and N is the total number of particles in the system.

Although Shannon does not explicitly mention it, H can also be seen to be similar to Gibbs' definition of entropy of the canonical ensemble. The probabilities in Gibbs' $\bar{\eta}$ however have a very different interpretation to the probabilities occurring in $\Omega(p)$: they give the probability that a system in the ensemble is in a particular cell of Γ-space. We may make the expression for $\bar{\eta}$ discrete just like we did with $\Omega(p)$:

$$\bar{\eta} = -\sum_i p_i \log p_i \tag{11}$$

where p_i is the probability of finding a system in the ensemble in cell i of the Γ-space and the sum is over all the cells in Γ-space. This may now be interpreted as the entropy of an information source. The set of messages contains elements of the form "there is a system of the ensemble in cell i of Γ-space" and the probability distribution is $p_i = n_i/N$ where n_i is the number of systems in cell i and N is the total number of systems[19]. Because of the formal similarity to Boltzmann and Gibbs entropies, H is now known as the *Shannon entropy.*

In this section we have seen how the Boltzmann and Gibbs entropies may be viewed as entropies of information sources under certain interpretations of the probability distributions and the messages. But this does not tell us that the statistical mechanical

[18] We do this to avoid a difficulty peculiar to the continuous Shannon entropy: while the discrete entropy measures *absolutely* the uncertainty in the message, the continuous entropy measures it *relative to a coordinate system.* Thus, changing variables will in general change the entropy. It does not, however, change entropy differences. For more detail, see Shannon and Weaver (1963, pp. 90–91)

[19] This highlights a major interpretational difficulty of the Gibbs entropy: how are we to interpret the probabilities in Gibbs' statistical mechanics if they appear to be about an imaginary ensemble of systems? For a general discussion of probabilities in physics see Ben-Menahem and Hemmo (2012) and Beisbart and Hartmann (2011).

entropies are ‘really’ about information or ‘should’ be interpreted in the context of information theory. The first attempt to actually do this and fully incorporate information theory and statistical mechanics was by Edwin Jaynes, whose ideas are the subject of the next section.

4.2 Link with entropy in statistical mechanics: Jaynes

In his 1957 *Information Theory and Statistical Mechanics*, Jaynes offered a reinterpretation of statistical mechanics which put the Shannon entropy at its foundation. This may be regarded as the first precise attempt to spell out precisely what the connection is between information theory and statistical mechanics, first hinted at in Shannon’s work. On this, Jaynes writes:

> “[T]he development of information theory has been felt by many people to be of great significance for statistical mechanics, although the exact way in which it should be applied has remained obscure.”(Jaynes 1957, p. 621)

Jaynes’ application of information theory to statistical mechanics takes the form of a proposal that the information theory entropy should be seen as the starting point of statistical mechanics. The need for this radical reinterpretation of statistical mechanics, he argues, comes from the lack of consensus among physicists of how to derive the macroscopic laws from the mechanics of the microscopic atoms.

The basic problem in statistical mechanics is how to make predictions about bodies consisting of large numbers of atoms without solving the equations of motion for each of the atoms. This is simply because we cannot possibly know the position and momentum of the all atoms. The next best thing we can do is determine the probability that particles will have certain positions and momenta based on what we know about the macroscopic properties of the body. What we want is a way of finding a probability distribution over microstates which is unbiased and consistent with what we know about the macroscopic properties of the system:

> “our problem is that of finding a probability assignment which avoids bias, while agreeing with whatever information is given.” (p. 622)

Jaynes argues that this unique probability distribution is given to us by maximising the Shannon entropy subject to constraints given to us by the macroscopic state of the system:

> “It is now evident how to solve our problem; in making inferences on the basis of partial information we must use that probability distribution which has maximum entropy subject 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.” (p. 623)

The reason we maximise the entropy is as follows. Recall that it was possible to interpret H as a measure of the uncertainty one may have about the message coming out of a source. Jaynes therefore takes entropy and uncertainty to be synonymous. The more uncertain we are, the more spread out the probability distribution, the higher the value of H. Therefore, to find the probability distribution over microstates which is unbiased but consistent with the macroscopic constraints, we have to find the probability distribution which is maximally uncertain within the constraints; this is the only way not to bias some microstates over others.

The idea of maximising the Shannon entropy subject to constraints in order to obtain the most unbiased probability distribution consistent with the macroscopic properties has come to be known as the *Maximum Entropy Principle* (MEP). The consequences of this thinking is that statistical mechanics should no longer be regarded as a physical theory trying to derive, via some physical argument, the macroscopic laws of of thermodynamics from the microscopic laws governing the atoms, but as a method of statistical inference; a method for making predictions about a macroscopic system from incomplete information about the microstate of the system[20]. This position has come to be known more generally as *Objective Bayesianism*, defined as: "any position which holds that the strengths of one's beliefs should be representable by a probability function, from all those that satisfy constraints imposed by evidence, that is maximally equivocal." (Williamson 2010, p. 25).

However, a great deal of controversy surrounds the maximum entropy principle and maximum entropy methods more generally. For an introduction to some of the controversy surrounding the MEP see Uffink (1995, 1996) and Howson and Urbach (1993, pp. 276–288) and for a defence of objective bayesianism more generally, see Williamson (2010).

4.3 Link with entropy in dynamical systems theory: Kolmogorov-Sinai Entropy

Dynamical systems theory deals with the long-term behaviour of systems whose time evolution is determined by the system's state at a given point in time and a law, or some kind of state transition rule, which governs how this state evolves in time. Newtonian dynamics is a paradigm example of a dynamical system. The study of dynamical systems in the abstract rather than the study of particular dynamical systems has proven to be useful and interesting. One of the reasons for this is that many physical theories are of this form, and so the study of the more general, abstract form of these theories can help us understand the individual features of the particular theories. Another reason is that it can help us understand *chaotic systems*, understood here as systems exhibiting random behaviour[21]. But what does 'random' mean in this context? It is claimed that the *ergodic hierarchy* (see Frigg, Berkovitz, and Kronz (2016) for an overview) provides the correct

[20] Jaynes' ideas were developed by many others in various fields (Levine and Tribus 1979). See Tribus (1961) and Hobson (1971) for a development of Jaynes' ideas in physics.

[21] Another characterisation of a system's chaotic behaviour is its sensitive dependence on initial conditions; a small change in the initial conditions can lead to a big change in the state at a later time. This characterisation, while relevant to dynamical systems, is not relevant to our information and entropy discussion.

and precise concepts that allow us to adequately characterise random behaviour. One of the most important concepts in this heirarchy that is taken to be indicative of random behaviour is a positive *Kolmogorov-Sinai entropy.* The justification for this, is that it can be shown (see Frigg (2004)) to be equivalent, under certain plausible assumptions, to a generalised version of the Shannon entropy[22], which can in turn can be a measure for randomness; roughly speaking, it can be a measure of the uncertainty in a future outcome (the outcome being where the dynamical system will be). Thus, the Shannon entropy is the 'bridge' connecting the KS entropy and some precise notion of randomness namely: uncertainty in the future outcome.

To see more clearly the link with the Shannon entropy, we first introduce the basics of dynamical systems theory. A dynamical system is a triple $\langle \Gamma, \mu, \Phi_t \rangle$ where Γ is the phase space of the system and μ is a measure on the space assigning volumes to regions of the phase space[23]. Φ_t is an *automorphism* sometimes also called a *phase flow*; it tells us, given the state of the system at $t = 0$ (the initial condition), what the state will be in the future. The automorphism not only maps points to points but also subsets to subsets. We may partition Γ into regions $\alpha = \{\alpha_1, ..., \alpha_n\}$ such that all these regions taken together form the entire phase space. The system may start in any of the partitions and move between them under the phase flow.

The analogies with communication theory is as follows. The partition of the phase space $\alpha = \{\alpha_1, ..., \alpha_n\}$ corresponds to the set of possible messages of an information source. The (suitably normalised) measure of the partitions $\{\mu(\alpha_1), ..., \mu(\alpha_n)\}$ correspond to the probabilities of the messages. The automorphism Φ_t corresponds to the information source, since this 'generates' the 'messages' α_i. In a dynamical system, the partition the system is in at time t will in general depend on which partition the system was in at previous times, i.e. on the history of the system. In a similar way, the message from the information source at time t may depend on previous messages from the source at previous times.

With these analogies, one can establish a rigorous link between dynamical systems theory and communication theory. I summarise some important steps in the argument (the details can be followed in Frigg (2004, §4)). Define the *entropy of a partition*:

$$H(\alpha) = -\sum_{i=1}^{n} \mu(\alpha_i) \log \mu(\alpha_i) \tag{12}$$

Notice that its functional form is identical to the Shannon entropy (Equation 9), hence justifying calling it entropy. By incorporating technical assumptions about how the past history of the system affects where the system will be in future partitions, we arrive at the following definition of the *entropy of an automorphism*:

$$H_{\Phi_t} = \sup_{\alpha} H_{\Phi_t}(\alpha) \tag{13}$$

[22] It can also be shown to be related to *Lyapunov exponents* (via Pessin's theorem) and to *algorithmic complexity* of a sequence (via Brudno's theorem). See Frigg, Berkovitz, and Kronz (2016) for an overview of these and their relation to the ergodic heirarchy.

[23] We may assume the measure to be normalised: $\mu(\Gamma) = 1$. Then we can interpret the measure as a probability measure.

where $H_{\Phi_t}(\alpha)$ is the entropy of the automorphism with respect to partition α. The motivation behind this definition is that it corresponds to the Shannon entropy of an information source when we understand the phase flow as the source generating the messages. However, there is a disanalogy between communication theory and dynamical systems theory in this regard: in communication theory the messages are given to us as part of the definition of the information source whereas the 'source' Φ_t does not give us the 'messages' α_i; the partition of Γ is up to us to choose. We therefore do not want the entropy to depend on the choice of α because otherwise our definition of entropy might end up telling us more about α than about Φ_t. TIt is for this reason that the dependence on α is eliminated by taking the supremum of $H_{\Phi_t}(\alpha)$ over all finite measurable partitions. Then, through an equivalence theorem (Frigg 2004, p. 428), one can show that the entropy of the automorphism is equivalent to the KS entropy.

The importance of this result is that it allows us to carry over results about unpredictability and uncertainty in an information source over to dynamical systems. Recall that when the Shannon entropy is zero there is no uncertainty about the output of the source; the message from the information source is completely predictable. When the Shannon entropy is greater than zero then we are uncertain about the message to come out. This carries over to dynamical systems. If the KS entropy is greater than zero this can now be interpreted as the unpredictability in the system; we are not able to predict which partition the system will be in next. Even if we know at every past time which partition the system was in, we will not be able to predict where it will be in the future if the KS entropy is positve[24].

5 Maxwell's demon

We leave the context of information theory behind now and consider other contexts in which information, no longer necessarily construed in Shannon's sense, meets entropy. We now move on to discuss what is seen as classic intersection between the concepts of entropy and information: Maxwell's demon. This thought experiment is often seen as indicating a powerful and deep link between entropy and information[25] although we will see in these next two Sections that this view is far from uncontroversial. The setup goes as follows (Maxwell 1875, pp. 328–329):

Imagine a container of gas of uniform temperature and pressure divided by a partition equally into two portions A and B. It is a fact that in such a container, the molecules are distributed according to the Maxwell-Boltzmann distribution. The important feature of this distribution for our purposes is that it means that some particles will be moving much faster than average and others will be moving much slower than average. Now suppose that there is a a being (or, if it has nefarious aims, a 'demon') which has control

[24] This is not strictly true but it gets the right idea. The point is that it is unpredictable on average. There may be some points in time where the next partition is predictable. But even if you collect information for ever, there will be some times in the future where the next partition is unpredictable.

[25] See for example Blundell and Blundell (2010, p. 150) who write: "The Maxwell demon [...] beautifully illustrates the connection between entropy and information."

over a small shutter in the partition and uses the shutter to allow the fast molecules to pass from A to B and the slow molecules to pass from B to A. The demon will then have created a temperature gradient without apparently doing any work. This directly contradicts the Clausius statement of the Second Law of thermodynamics (see Section 2). Faced with this problem, we have two options:

1. We can allow exceptions to the Law. This would mean that the Law, strictly speaking, is false or is only true in limited situations.

2. We can try to save the Law. This would involve trying to explain why the demon cannot in principle affect an entropy decrease in the gas.

Let us first consider efforts to save the Law. These efforts by far make up the majority of responses to Maxwell's demon but come in various flavours. All these efforts, however, have this in common: they all attempt to exorcise Maxwell's demon by trying to locate the compensating entropy increase. The first clarification such responses tend to make is to specify what is meant by the 'system'. The system is not just the box of gas but must include the demon as well. This is because the Second Law refers to the entropy of an *isolated* system. The box of gas is clearly not isolated since it is in contact with the demon. However, the box of gas plus demon can be considered an isolated system. So, the question becomes: how can the demon provide the compensating entropy increase, so that, at the very least, the entropy of the system does not increase?

Typical answers involve appealing to information. The first part of the answer involves arguing that in order for the demon to be able to do its job, it must be some sort of computational device which is able to measure and store information about the position and momenta of the particles in the gas in order to decide whether to open the shutter or keep it closed. Let us now consider some specific responses to the demon.

Szilard (1929) argued that the entropy increase happens when the demon measures the position and momentum of the particles; acquiring this information requires dissipation of energy or producing entropy. This line of thought was later developed by Brillouin (1951) and Gabor (1961). Later on however Bennett in his 1982; 1987 showed that measurement could take place in principle without entropy production. The purported entropy increase therefore had to come from elsewhere. Bennett, using the ideas of Rolf Landauer, claimed to have exorcised the demon by arguing that it came from the erasure of information from the hardware storing the information.

Landauer's ideas connecting entropy and information are explored and discussed in more detail in Section 6. Here, we state Landauer's conclusion and explain how Bennett uses it to exorcise Maxwell's demon. Bennett argues: in order for the demon to decide whether or not to open the shutter, it must be able to store information about the position and momentum of particles and perform computations on this information. Once the demon has decided whether to open or close the shutter on the basis of the position and momentum of that particle, it erases that information and stores information about the next approaching particle. Landauer claimed to show that the erasure is necessarily accompanied by an entropy increase. What prevents the demon from breaking the Law, writes Bennett:

> “is not the making of a measurement (which in principle can be done reversibly) but rather the logically irreversible act of erasing the record of one measurement to make room for the next.” (Bennett 1982, p. 906)

This point of view has generated a great deal of debate: see Bennett (2003), Norton (2005), Norton (2013), and Earman and Norton (1998, 1999) and references therein for the details.

We now turn to the first option: allowing exceptions to the Law. Maxwell’s response to his own demon is an example:

> “This is only one of the instances in which conclusions which we have drawn from our experience of bodies consisting of an immense number of molecules may be found not to be applicable to the more delicate observations and experiments which we may suppose made by one who can perceive and handle the individual molecules which we deal with only in large masses.” (Maxwell 1875, p. 329)

The ‘conclusion’ Maxwell refers to is the Second Law; we have deduced from our observations of macroscopic bodies that spontaneous temperature gradients do not form. His point is that, if we stop thinking about macroscpoic bodies as homogeneous masses (as we do in thermodynamics) and start thinking about bodies in terms of the motion an interaction of the individual molecules, then our conclusions drawn from the former characterisation will not be applicable to the latter. In other words, the Second Law as applied to homoegeneous masses is true but, strictly speaking, the Second Law is false when applied to bodies consisting of a large number of molecules. Since bodies do in fact consist in a large number of atoms, the Second Law is strictly false and only appears true to us due to the very large number of molecules. The dynamics governing the atoms do in fact allow that temperature gradients may spontaneously form; this is known as *Poincaré recurrence.*

The literature on the connection between Maxwell’s Demon and information is enormous so there is not sufficient space to assess all the commentary. For an extremely comprehensive survey of the topic, see Leff and Rex’s 1990; 2002. We now turn to discuss Landauer’s Principle in the thermodynamics of computation, a key result used by Bennett in his analysis of Maxwell’s demon concerning the connection between entropy and information.

6 Landauer’s Principle

In his 1961 paper *Irreversibility and Heat Generation in the Computing Process*, Landauer developed the idea that the erasure of information from the memory of a computing machine gives rise to an entropy production. We think of a computer as a finite array of N binary elements which can hold information in the form of strings of 0 and 1. Computers work by performing logical operations on these values. The results of these operations are then stored in the binary elements of the computer. However, since the memory of a computer is finite, there must come a point where information no longer needed has to be erased in order for the computer to store the results of its logical operations. This erasure normally involves resetting the binary elements to one of the values, say 1, so that the computer can store information on these elements. It is this necessary erasure, performed by the logical operation RESTORE TO ONE, which Landauer argues necessarily produces entropy. His argument for this, in his own words, goes as follows:

> "We shall call a device *logically irreversible* if the output of a device does not uniquely define the inputs. We believe that devices exhibiting logical irreversibility are essential to computing. Logical irreversibility, we believe, in turn implies physical irreversibility, and the latter is accompanied by dissipative effects." (Landauer 1961, p. 186)

In other words, he claims that logical irreversibility is a necessary condition for computing. He then goes on to argue that logical irreversibility implies physical irreversibility, meaning that entropy is produced. This increase in entropy is made manifest as heat dissipation into the environment.

Let us look at the steps of the argument in more detail. Let us first examine his claim that computing devices are logically irreversible[26]. An example of a logically reversible device is one which negates the input; so if the input is 1 the output is 0 and vice versa. In this case, the output uniquely defines the input and the device is logically reversible. A simple example of a logically irreversible device would be one which takes the conjunction of two values; in this case, the device is not logically reversible because a result of 0 might have been obtained from the conjunction of 0 and 1 or the conjunction of 0 and 0.

The next step is to argue that this logical irreversibility implies physical irreversibility and hence an increase in entropy due to the dissipative effects. Suppose that we have an assembly of bits, all with value 0. This state, writes Landauer,

> "corresponds, by the usual statistical mechanical definition of entropy, $S = k \ln W$, to zero entropy. The degrees of freedom associated with the information can, through thermal relaxation, go to any one of 2^N states (for N bits in the assembly) and therefore the entropy can increase by $kN \ln 2$ as the initial information becomes thermalized." (p. 187)

Landauer does not explicitly state exactly what he means by W in this case but he seems to take it to mean the number of states available to the system. If only one state is available to the system, then $W = 1$ and so the entropy of the system has zero entropy. If all the bits are capable of taking either of their two values, ZERO or ONE, then there are 2^N states available to the system, hence the claim that the system can increase its entropy by up to $kN \ln 2$; the system can go from having only one state available to it

[26] In fact, Landauer does concede that there could exist logically reversible devices using a device called the *Toffoli gate* (Toffoli 1980), but that they would be practically useless. See Landauer (1961, p. 187). Although Bennett (1982) claims that (useful) reversible computing is in fact possible.

to having all possible states available to it. The process by which a system can increase its entropy in this way, Landauer calls “thermal relaxation” or “thermalization”.

Now we consider the reverse process; we go from a state in which each bit can take either value to a state in which each bit can only take one value, say 1. This is called the RESTORE TO ONE operation. This is exactly the kind of operation that would be required in any logically irreversible computer; information which is no longer needed for the computer program would be erased in order to make way for the output of the program. This erasure is performed by the RESTORE TO ONE operation. In this operation, the entropy of each bit has been reduced by $k \ln 2$. Given that the computer is a closed system, this reduction in entropy of each bit must appear somewhere as heat supplied to the surroundings. This dissipation of heat and resulting production of entropy is a physically irreversible process. Hence, logical irreversibility implies physical irreversibility. Landauer provides a specific example on Landauer 1961, p. 188.

It is worth emphasising here as a closing remark that the connection between entropy and information suggested by Landauer's work, has nothing to do with the connection suggested by Shannon's work. In Landauer's words:

> “Note that our argument here does not necessarily depend upon connections, frequently made in other writings, between entropy and information. We simply think of each bit as being located in a physical system, with perhaps a great many degrees of freedom, in addition to the relevant one.” (p. 187)

The ‘other writings’ he refers to is Shannon's work. We have seen that Shannon means a very particular thing by information. Although not explicitly defined, Landauer intends information to be the bit value or string of bit values. Neither is more correct than the other, they just mean different things by the same word.

By way of concluding this Section, recall what we concluded in Section 2 namely: the thermodynamic entropy does not rely on any information-related concept in its definition or motivation. Landauer's ideas have linked the thermodynamic entropy and information in a curious way although they have drawn a great deal of interest and controversy from the physical and philosophical community. See Ladyman et al. (2007), Maroney (2005), and Norton (2005) and references therein for further in-depth discussion of these issues.

7 Entropy and information in quantum theory

It remains to discuss entropy in quantum theory and its link with information. Developments in this field have happened relatively recently, with quantum theory getting its mathematical rigour with von Neumann's *Mathematical Foundations of Quantum Mechanics* in the beginnings of the twentieth century and then with the field of quantum information emerging towards the end. We cannot hope to do justice to the whole of this enormous and growing field. We will content ourselves with highlighting some of the major developments and issues.

The quantum analogue of entropy was introduced by von Neumann (1955) and is now known as the *von Neumann entropy* defined as:

$$S(\rho) = -\mathrm{Tr}(\rho \ln \rho) \tag{14}$$

In order to understand this formula, we need to introduce some formalism and definitions. $\rho = \sum_i p_i |\psi_i\rangle\langle\psi_i|$ is the *density operator* of the ensemble of states $\{|\psi_i\rangle\}$ with probabilities $\{p_i\}$ and describes the state of a quantum system. ρ is said to be *pure* if $p_i = 1$ for state i the system and *mixed* otherwise[27]. The density operator encodes the probability of an outcome of a measurement on the quantum system. Tr is the trace operation. The trace of operator A is defined as $\mathrm{Tr}A := \sum_n \langle n|A|n\rangle$ where $|n\rangle$ is an orthonormal basis of the Hilbert space. If ρ is maximally mixed, i.e. all the p_i are the same, then $S(\rho)$ takes on its maximum value. If ρ is pure it takes its minimum value[28].

Von Neumann arrives at his expression for the entropy by running a thermodynamic-style argument: he considers the cyclic transformation of the state of a quantum gas[29] and argues that, since entropy is a function only of the state of the gas as it is in thermodynamics, whatever expression it has, it must have value zero at the end of the cyclic transformation. Von Neumann showed that the only expression for the entropy which does this is given by $-\mathrm{Tr}\rho \ln \rho$. For the details of this argument, see von Neumann (1955, pp. 358–379) or Petz (2001). Von Neumann's correspondnace between his entropy and the thermodynamic entropy has been critised: see Hemmo and Shenker (2006) and references therein for this debate.

We can see that it bears great formal similarity to the statistical mechanical expressions for entropy and to the Shannon entropy. But this formal link is not sufficient to establish a conceptual one. Such a conceptual link, first given by Schumacher (1995), can be made between the von Neumann entropy and the Shannon entropy, thus putting the von Neumann entropy to use in information theory, giving rise to *quantum information theory.* For a detailed overview of quantum information theory, see Timpson (2013, Ch. 3). To motivate the idea of quantum information theory somewhat: a classical bit can be in one of two states: 0 or 1. These might be physically realised by some sort of electronic switch that can either be on or off. A quantum bit (*qubit*) can be in states denoted by $|0\rangle$ and $|1\rangle$. But since a linear combination of quantum states is also a quantum state, the fully general state of a qubit can be written as the linear superposition $\alpha|0\rangle + \beta|1\rangle$ where α and β are complex numbers whose squares are interpreted as the probability that the qubit is measured to be in state $|0\rangle$ or $|1\rangle$. Qubits may be physically realised by, for example, an atom's spin along an axis; the state of the atom is then either spin up or spin down or a linear superposition of the two. Therefore, while a classical bit can only be in one of two states, a qubit can occupy a continuum of states and so, in

[27] We can further disambiguate mixed states into *proper* and *improper* mixtures. The probabilities in the former can be given an ignorance interpretation while the latter cannot. This distinction is due to D'Espagnat (2018).

[28] Note that this corresponds to a property of the Shannon entropy we have already noted, namely that if all the probabilities are equal, the Shannon entropy is 1 and if one of the probabilities is equal to one and the the rest are zero, then the entropy is 0.

[29] This is a transformation of the quantum state of a gas such that the beginning and end states are identical.

a sense that can be made more precise, can contain and process more information. This is one of the reasons for the excitement around quantum information.

To make the link with entropy: Schumacher proved the quantum version of Shannon's *noiseless coding theorem* which, roughly stated, places limits on the compression of data. Just as the classical Shannon entropy gives us a measure of how much the output of the source may be compressed (see Section 4), so the von Neumann entropy analogously provides a measure in the case of quantum information.

8 Conclusion

This article has discussed the concepts of, and links between, entropy and information in the context of thermodynamics, statistical mechanics, dynamical systems theory, information theory, computation theory and quantum theory. What we have seen is the extremely multi-faceted and pervasive nature of the concept of entropy. Some of its extensions into other areas of study are not all that surprising; statistical mechanics, viewed as an attempt to reduce the theory of thermodynamics to the mechanics of atoms, is a natural home for an extension of the concept of thermodynamic entropy. But from statistical mechanics, entropy spread into a number of apparently unconnected domains. The first really striking example of this was in Shannon's work in communication theory, whose expression for the amount of information in a source bears uncanny formal similarity to the expressions for entropy in statistical mechanics. While it was possible, before the rise of information theory, to interpret the statistical mechanical entropies in terms of an every day concept of information, it also became possible retrospectively, to interpret the statistical mechanical entropies in terms of Shannon's technical concept of information. While these interpretations are related insofar as Shannon information does capture some aspects of everyday information, there are other features they do not share, and we would do well to clearly distinguish between these two interpretations of the statistical mechanical entropies, not least because the everyday sense of information is not at all precisely defined. We also saw Shannon's ideas making roads into quantum theory via the work of Schumacher which led to the field of quantum information theory. Yet another sense of information arose with the discussion of Landauer's principle and Bennett's use of it to exorcise Maxwell's demon. Landauer did not precisely define the concept of information he used, but it seems slightly more precise than the everyday concept but not quite so precise as Shannon's. Indeed Landauer, insofar as he defined information at all, explicitly distanced his idea of it from Shannon's. This discussion highlighted the link that Landauer's principle purportedly makes between information and the thermodynamic entropy, something that seems quite surprising, given that no hint of information was present in the development of the thermodynamic entropy.

Much philosophical and clarificatory work remains to be done. The senses in which the statistical mechanical entropies reduce the thermodynamic entropy is still very much an open question. Related to this are questions concerning the relation between the Boltzmann and Gibbs versions of statistical mechanics[30] and making the notion of re-

[30] See Frigg and Werndl (2019) for a recent proposal on this front.

duction more clear and precise. Then there is the question of the role of information theory in statistical mechanics, first proposed by Jaynes; can this be made into a successful and coherent reinterpretation of classical statistical mechanics? Finally, the controversy of the status of Landauer's principle and its connection with thermodynamics and Maxwell's demon is still live. All of this, together with the work that is being done in the emerging field of quantum information and computation, indicates much interesting work still to be done.

References

Beisbart, Claus, and Stephan Hartmann. 2011. *Probabilities in Physics.* Oxford University Press.

Ben-Menahem, Y., and M. Hemmo. 2012. *Probability in Physics.* The Frontiers Collection. Springer Berlin Heidelberg.

Bennett, C. H. 1982. "The thermodynamics of computation—a review." *International Journal of Theoretical Physics* 21, no. 12 (December): 905–940.

———. 1987. "Demons, Engines and the Second Law." *Scientific American* (November).

———. 2003. "Notes on Landauer's Principle, Reversible Computation, and Maxwell's Demon." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 34 (3): 501–510.

Blundell, Stephen J., and Katherine M. Blundell. 2010. *Concepts in Thermal Physics.* 2nd ed. Oxford University Press.

Boltzmann, L. 1877. "On the relation between the second law of the mechanical theory of heat and the probability calculus with respect to the theorems on thermal equilibrium." *Kais. Akad. Wiss. Wien Math. Natumiss. Classe* 76:373–435.

Brillouin, L. 1951. "Maxwell's Demon Cannot Operate: Information and Entropy. I." *Journal of Applied Physics* 22 (3): 334–337.

Callen, H.B. 1960. *Thermodynamics.* J. Wiley.

Carnot, S. 1890. *Reflections on the Motive Power of Heat and on Machines Fitted to Develop that Power.* Landmarks of science. J. Wiley.

Clausius, R. 1879. *The Mechanical Theory of Heat.* Translated by Walter R. Browne. Macmillan & Co.

D'Espagnat, B. 2018. *Conceptual Foundations Of Quantum Mechanics: Second Edition.* CRC Press.

Earman, John, and John D. Norton. 1998. "Exorcist XIV: The Wrath of Maxwell's Demon. Part I. From Maxwell to Szilard." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 29 (4): 435–471.

———. 1999. "Exorcist XIV: The Wrath of Maxwell's Demon. Part II. From Szilard to Landauer and Beyond." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 30 (1): 1–40.

Frigg, Roman. 2004. "In What Sense Is the Kolmogorov-Sinai Entropy a Measure for Chaotic Behaviour? Bridging the Gap between Dynamical Systems Theory and Communication Theory." *The British Journal for the Philosophy of Science* 55 (3): 411–434.

———. 2008. "A Field Guide to Recent Work on the Foundations of Statistical Mechanics." In *The Ashgate Companion to Contemporary Philosophy of Physics,* edited by Dean Rickles, 99–196. London: Ashgate.

Frigg, Roman, Joseph Berkovitz, and Fred Kronz. 2016. "The Ergodic Hierarchy." In *The Stanford Encyclopedia of Philosophy,* Summer 2016, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Frigg, Roman, and Charlotte Werndl. 2011. "Entropy: A guide for the perplexed." *Probabilities in physics,* 115.

———. 2018. "Can somebody please say what Gibbsian statistical mechanics says?" *arXiv preprint arXiv:1807.04218.*

———. 2019. "Statistical Mechanics: A Tale of Two Theories." *The Monist* 102, no. 4 (September): 424–438.

Gabor, D. 1961. "Light and Information," 1:109–153. Progress in Optics.

Gibbs, Josiah Willard. 1902. *Elementary Principles in Statistical Mechanics.* New Haven: Yale University Press.

Goldstein, Sheldon. 2001. "Boltzmann's approach to statistical mechanics." In *Chance in physics,* 39–54. Springer.

Hájek, Alan. 2019. "Interpretations of Probability." In *The Stanford Encyclopedia of Philosophy,* Fall 2019, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

Hemmo, Meir, and Orly Shenker. 2006. "Von Neumann's Entropy Does Not Correspond to Thermodynamic Entropy." *Philosophy of Science* 73 (2): 153–174.

Hobson, Arthur. 1971. *Concepts in Statistical Mechanics.* Gordon & Breach.

Howson, Colin, and Peter Urbach. 1993. *Scientific Reasoning: The Bayesian Approach.* Open Court.

Jaynes, E. T. 1957. “Information Theory and Statistical Mechanics.” *Phys. Rev.* 106 (4): 620–630.

Ladyman, James, Stuart Presnell, Anthony J. Short, and Berry Groisman. 2007. “The connection between logical and thermodynamic irreversibility.” *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 38 (1): 58–79.

Landauer, R. 1961. “Irreversibility and Heat Generation in the Computing Process.” *IBM Journal of Research and Development* 5, no. 3 (July): 183–191.

Lebowitz, Joel L. 1999. “Statistical mechanics: A selective review of two central issues.” *Reviews of Modern Physics* 71 (2): S346.

Leff, H., and A.F. Rex. 1990. *Maxwell's demon: entropy, information, computing.* A. Hilger.

———. 2002. *Maxwell's Demon 2 Entropy, Classical and Quantum Information, Computing.* CRC Press.

Levine, R.D., and M. Tribus. 1979. *The Maximum Entropy Formalism: A Conference Held at the Massachusetts Institute of Technology on May 2-4, 1978.* MIT Press.

Maroney, O. J. E. 2005. “The (Absence of a) Relationship Between Thermodynamic and Logical Reversibility.” *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 36 (2): 355–374.

Maxwell, J.C. 1875. *Theory of Heat.* Longmans, Green, & Co.

Norton, John. 2005. “Eaters of the Lotus: Landauer's Principle and the Return of Maxwell's Demon.” *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics* 36 (June): 375–411.

Norton, John D. 2013. “The End of the Thermodynamics of Computation: A No-Go Result.” *Philosophy of Science* 80 (5): 1182–1192.

Petz, Dénes. 2001. “Entropy, von Neumann and the von Neumann Entropy.” In *John von Neumann and the Foundations of Quantum Physics,* edited by Miklós Rédei and Michael Stöltzner, 83–96. Dordrecht: Springer Netherlands.

Rushbrooke, G S. 1949. *Introduction to Statistical Mechanics.* Oxford: Clarendon Press.

Schrödinger, Erwin. 1989. *Statistical Thermodynamics.* New York: Dover.

Schumacher, Benjamin. 1995. “Quantum coding.” *Phys. Rev. A* 51 (4): 2738–2747.

Shannon, C. E. 1948. “A Mathematical Theory of Communication.” *Bell System Technical Journal* 27 (3): 379–423.

Shannon, C.E., and W. Weaver. 1963. *The Mathematical Theory of Communication.* University of Illinois Press.

Sharp, Kim, and Franz Matschinsky. 2015. "Translation of Ludwig Boltzmann's Paper "On the Relationship between the Second Fundamental Theorem of the Mechanical Theory of Heat and Probability Calculations Regarding the Conditions for Thermal Equilibrium" Sitzungberichte der Kaiserlichen Akademie der Wissenschaften. Mathematisch-Naturwissen Classe. Abt. II, LXXVI 1877, pp 373-435 (Wien. Ber. 1877, 76: 373-435). Reprinted in Wiss. Abhandlungen, Vol. II, reprint 42, p. 164-223, Barth, Leipzig, 1909." *Entropy* 17 (4): 1971–2009.

Szilard, L. 1929. "On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings." *Zeitschrift für Physik* 53 (11–12): 840–856.

Timpson, Christopher Gordon. 2013. *Quantum Information Theory and the Foundations of Quantum Mechanics.* Oxford University Press.

Toffoli, Tommaso. 1980. "Reversible computing." In *International colloquium on automata, languages, and programming,* 632–644. Springer.

Tolman, Richard C. 1979. *The Principles of Statistical Mechanics.* Dover.

Tribus, Myron. 1961. *Information theory as the basis for thermostatics and thermodynamics.* D. van Nostrand Company.

Uffink, Jos. 1995. "Can the maximum entropy principle be explained as a consistency requirement?" *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 26 (3): 223–261.

———. 1996. "The constraint rule of the maximum entropy principle." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 27 (1): 47–79.

———. 2001. "Bluff Your Way in the Second Law of Thermodynamics." *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics* 32 (3): 305–394.

———. 2007. "Compendium of the Foundations of Classical Statistical Physics." In *Philosophy of Physics,* edited by Jeremy Butterfield and John Earman, 923–1074. Handbook of the Philosophy of Science. Amsterdam: North-Holland.

———. 2017. "Boltzmann's Work in Statistical Physics." In *The Stanford Encyclopedia of Philosophy,* Spring 2017, edited by Edward N. Zalta. Metaphysics Research Lab, Stanford University.

von Neumann, J. 1955. *Mathematical Foundations of Quantum Mechanics.* Translated by Robert T. Beyer.

Williamson, Jon. 2010. *In Defence of Objective Bayesianism.* Oxford University Press.

Zemansky, Mark W., and Richard H. Dittman. 1997. *Heat and Thermodynamics.* McGraw-Hill.