Explaining Thermodynamics: What remains to be done?

Wayne C. Myrvold
Department of Philosophy
The University of Western Ontario
wmyrvold@uwo.ca

July 12, 2019

Abstract

In this chapter I urge a fresh look at the problem of explaining equilibration. The process of equilibration, I argue, is best seen, not as part of the subject matter of thermodynamics, but as a *presupposition* of thermodynamics. Further, the relevant tension between the macroscopic phenomena of equilibration and the underlying microdynamics lies not in a tension between time-reversal invariance of the microdynamics and the temporal asymmetry of equilibration, but in a tension between preservation of distinguishability of states at the level of microphysics and the continual effacing of the past at the macroscopic level. This suggests an open systems approach, where the puzzling question is not the erasure of the past, but the question of how reliable prediction, given only macroscopic data, is ever possible at all. I suggest that the answer lies in an approach that has not been afforded sufficient attention in the philosophical literature, namely, one based on the temporal asymmetry of causal explanation.

1 Introduction

Early in the previous century, Josiah Willard Gibbs wrote,

> A very little study of the statistical properties of conservative systems of a finite number of degrees of freedom is sufficient to make it appear, more or less distinctly, that the general laws of thermodynamics are the limit towards which the exact laws of such systems approximate, when their number of degrees of freedom is indefinitely increased (Gibbs, 1902, 166).

From the vantage point of the twenty-first century, Gibbs' optimism may seem naïve, as the relation between thermodynamics and statistical mechanics continues to be a topic of philosophical discussion, with no consensus in sight.

In this chapter I will argue that Gibbs is right. The laws of thermodynamics have been satisfactorily explained on the basis of statistical mechanics, and, indeed, the seeds of the explanation were already present in Gibbs' work, as (by a not entirely surprising bit of good fortune) the relevant parts of classical statistical mechanics can be taken over with little ado into quantum statistical mechanics. The explanation of the laws of thermodynamics consists of finding appropriate statistical mechanical analogues of thermodynamic concepts and deriving relations between them that approximate the laws of thermodynamics for systems of many degrees of freedom.[1] Along the way, something interesting and subtle happens to the temporal irreversibility of the second law of thermodynamics. The statistical analogue of the second law of thermodynamics is, unlike its thermodynamic counterpart, *not* temporally asymmetric; in place of temporal asymmetry is a distinct asymmetry related, but not identical to it. This will be made clear in section 3, below. This removes what has been seen as the chief stumbling-block for the reduction of thermodynamics to statistical mechanics, namely, the *prima facie* tension between a temporally asymmetric second law of thermodynamics and time-reversal invariance of the underlying dynamics.

[1] I hope it goes without saying that the project of investigating behaviour asymptotically approached by systems with a finite number of degrees of freedom as the number of degrees of freedom is increased indefinitely is a distinct project from that of investigating an idealized system with infinitely many degrees of freedom, though, with care, the latter project might be informative about the former.

This doesn't mean that we're out of the woods, though. There still remains an important task, not yet fully accomplished, the task of explaining the process of equilibration in statistical mechanical terms. This involves both an explanation of the general tendency for systems out of equilibrium to relax to an equilibrium state unless maintained in a non-equilibrium state by an external influence, and an explanation of the paths to equilibrium and the rates at which those paths have been traversed. This task has, of course, been regarded as part of thermodynamics, and a principle of equilibration has been counted by some as a law of thermodynamics (Uhlenbeck and Ford, 1963; Brown and Uffink, 2001). Compared to the other laws of thermodynamics, a law to the effect that, left to themselves, systems tend to relax to equilibrium, is an outlier. Though long recognized as an important principle, it was a late entry to the list of laws of thermodynamics: it was not referred to as a law of thermodynamics until the 1960s, more than a century after Kelvin initiated talk of laws, or fundamental principles, of thermodynamics. There is also an important conceptual distinction between the equilibration principle and the other laws of thermodynamics. It is unique among the so-called laws of thermodynamics in that its formulation does not require to make a distinction between energy transfer as heat and energy transfer as work. It is also the chief locus of temporal asymmetry.

Though, of course, this is ultimately a matter of choice of terminology, and nothing more, it seems to be that it is helpful to highlight the differences between the equilibration principle and the more traditional laws of thermodynamics by restricting the scope of "thermodynamics" to something like what its founders intended, and to regard the equilibration principle, not as belonging to thermodynamics proper, but as a *presupposition* thereof. A payoff in conceptual clarity of this choice is that it will perhaps mitigate somewhat the tendency to conflate the equilibration principle with the second law.

2 What is thermodynamics?

In contemporary physical parlance, to speak of the *dynamics* of a system is to speak of the laws according to which its state evolves over time. This has given rise to a folk etymology for the term "thermodynamics," according to which thermodynamics should mean the dynamical laws governing heat transfer. This folk etymology is incor-

rect. Understanding the actual etymology of the term is not merely a matter of historical interest. Contrary to the impression that the folk etymology would give, thermodynamics is aptly named, as the term stems from, and highlights, the two concepts that are at the core of the subject.

The term *thermodynamics* is composed from the Greek words for *heat* and *power*. It refers to the study of the ways in which heat can be used to generate mechanical action and heat can be generated via mechanical means. The word's first appearance is in Part VI of Kelvin's "On the Dynamical Theory of Heat" (Thomson 1857, read before the Royal Society of Edinburgh on May 1, 1854). There he recapitulates what in 1853 he had called the "Fundamental Principles in the Theory of the Motive Power of Heat," now re-labelled "Fundamental Principles of General Thermo-dynamics." The context makes clear that the new term is intended to denote the study of the relations between mechanical action and heat. It is worth noting that the term "thermodynamics" had not been used in connection with the earlier investigations, by Fourier, Kelvin, and others, of the laws of heat transport.

The term "thermodynamics" flags the distinction that is at the heart of the subject, the distinction between two modes of energy transfer: as work, and as heat. This is not a distinction that belongs to fundamental physics. Pure mechanics, whether classical or quantum, employs the concept of energy, but not this distinction between ways in which energy can be transferred from one system to another.

The two laws identified by Kelvin as laws of thermodynamics crucially invoke the heat-work distinction, and cannot be formulated without it. The first law states that total energy is conserved, whether transferred as heat or work; the net change in the internal energy of a system during any process is the net result of all exchanges of energy as work or as heat. The second law, in any of its formulations, also invokes the distinction. This is obvious in the Clausius and Kelvin formulations.

> Heat can never pass from a colder to a warmer body without some other change, connected therewith, occurring at the same time (Clausius 1856, 86, from Clausius 1854, 488).

> *It is impossible, by means of inanimate material agency, to derive mechanical effect from any portion of matter by cooling it below the temperature of the coldest of the sur-*

rounding objects (Thomson 1853, 179; reprinted in Thomson 1882, 265).

There is also an entropy formulation of the second law, that says that, in any process, the total entropy of all systems involved in the process does not decrease. At first glance, this might seem not to depend on the distinction between energy transfer as work and energy transfer as heat. But recall the definition of thermodynamic entropy. The entropy difference between two thermodynamic states of a system is calculated by considering some thermodynamically reversible process that links the two states and the associated heat exchanges between the system and the external world; the entropy difference between the two states of the system is the integral of $đQ/T$ over any such process (which cannot depend on which reversible process is chosen for consideration, on pain of violation of the second law in Clausius or Kelvin form).

In addition to the two laws of thermodynamics identified as such by Kelvin, there is another, more basic law, called the *zeroth law*, on which the definition of temperature depends.[2] The zeroth law has to do with the behaviour of systems when brought into thermal contact. Thermal contact (taken as a primitive notion) between two systems is a condition under which heat may flow between the systems. Under conditions of thermal contact, heat may flow from one system to the other, or else there may be no heat flow, in which case the systems are said to be in thermal equilibrium with each other. This induces a relation between states of systems that can be put into thermal contact with each other, which holds between the states (whether or not the bodies are actually in thermal contact) if there would be no heat flow if the systems were brought into thermal contact. This is obviously a symmetric relation. If we take a system in equilibrium to be in thermal contact with itself, it is also a reflexive relation. It is, therefore, an equivalence relation if and only if it is transitive. The zeroth law states that the relation is, indeed, transitive. If the zeroth law holds, we can partition thermodynamic states of systems

[2]The phrase occurs in Fowler and Guggenheim (1939, 56), and became a textbook staple in the years that followed. This was not, however, its first occurrence. However, a few years earlier, in a note on terminology, Charles Galton Darwin (1936) mentions the "*zeroth* law of thermodynamics" in a way that suggests that he expects it to be familiar to his readers; it is brought up only to illustrate the use of the word "zeroth," in Darwin's estimation a "terrible hybrid." Sommerfeld (1956, 1) attributes the coinage to Fowler, when giving an account of Saha and Srivastava (1931, 1935).

into equivalence classes under this relation, which we may regard as the relation of being of the same temperature. The zeroth law requires the notion of thermal contact, and hence the notion of heat flow, and thus, like the first and second laws, requires the heat-work distinction for its formulation.

These three laws have been recognized as laws of thermodynamics, and presented as such, in textbooks since the 1930s. There is also a late-comer, more basic than the rest. This is what Brown and Uffink (2001) call the *Equilibrium Principle* (though perhaps *Equilibration Principle* would be better):

> An isolated system in an arbitrary initial state within a finite fixed volume will spontaneously attain a unique state of equilibrium (Brown and Uffink, 2001, 528).

To signal that this is more fundamental than the traditional zeroth, first, and second laws, Brown and Uffink label this the *minus first law*. They point out that a principle of this sort had been recognized as a law of thermodynamics earlier, by Uhlenbeck and Ford (1963, 5).[3]

The equilibration principle has often been conflated with the second law. The two are distinct, however. It is a consequence of the second law that, *if* an isolated system makes a transition from one thermodynamic state to another, the entropy of the final state will not be higher than that of the initial state. This does not entail that its behaviour will be that dictated by the equilibration principle. As far as the second law is concerned, there might be some set of distinct thermodynamic states of the same entropy that the system cycles through. Or else it might fail to equilibrate when isolated, remaining in some quasistable nonequilibrium state until some external disturbance triggers a slide towards equilibrium.

The equilibration principle is unlike the zeroth, first, and second laws. The others have to do with exchanges of energy between systems, and rely on a distinction between energy exchange as heat and energy exchange as work. The equilibration principle, on the other hand, has to do with the spontaneous behaviour of an isolated system, and, *ipso facto* has nothing at all to do with energy exchange, in any mode.

Should we regard the equilibration principle as a law of thermodynamics? A case can be made for not doing so, as refraining from doing

[3]Uhlenbeck and Ford called it the *zeroth law* to emphasize its priority over the first and second laws. We will continue to follow standard terminology in taking transitivity of thermal equilibrium to be the zeroth law.

so provides a neat separation of two sorts of theoretical investigations.

There is a tradition, which can be traced back to Maxwell[4] and which has undergone renewed interest in recent years, of thinking of thermodynamics as a *resource theory*.[5] That is, it is a theory that investigates how agents with specified powers of manipulation and specified information about physical systems can best use these to accomplish certain tasks. A theory of this sort involves physics, of course, because it is physics that tells us what the effects of specified operations will be. But not *only* physics; considerations not contained in the physics proper, having to do with specification of which operations are to be permitted to the agents, are brought to bear. On such a view, the work-heat distinction rests on a distinction between variables that the agent can keep track of and manipulate, and variables that are not amenable to such treatment. If we take this view, we would not expect to capture concepts such as heat, work, and entropy within physics proper, and would not expect there to be statistical mechanical analogues of these concepts definable in purely physical terms.

The equilibration principle is in a different category. Though it, too, requires for its formulation a distinction not found in the fundamental physics, a distinction between macrovariables, used to define the thermodynamic state, and microvariables, required to specify the complete physical state of a system, it doesn't require the heat-work distinction. Someone who (like Maxwell) holds that the distinction between heat and work vanishes when all limitations on knowledge and manipulation are removed would expect the zeroth, first, and second laws to make sense only in the presence of such limitations. The same cannot be said of the equilibration principle.

If we take the word *thermodynamics* in its originally intended sense, as the science of heat and work, then the equilibration principle is not a law of thermodynamics. Instead, since it delivers the equilibrium states with which thermodynamics deals, it is a *presupposition* of all thermodynamics. Though this is merely a terminological issue, regarding the scope of the term *thermodynamics*, there is something to be said for flagging the relation between the equilibration principle and the traditional laws of thermodynamics by saying that only the latter, not the former, comprise thermodynamics proper. And if we

[4]See Myrvold (2011).

[5]See del Rio et al. (2015) for a general framework for resource theories, Wallace (2016) and Bartolotta et al. (2016) for some recent work, and Goura et al. (2015) for a review.

take the scope of thermodynamics to comprise only the zeroth, first, and second laws, then the task of explaining thermodynamics in statistical mechanical terms is a much less daunting one. It is of this more modest goal that Gibbs speaks, in the quotation with which we started.

This, of course, leaves with the deep, important, and interesting task of explaining equilibration. This is a matter to which much valuable and interesting work has been, and continues to be, devoted. The literature on equilibration deserves more attention from philosophers than it has heretofore received. Seeing equilibration as not a matter of thermodynamics but, rather, a process that thermodynamics presupposes, helps one to view this work more clearly, as the enterprise of studying equilibration is thereby freed of extraneous thermodynamic concepts (and in particular, of the concept of entropy).

3 Explaining thermodynamics in terms of statistical mechanics

Statistical mechanics deals with systems composed of a large number of interacting subsystems, and examines their aggregate behaviour, eschewing a detailed description of the microstate of the system. In the decade from 1867 to 1877, the major figures working to lay the foundations that Gibbs would later call *statistical mechanics*—Maxwell, Kelvin, and others in Britain, Gibbs in the U.S., and Boltzmann on the continent—came to realize that what was to be recovered from statistical mechanics was not the laws of thermodynamics as originally conceived, but a modified version on which what the original version of the second law declares to be impossible should be regarded as possible but, when dealing with things on the macroscopic scale, highly improbable. Because of random fluctuations of molecules, on a given run a heat engine operating between two reservoirs might yield more work than the Carnot limit on efficiency permits, but, by the same token, it might also yield less than expected. What we can expect from statistical mechanics is some statement to the effect that we can't *predictably and reliably* exceed the Carnot efficiency.

This suggests that we will have to traffic in probabilities, and consider, in the classical context, properties of probability distributions on the phase spaces of classical systems, and, in the quantum context, of density operators.

Construing a macroscopic system as composed of a large number of molecules also motivates a reconsideration of the notion of an equilibrium state. Though systems may settle into a state in which macroscopically measurable quantities are not changing perceptibly, this state cannot be a state of quiescence at the microphysical level. At the level of individual molecules, a system in thermodynamic equilibrium is seething with activity, and the observed macroscopic repose is the net result of averaging over large numbers of rapidly changing microphysical parameters. There is, of course, no sharp line between macroscopic and microscopic, and, at the mesoscopic scale (illustrated by Brownian motion) the state into which a system settles is one in which some measurable parameters are continually fluctuating, with a stable pattern of fluctuations. Here, again, probabilistic considerations come into play; we want to say that large fluctuations of macroscopic parameters during a time scale typical of our observations are not impossible, but merely improbable. Considerations such as this suggest that, in statistical mechanics, what we should associate with the condition of equilibrium, is, not an unchanging state, but a stable probability distribution. The process of equilibration will then be one in which non-equilibrium probability distributions over initial conditions converge (in an appropriate sense) towards the equilibrium distribution.

We also need, in order to apply thermodynamic concepts to statistical mechanical systems, a way of partitioning energy exchange between systems into heat exchange and work. The standard way to do this is to treat certain parameters (think of, for example, the position of a piston) as exogenously given, not treated dynamically and not subject to probabilistic uncertainty. The Hamiltonian of a system may depend on such parameters, and so a change in an exogenous parameter can result in a change in the energy of the system. Change to the energy of a system due to changes in the exogenous parameters on which its Hamiltonian depends are to be counted as work done on or by the system, and all other changes of energy counted as heat exchanges.

In thermodynamics, *thermal states* play an important role. A system that has thermalized (relaxed to thermal equilibrium) is one from one can extract no energy as work; the only way to extract energy from it is via heat flow. In such a state, a system has a definite temperature, uniform throughout the system. When two thermalized systems are placed in thermal contact with each other, the expected heat flow is

from the hotter to the colder. These things continue to be true if we consider a number of systems at the same temperature; an assembly of thermal systems at the same temperature is also a thermal system at that temperature. There are arguments (see Maroney 2007 for a lucid exposition) that these considerations lead to the conclusion that the appropriate probability distributions to associate with thermal states are what Gibbs named *canonical distributions*. In the classical case, the canonical distribution has density function (with respect to Liouville measure),[6]

$$\rho_\beta(x) = Z^{-1} e^{-\beta H(x)}, \tag{1}$$

where x ranges over phase-space points, H is the Hamiltonian of the system, and Z a normalization constant chosen to make the integral of ρ_β over the accessible region of phase space equal to unity. The parameter β indicates the temperature of the thermal state, and is inversely proportional to the absolute (Kelvin) temperature.

For a quantum system, a canonical state is represented by the density operator,

$$\hat{\rho} = Z^{-1} e^{-\beta \hat{H}}. \tag{2}$$

We want an analogue of the second law of thermodynamics. The second law of thermodynamics is equivalent to the statement that no heat engine can operate with an efficiency exceeding the Carnot efficiency. We can state this as follows.

> If a system undergoes a cyclic process, exchanging heats Q_i with thermal systems at temperatures T_i (and exchanging no heat with any other system), and returning to its original state, then
>
> $$\sum_i \frac{Q_i}{T_i} \leq 0. \tag{3}$$
>
> Moreover, if the process is thermodynamically reversible, it can be run in the opposite direction, with heat exchanges $-Q_i$. This entails that, for such a process,
>
> $$\sum_i \frac{Q_i}{T_i} = 0. \tag{4}$$

[6] We're assuming that the system is confined to a bounded region of phase space or otherwise subjected to conditions that render this a normalizable distribution.

This, in turn, entails that, if two thermodynamic states a and b of the system can be connected by a reversible process, then the heat exchanges the system makes with thermal systems must be such that the sum of Q_i/T_i over all such exchanges is the same for any reversible process connecting the states. We can use this to define a state function S, the thermodynamic entropy, such that

$$S(b) - S(a) = \sum_i \frac{Q_i}{T_i}, \tag{5}$$

where the sum may be taken over any thermodynamically reversible process taking state a to state b. This uniquely determines the entropy difference of two thermodynamic states, as long as they can be connected by a thermodynamically reversible process.

It turns out that there is something analogous in the form of a theorem about probability distributions, a theorem that is provable in two versions, classical and quantum.

Since we're dealing with multiple systems that may interact, we have to consider probability distributions over the state space of the composite system. Given a probability distribution P_{AB} over the state space of a composite system consisting of disjoint subsystems A and B, we can define marginal distributions P_A and P_B as the restrictions of P_{AB} to the degrees of freedom of A and B, respectively. Now consider a system A, that interacts with a number of thermal systems $B_i, i = 1, \ldots, n$, at temperatures T_i. Suppose that at time t_0 there is no interaction between the system A and the thermal systems B_i, and that the probability distribution of the joint system consisting of A and the thermal systems B_i is such that there are no correlations between A and the thermal systems. Suppose that, between t_0 and a time t_1, A interacts with the B_is successively, possibly exchanging energy with them. During this time, the energy of A may also change via manipulation of exogenous variables, but the only heat exchanges are with the thermal systems B_i. We also assume that at time t_1 there is no interaction between A and the thermal systems. Let us take the process to a cyclic one, in the sense that the marginal probability distribution of A at t_1 is the same as at t_0. As the system A has interacted with the thermal systems B_i in the interim, we do not assume that at t_1 A is uncorrelated with them.

As we are dealing with probability distributions, the heats Q_i that the system A exchanges with the thermal systems B_i is not determined by the setup. Instead, there will be a probability distribution

over the energy exchanges. We can consider the *expectation values* of these energy exchanges. It is easy to show that, provided that A is uncorrelated with the B_is at t_0 and has the same marginal at t_0 and t_1, the expectation values $\langle Q_i \rangle$ of heat exchanges satisfy

$$\sum_i \frac{\langle Q_i \rangle}{T_i} \leq 0. \tag{6}$$

This means that, even if fluctuations might yield, on an individual run, a greater yield of work than permitted by the Carnot limit on efficiency of heat engines, one cannot consistently and reliably violate the Carnot limit. Suppose I operate a heat engine between two heat reservoirs, a hot one of temperature T_1 and a cooler one, of temperature T_2, and I operate the engine in a cycle, restoring the marginal distribution of the engine at the end of the cycle. Prior to the beginning of the cycle my engine is uncorrelated with the heat baths. Suppose my process is such that the expectation value of heat extracted from the hot bath is $\langle Q_1 \rangle$. Then the expectation value of work obtained must satisfy

$$\langle W \rangle \leq \left(1 - \frac{T_2}{T_1}\right) \langle Q_1 \rangle. \tag{7}$$

If the cycle is a reversible one, that is, if it can be run in the opposite direction with the signs of the expectation values reversed, we must have equality in (6) and (7). For a reversible cycle, the expectation value of the work obtained is proportional to the expectation value of heat extracted, with the factor of proportionality equal to the Carnot efficiency of a heat engine operating between reservoirs at temperatures T_1 and T_2.

Equation (6) can be thought of as a statistical mechanical analogue of the second law of thermodynamics. It should be stressed that it is a *theorem* of statistical mechanics, which follows from the stated conditions on probability distributions, that, at time t_0, the distributions of the systems B_i be canonical, with temperatures T_i, and uncorrelated with A.

In thermodynamics, the fact that relation (4) holds for *any* reversible process linking two thermodynamic states a, b permits us to define the entropy difference $S[b] - S[a]$, provided that there is at least one reversible process linking the two states, and this uniquely defines the state function S, up to an additive constant. Similarly, in statistical mechanics, the fact that (6) holds for arbitrary probability distributions P_A entails that there is a functional $S[P_A]$, that takes

probability distributions over the state of A as input and yields real numbers, with the following properties. If, at t_0, A is uncorrelated with thermal systems B_i, whose probability distributions are given by canonical distributions at temperatures T_i, and if A exchanges heat in the interval between t_0 and t_1 with thermal systems B_i and with nothing else, then the expectation values of these heat exchanges, $\langle Q_i \rangle$, satisfy[7]

$$S[P_A(t_1)] - S[P_A(t_0)] \geq \sum_i \frac{\langle Q_i \rangle}{T_i}. \tag{8}$$

Suppose, now that there is a reversible process connecting $P_A(t_1)$ and $P_A(t_0)$. In this context this means that there is a process taking $P_A(t_0)$ to $P_A(t_1)$ and another process starting at t_1 and ending at a time t_2 such that the marginal of A at t_0 is restored at t_2 (that is, $P_A(t_2) = P_A(t_0)$), such that

$$\sum_i \frac{\langle Q_i \rangle}{T_i} = 0, \tag{9}$$

with the sum taken over the entire process. For a reversible process, we have equality.

$$S[P_A(t_1)] - S[P_A(t_0)] = \sum_i \frac{\langle Q_i \rangle}{T_i}. \tag{10}$$

Thus, if two probability distributions can be linked by a reversible process this *uniquely* determines the functional S, up to an additive constant. In the classical case, this is the Gibbs entropy,

$$S_G[P] = \int \rho(x) \log \rho(x)\, dx, \tag{11}$$

where $\rho(x)$ is a density for P with respect to Liouville measure. In the quantum case, S is the von Neumann entropy.

$$S_{vN}[\hat{\rho}] = \mathrm{Tr}[\hat{\rho} \log \hat{\rho}]. \tag{12}$$

Thus, we find that the Gibbs entropy, in the classical context, and von Neumann entropy, in the quantum context, play the role played by

[7]This is valid for both classical and quantum mechanics. The classical version is found in Gibbs (1902, 160–164), and the quantum version, in Tolman (1938, §§128–130).

thermodynamic entropy in the second law, if we consider expectation values of heat exchanges rather than actual values.[8]

The proof that (8) obtains relies only on fairly basic facts about evolution of states, classical or quantum, facts that are indifferent to the temporal order of t_0 and t_1. And, indeed, the result itself is indifferent. Yet equation (8) itself is not symmetric under interchange of t_0 and t_1. The two times enter into the statement of the theorem asymmetrically because it is assumed that the system A is uncorrelated with the thermal systems B_i at time t_0, and it is not assumed that this holds at time t_1. In application, we expect that this is reasonable if t_0 is the earlier time, prior to interactions between A and the thermal systems B_i. That is, it is assumed that we have available to us thermal systems, and that the process by which these come about—relaxation to equilibrium—effectively effaces any correlations there might be between the thermal system and other systems. Thus, to understand the source of temporal asymmetry in thermodynamics, we should look to the process of relaxation to equilibrium.

4 The Ubiquity of Forgetfulness

Much of the discussion surrounding the relation of microscopic dynamics to the behaviour of macroscopic objects has focussed on the issue of time reversal invariance: how can we reconcile temporally irreversible behaviour at the macroscopic level with time-reversal invariance of the fundamental dynamics?

Time reversal invariance isn't really the key issue, however. A symptom of this is that, although there is literature on the proper characterization of time reversal invariance in classical electromagnetism and in quantum mechanics (Albert, 2000; Earman, 2002; Malament, 2004; Callender, 2000), none of it (as the authors would be the first to acknowledge) gets us any closer to understanding the phenomenon of equilibration. Nor does the actual breakdown of time-reversal invariance of the weak nuclear force. Though not time-reversal invariant, the weak force is thought to be invariant under a combination of charge conjugation, parity inversion, and time inversion (CPT). A universal tendency to equilibrium, encompassing both matter and

[8]It should be noted that we are not assuming that the actual values of heat exchanges are even close to the expectation values. The theorem is valid without restriction, and holds even in cases in which the variances of the random variables Q_i are appreciable.

antimatter, breaks CPT symmetry every bit as much as it breaks T symmetry. Similarly, even if claims by Albert and Callender of failure of T-invariance of classical electromagnetism and quantum mechanics were correct, this would get us no closer to understanding equilibration. What is striking about equilibration is that the state of equilibrium approached by a system is independent of the precise details of its initial state. This is a temporal asymmetry of a sort different from a mere violation time-reversal invariance. It is a failure of *invertibility*: distinct initial states lead to the same final state.

This is a phenomenon that is familiar to us in the macroscopic world. The relentless processes of decay lead inexorably to erasure of evidence about the past. Put two objects at different temperatures into thermal contact with each other, allow them to equilibrate, coming to the same temperature, and you will be unable to tell, by experiments performed on the systems, which had been the hotter, and which, the cooler. Everything we take to be a record of the past is subject to the same decay; a book will eventually crumble into dust, and the words that its pages once held will be lost.

Moreover, not just any failure of invertibility suffices to permit this sort of forgetting. Suppose that we have some measure μ on the state space of the system that is conserved under the system's temporal evolution.[9] Then, whether or not the dynamics are invertible, any two disjoint sets will be mapped, by the dynamics, into sets whose overlap has zero measure. In this sense they remain as distinguishable as they were before.

[9]Some terminology. Suppose we have a dynamics on the state space of a system; that is, for each t in some interval of the real line, a mapping T_t of the system's state space into itself, that takes a state at a time t_0 to the state at time $t_0 + t$. Given the dynamics, a measure μ_0 on the state of the system induces a measure μ_t on its state at $t_0 + t$: the measure μ_t assigns to any measurable set of states A the measure, on μ_0, of the set of states at time t_0 that get mapped into A.

$$\mu_t(A) = \mu_0(T_t^{-1}(A)).$$

We will say that the measure μ_0 is *invariant* under the evolution T_t iff $\mu_t = \mu_0$.

Given a set of states A, we can also track the changes (if any) of the measure of its image $T_t(A)$ under evolution. We will say that the measure μ_0 is *conserved* under the evolution T_t iff $\mu_0(T_t(A)) = \mu_0(A)$ for all measurable A.

Any conserved measure is an invariant measure. If T_t is an invertible map, then any invariant measure is also a conserved measure. If T_t is not invertible, then an invariant measure might not be a conserved measure, though it will be non-decreasing under the action of T_t.

Equilibration requires the temporal evolution of the system to erase distinctions between initial states. If there is a measure on the state space of the system that is conserved under temporal evolution, then we have preservation of distinguishability with respect to that measure. In the quantum context, what matters is that the evolution of an isolated system cannot decrease the absolute value of the inner product of two state vectors.

The salient tension between equilibration and the underlying microdynamics, in the classical context, is not the tension between time-reversal symmetry of the microdynamics and temporal asymmetry at the macroscopic level, but, rather, the tension between the existence of a conserved measure at the microphysical level and obliteration of traces of the past at the macrophysical level. In the quantum context, the salient tension is between forgetfulness at the macroscopic level an the existence of a conserved inner product on the Hilbert space of the system. Time-reversal invariance of the microphysics is a red herring.

5 The open systems approach

There are two approaches, at first glance strikingly different, towards the study of equilibration. On one approach, one considers an isolated system, but focusses attention on a limited set of dynamical variables of the system, typically thought of as its macrovariables. The other considers a nonisolated system, in interaction with its environment, and tracks the evolution of the state of the system.

The two approaches are not as different from each other as might seem at first glance. In each case, we are investigating the evolution of a limited set of degrees of freedom of a larger system and disregarding the rest. The larger system is itself treated as isolated, and hence undergoing Hamiltonian evolution. There is no forgetting in the large system; if its full set of degrees of freedom is considered, distinguishability of sets of states is preserved under evolution.

Obviously, this will not hold for subsystems. Though the present state of the whole may determine the past state of a subsystem, it is clear that the present state of a subsystem, disregarding the state of the environment with which it interacts, will not suffice to determine either the subsystem's past or future. The evolution of the subsystem can be a process of forgetting, because details relevant to its past state may have been exported to the environment. Similarly, details

relevant to the past macrostate of an isolated system may become embedded in inaccessible details of its microstate.

Therefore, if we consider a nonisolated system, there is no mystery of reconciling nonconservation of distinguishability of its states with Hamiltonian dynamics of the whole of which it is a part; one would expect this to be ubiquitous. What becomes puzzling instead is the temporal unidirectionality of this phenomenon.

If the description we are working with is a partial description of a total system, it is no surprise that, even though the dynamics together with a *complete* specification of the present state uniquely determine the past, a partial description contains less than complete memory of the past, and the present is compatible with more than one past. By the same token, it would not be surprising if a partial description of the present radically and drastically underdetermined the future evolution. However, in a whole host of cases, we do expect to be able to evolve the present macrostate forward, and make reliable predictions about future macrostates. In some cases, researchers are able to provide reliable, autonomous equations for the forward evolution of the macrostate. What is puzzling is not how one obtains macrodynamics that does not preserve distinguishability from microdynamics that does. The puzzle is why the phenomenon is not as ubiquitous in the forward direction: why does the present macrostate of a system not underdetermine its future as radically as it does its past? To the extent that equilibration is occurring, the evolution merges distinct past states into a single present. Why, then, does this present not bifurcate into distinct possible futures? That is, the real puzzle is: how is prediction ever possible, given that we only have access a partial description of the state of a system?

6 Obtaining autonomous dynamics for subsystems

To get a sense of circumstances under which one can obtain autonomous dynamics for a system in interaction with its environment, rendering prediction of the system's state possible, it is worthwhile to consider a sufficient condition of particular interest.

Consider a system, A, that is in interaction with another system, B. Suppose that, for the duration of the time interval considered, the joint system AB can be treated as isolated, at least as far as the

evolution of A is concerned.[10] Given the state ρ_{AB} of the combined system at some time t_0, we can consider the reduced state ρ_A, which is the restriction of ρ_A to the dynamical variables of A, the reduced state ρ_B, and also the product state $\rho_A \otimes \rho_B$, a state in which there are no correlations between the two systems. We can consider two mathematical operations:

1. Apply the system's dynamics to evolve $\rho_{AB}(t_0)$ forward to $\rho_{AB}(t)$, and then obtain from that $\rho_A(t)$.
2. Apply the system's dynamics to the product state $\rho_A(t_0) \otimes \rho_B(t_0)$ to obtain a state that we will call $\tilde{\rho}_{AB}(t)$, and then obtain from that $\tilde{\rho}_A(t)$.

Now it might so happen that, for some range of values of t, $\rho_A(t)$ is equal to $\tilde{\rho}_A(t)$, or, at least, sufficiently close that the difference is negligible. In such a case, we will say that the correlations between A and B are irrelevant for A's evolution. Suppose that this holds for all t in some interval $[t_0, t_1]$. Suppose also that the influence of A on its environment is such that $\rho_B(t)$ is only negligibly changed during the interval. In such circumstances, we obtain an evolution of ρ_A that can be applied to any initial state of A, that yields $\rho_A(t)$ as a function of $\rho_A(t_0)$ (holding $\rho_B(t_0)$ fixed).

Under what circumstances will this obtain? If B is a large, noisy environment, that can be regarded as a heat bath, it might be the case that, though correlations build up between A and B, the traces of interactions with B become so distributed throughout the system that they play no significant role in future interactions. You may imagine B to be a gas or liquid composed of a large number of molecules undergoing chaotic motion. Molecules that interact with A wander off into the environment and interact with a great many other molecules before they interact with A again, and correlations between A and its environment become so diffused that they are irrelevant to the influence of B on A.

It is this effective lack of correlations that leads to equilibration. Suppose, for example, that the systems A and B are initially at different temperatures. Nothing in the dynamics of the system forbids a steady transfer of energy from the colder to the hotter. But, on average, molecules of the hotter system will have greater energy than

[10] This is a much weaker condition than the condition that the joint system be approximately isolated. It might have considerable interactions with its environment, as long as those interactions don't affect the evolution of A in an appreciable way.

those of the cooler system, so, if interactions between molecules of the system are anything like a random sample of molecules in the two systems, then, with high probability, on any appreciable time scale the net effect will be a transfer of energy from the hotter to the colder system.

The presentation adopted here differs somewhat from the usual presentation in textbooks. In the usual presentations one assumes that the state of the combined system at some time t_0 is a product state, and evolves the joint system forward.[11] Under certain conditions, one will obtain autonomous evolution for ρ_A. This raises the question of what justification one might have for taking systems A and B to be uncorrelated at time t_0.

It might seem that there's a simple answer to this. In the sorts of situations routinely studied in the laboratory, one subjects the systems A and B to independent preparations, and, at some time t_0, places them into thermal contact with each other, examining the subsequent evolution of A. One could invoke some sort of principle to the effect that systems that have not yet interacted be uncorrelated.

This is too quick, however. In the situation considered, there is considerable overlap in the causal pasts of the two systems, and plenty of opportunity for correlations to have built up. What is really being assumed is that the preparation procedures efface those correlations. For example: suppose that our system A consists of a block of ice, and is placed in a bath of warm water, both sourced from the same bottle of distilled water. It is quite possible that some of the molecules that end up in the block of ice are ones that have interacted with molecules that end up in the heat bath. But any entanglement between such molecules will have been very thoroughly and effectively erased by subsequent interactions. The sense that we may have, that a product state is the default state for a pair of systems, in the absence of a process to induce and maintain entanglement, is warranted by the ubiquity of mechanisms of decoherence continually erasing such entanglement. The guiding principle should be: if A and B are two systems interacting with a large noisy environment but not directly with each other, we should expect their state to be effectively a product state in the absence of some process counteracting the effects of decoherence.

The rationale for employing a product state at time t_0, the time

[11] This is one special case of the procedure discussed by Wallace (ming).

at which the systems are brought into thermal contact, is, therefore, much the same as the rationale for employing a product state at later times. We are *not* committed to a full specification of the quantum state of A and its environment being a product state, at t_0 or any other time. We are only committed to the much weaker claim that any entanglement that might exist between A and B at time t_0 be irrelevant to the subsequent development of the system A. The usual textbook treatment, which gives the impression that we are assuming a product state at some time t_0, a condition that could (if there are nontrivial interactions between the systems) hold only for isolated instants, is misleading, as it suggests that the moment t_0 must be singled out as a special moment. This obscures the fact that the rationale for employing a product state (which is not the same as assuming that the state *is* a product state) at time t_0 is of the same sort as the rationale for employing a product state at other times.

7 Temporal asymmetry and the open systems approach

Consider again the example of a system that is in thermal contact with a large, noisy environment, that may be treated as a heat bath. Suppose that, at time t_0, the system is not in thermal equilibrium with its environment, and suppose that we are convinced that, for some time t_0 (not the initial moment at which the system and its environment are brought into thermal contact), employment of a product state at that time yields correct results for times after t_0, and that calculations based on that state yield an approach to equilibrium, with the hotter system cooling and the colder system warming.

Could the procedure be applied in the opposite temporal direction? Could it be the case that taking an uncorrelated state at time t_0 and applying the dynamics in the reverse temporal direction yields a correct, or approximately correct, description of the evolution of the reduced state of A? Evolving a product state backward leads to correlations prior to t_0, correlations that, moreover, are precisely balanced in such a way that the interactions between two systems leads, not merely to effective disentanglement, but to actual disentanglement at t_0. Moreover, typically this procedure would yield approach to equilibrium as one moves away from t_0 in both forward and backwards directions. If A is hotter than B, such an evolution would be, prior

to t_0, one in which energy is transferred from the cooler environment to the hotter body A. Although, on average, the molecules of B have less energy than those of A, the ones that interact with A are systematically those with higher energy than average.

A natural reaction to such a scenario is that there is something uncanny, conspiratorial about the states prior to t_0. The sorts of correlations that lead to transfer of energy from B to A are not the sort that are explicable by appeal to events in their common history, and the erasure of the correlations at t_0 not at all like the sort of erasure of correlations that we expect decoherence to produce.

Some readers will have attempted to train themselves to dismiss such judgments as misguided prejudices; the proper metaphysical attitude, it might be thought, is one of temporal democracy. Any metaphysically respectable explanation, on this mode of thought, ought to work equally well in both temporal directions.

Is this attitude warranted? The intuition that motivates it seems to be that, at some deep level, there is no real difference between past and future temporal directions, no difference that makes a metaphysical difference. We should ask what grounds, if any, we might have for believing this to be true. A conviction of this sort cannot be based on empirical evidence, as the empirical phenomena exhibit a profuse abundance of temporal asymmetries (and this may even be a precondition for the existence of empirical phenomena at all, as it may be a precondition for the existence of cognizing subjects). It might be said that a conviction of this sort is mandated by the time-reversal invariance (or CPT-invariance) of fundamental physical laws. To reach such a conclusion requires an additional premise, one that often goes unstated; the conclusion only follows if we add the stipulation that our explanation can only invoke considerations that follow from fundamental dynamical laws.

Such a stipulation would be unwarranted. To see this, let's step back a moment, and think about the nature of explanation. Consider, for vividness, an example suggested by Maxwell's remark that "The 2^{nd} law of thermodynamics has the same degree of truth as the statement that if you throw a tumblerful of water into the sea, you cannot get the same tumblerful of water out again."[12] Imagine yourself standing by the seaside. You have a tumbler containing a half-litre of fresh water in your hand, and you toss the water into the sea. You then

[12]This appears in a letter to John William Strutt, Baron Rayleigh, dated Dec. 6, 1870. It can be found in Garber et al. (1995, 205) and Harman (1995, 583).

hold the tumbler above the surface of the sea, and wait for a half-litre of fresh water to leap from the sea into the tumbler.

Of course, we expect it to be a long wait; you could stand there until the sun burns out, and you would not expect to see the temporal reverse process of fresh water being poured into to ocean.

Why not?

Consider two possible answers to this question. One is the one that would most readily come to mind to most people. It is, I claim, the correct answer, one that has been too hastily dismissed in the literature on the philosophy of statistical mechanics. The other is the one that is, perhaps, most prevalent in the literature on the philosophy of statistical mechanics.

For the first answer, consider what sort of process the temporal inverse of your tossing the tumbler-full of fresh water into the sea would be. It is one in which the seemingly random thermal motion of molecules becomes a coordinated one, and a very large number of water molecules (and none at all of the molecules of dissolved matter with which they are continually interacting) coalesce in one area of the sea, all possessing an upward velocity that takes them, all in the same general direction, away from the surface of the water, on trajectories that happen to land exactly in the tumbler. We are apt to find the prospect of this concatenation of events unlikely. Asked why, a natural response would be, "What would make them behave like that? The presence of the tumbler can't have that kind of effect on the water molecules."

This sort of thinking has temporal asymmetry deeply embedded in it. This can be seen vividly by considering again the process of hurling the water into the sea. Consider a moment (or a brief time interval) at which the water is in the air, having left the tumbler but not yet hit the sea. In this situation there is coordinated motion of the bits of water; the velocities of the several parts of the blob of water are such that, were all of them reversed, the water would return into the place where the tumbler was when the water left it. If one asked for an explanation of this remarkable coordination between the positions and instantaneous velocities of the bits of water, and their relation to the former position of the tumbler, an adequate explanation of the phenomenon would be: a velocity-reversal of the water would put it on a trajectory that leads back to the former location of the tumbler *because that is where it came from.*

This, I claim, is a perfectly adequate explanation, and ought to be

counted as such on any account of explanation.[13] The same cannot be said of its temporal inverse. If asked why a blob of fresh water that has just emerged from the sea is on a trajectory that will take it into an awaiting tumbler, an answer of "*because that's where it is going*" would not be counted as an adequate explanation.

Underlying these judgments is a concept of explanation on which an adequate explanation can be provided by citing a cause in the recent past that, in conjunction with the laws of physics, accounts for the coordination in question. The concept of cause invoked is one on which the cause-effect relation is temporally asymmetric: a cause must be in the past of its effect. That is, we are invoking a temporally asymmetric notion of what it is to explain something. A causal explanation may invoke dynamical laws that are themselves temporally symmetric; the temporal asymmetry of the explanation lies in the temporal asymmetry of the notion of cause being employed.

Considerations of this sort are not new, of course, and an account of temporal asymmetry in physics invoking considerations of this sort has been defended by Penrose and Percival (1962) and Penrose (2001). The underlying idea is that there be no correlations between systems not attributable to common causes in their past. Acceptance of this basic idea does not necessitate acceptance of Reichenbach's formalization of it, on which the common cause screens off correlations.[14] This sort of reasoning must be applied with due caution and some finesse. It becomes empty if it can only be applied to events with no common past. The fact that there are systems that are effectively independent relies on a process by which events in their common past are rendered irrelevant to their future state.

The other sort of explanation, favoured by neo-Boltzmannians such as Lebowitz (1993, 1999a,b), Goldstein (2001), Price (1996, 2002), and Albert (2000), eschews a temporal asymmetric postulate about probabilities. One chooses a time t_0, and imposes a probability distribution over the state of the system at time t_0 that is the restriction of Liouville measure to its macrostate at that time. Little, if anything,

[13] One could, of course, ask further questions, such as how the water came to be in the tumbler, or why there are tumblers containing fresh water, but these would be requests for explanation of other matters. The original why-question has been answered.

[14] This remark is there because, in Bell-type experiments, it does seem to make sense that the entanglement exhibited by distant particles can be attributed to the circumstances of the generation of the particles pairs at their common source, though there is no Reichenbachian screening-off common cause.

is said about the rationale for this choice of probability distribution, other than that it seems to work, as long as one considers its implications only for events to the future of t_0.[15]

There is a view of probability, often wrongly attributed to Laplace, and often associated with the Principle of Indifference, that probability can be reduced to mere counting of possibilities. The probability of an event A is the number of ways that A can occur, divided by the total number of ways things can be. This sort of thinking has been roundly critiqued in the literature on the philosophy of probability, and has (rightly, in my view) been widely regarded as untenable. Yet it seems to linger in the way some advocates of the neo-Boltzmannian approach talk.

The fundamental problem with a mere-possibility-counting conception of probabilities is that, as Laplace himself stressed, it requires, in order to get off the ground, a judgment about which possibilities are to be regarded as equally probable, and any such judgment will require grounds for favouring that choice over others. In the case of a continuum of possibilities, the Principle of Indifference is thought to enjoin us to adopt a probability distribution on which the parameters we are using to characterize the state of the system are uniformly distributed. This requires a choice of parametrization. Liouville measure, which plays a central role in equilibrium statistical mechanics, is uniform in canonical phase-space coordinates, but not in others.

If a probabilistic postulate of the sort invoked by neo-Boltzmannians, or any other time-symmetric postulate, is to be applied out of equilibrium, then it must be applied at some special time t_0, which is either the beginning of all things (if there is such a time), or else a turning point, with approach toward equilibrium as one moves away from this time, in both temporal directions. One could, for example, apply it to an isolated system that has equilibrated but is at the peak of a

[15]To be sure, Liouville measure, applied to the whole of an isolated system's phase space, is distinguished from other measures in various ways. For one thing, it is a conserved (hence invariant) measure, no matter what the system's Hamiltonian is. Invariance is sometimes mentioned by neo-Boltzmannians (see Lebowitz 1999a, 520; 1999b, S348; Goldstein 2001, 53) as a reason for privileging Liouville measure. This property was invoked by Gibbs as a necessary condition for a measure to represent equilibrium. It cannot be invoked as a rationale for preference out of equilibrium. For a system that is out of equilibrium and undergoing a process of relaxation towards equilibrium, the probability to be attached to a given set of microstates is not unchanging; presumably, one wants to say, of a such system, that it is more likely to exhibit the macroscopic indications of equilibrium at later times in the process than at earlier ones.

fluctuation from the equilibrium mean values of its macrovariables. This is a marked difference from the approach considered here, which involves time-asymmetric considerations that can be applied at *any* time. Moreover, the same work needs to be done on either approach, that of explaining why the sorts of correlations that have built up between systems as a result of events in their common path tend to become irrelevant for prediction of the forward evolution of the system, though remain relevant to retrodiction of the systems' past.

Obviously, a full-scale critique of this sort of approach is beyond the scope of this chapter. I will say only that, insofar as it takes motivation from the idea that probability of an event can be thought of a matter of mere counting, it should be viewed with suspicion. In addition, making reference to a special time in the remote past to explain the ubiquitous and mundane phenomenon of equilibration strikes me as an act of desperation. For much of the period of the development of statistical mechanics, a steady-state cosmology was a live option among serious cosmologists. An advantage of the approach advocated here is that it does not make explanation of events in the laboratory or in our homes, such as the cooling of a cup of coffee, sensitive to large-scale cosmological questions, and can be applied in exactly the same way whether or not there is a cosmologically privileged instant t_0. It also fits better with nonequilibrium statistical mechanics as practiced; one will search in vain textbooks of nonequilibrium statistical mechanics for the sorts of cosmological considerations so frequently found in the philosophical literature on statistical mechanics.

8 Conclusion

If one construes thermodynamics as it was construed at the time that Gibbs was writing, as the science of work and heat, then Gibbs' remark, quoted at the beginning of this chapter, does not seem so naïve. We do have satisfactory statistical mechanical analogues of the first and second laws of thermodynamics. Moreover, these analogues are not, in and of themselves, asymmetric in time. Their formulation presupposes, however, the availability of thermal systems, that is, systems that have relaxed to thermal equilibrium. Understanding circumstances in which this does and does not occur is an active and ongoing research project in physics, and it is one that philosophers would do well to pay attention to. Much of this work involves inves-

tigation of conditions under which one obtains autonomous dynamics either for the state of a subsystem of a larger system, or for a limited set of degrees of freedom of an isolated system. What is needed is an explanation of why the sorts of states that we can reliably produce in the laboratory or can reasonably expect to find in nature are states that afford the autonomous dynamics we seek.

The phenomena to be explained are temporally asymmetric. I have suggested that we may nevertheless obtain an explanation invoking only time-symmetric dynamical laws, because of a temporal asymmetry in the very notion of explanation. This is a temporal asymmetry that most philosophers have shied away from, perhaps taking it as axiomatic that we have not reached a satisfactory explanation until temporal asymmetries have not merely been explained, but have been *explained away*, having been shown to be either merely apparent, or else merely local facts about our limited region of the universe. As a result, proposals such of that of Percival and Penrose have received short shrift in the philosophical literature. At the very least I hope, in this chapter, to have persuaded readers that they deserve serious consideration.

References

Albert, D. (2000). *Time and Chance*. Cambridge: Harvard University Press.

Bartolotta, A., S. M. Carroll, S. Leichenauer, and J. Pollack (2016). Bayesian second law of thermodynamics. *Physical Review E 94*, 022102.

Bricmont, J., D. Dürr, M. Galavotti, G. Ghirardi, F. Petruccione, and N. Zanghì (Eds.) (2001). *Chance in Physics*. Berlin: Springer.

Brown, H. R. and J. Uffink (2001). The origins of time-asymmetry in thermodynamics: The minus first law. *Studies in History and Philosophy of Modern Physics 32*, 525–538.

Callender, C. (2000). Is time 'handed' in a quantum world? *Proceedings of the Aristotelian Society 100*, 246–269.

Clausius, R. (1854). Ueber eine veränderte Form des zweiten Hauptsatzes der mechanischen Wärmetheorie. *Annalen der Physik 93*, 481–506. Reprinted in Clausius (1864, 127–154); English translation in Clausius (1856), and in Clausius (1867).

Clausius, R. (1856). On a modified form of the second fundamental theorem in the mechanical theory of heat. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science 12*, 81–88. English translation of Clausius (1854).

Clausius, R. (1864). *Abhandlungen über die mechanische Wärmetheorie*, Volume 1. Braunschweig: Friedrich Vieweg und Sohn.

Clausius, R. (1867). *The Mechanical Theory of Heat, with its Applications to the Steam Engine, and to the Physical Properties of Bodies*. London: John van Voorst. English translation, with one additional paper, of Clausius (1864).

Darwin, C. G. (1936). Terminology in physics. *Nature 138*, 908–911.

del Rio, L., L. Krämer, and R. Renner (2015). Resource theories of knowledge. arXiv:1511.08818 [quant-ph].

Earman, J. (2002). What time reversal invariance is and why it matters. *International Studies in the Philosophy of Science 16*, 245–264.

Fowler, R. and E. A. Guggenheim (1939). *Statistical Thermodynamics: A Version of Statistical Mechanics for Students of Physics and Chemistry*. Cambridge: Cambridge University Press.

Garber, E., S. G. Brush, and C. W. F. Everitt (Eds.) (1995). *Maxwell on Heat and Statistical Mechanics: On "Avoiding All Personal Enquiries" of Molecules*. Bethlehem, Pa: Lehigh University Press.

Gibbs, J. W. (1902). *Elementary Principles in Statistical Mechanics: Developed with Especial Reference to the Rational Foundation of Thermodynamics*. New York: Charles Scribner's Sons.

Goldstein, S. (2001). Boltzmann's approach to statistical mechanics. See Bricmont et al. (2001), pp. 39–54.

Goura, G., M. P. Müller, V. Narasimhachar, R. W. Spekkens, and N. Y. Halpern (2015). The resource theory of informational nonequilibrium in thermodynamics. *Physics Reports 583*, 1–58.

Harman, P. M. (Ed.) (1995). *The Scientific Letters and Papers of James Clerk Maxwell, Volume II: 1862-1873*. Cambridge: Cambridge University Press.

Lebowitz, J. L. (1993). Boltzmann's entropy and time's arrow. *Physics Today 46*(September), 33–38.

Lebowitz, J. L. (1999a). Microscopic origins of irreversible macroscopic behavior. *Physica A 263*, 516–527.

Lebowitz, J. L. (1999b). Statistical mechanics: A selective review of two central issues. *Reviews of Modern Physics 71*, S346–S357.

Malament, D. B. (2004). On the time reversal invariance of classical electromagnetic theory. *Studies in History and Philosophy of Modern Physics 35*, 295–315.

Maroney, O. (2007). The physical basis of the Gibbs-von Neumann entropy. `arXiv:quant-ph/0701127v2`.

Myrvold, W. C. (2011). Statistical mechanics and thermodynamics: A Maxwellian view. *Studies in History and Philosophy of Modern Physics 42*, 237–243.

Penrose, O. (2001). The direction of time. See Bricmont et al. (2001), pp. 61–82.

Penrose, O. and I. C. Percival (1962). The direction of time. *Proceedings of the Physical Society 79*, 605–616.

Price, H. (1996). *Time's Arrow and Archimedes' Point*. Oxford: Oxford University Press.

Price, H. (2002). Boltzmann's time bomb. *The British Journal for the Philosophy of Science 53*, 83–119.

Saha, M. N. and B. N. Srivastava (1931). *A Text Book of Heat*. Allahabad: Indian Press.

Saha, M. N. and B. N. Srivastava (1935). *A Treatise of Heat*. Allahabad: Indian Press.

Sommerfeld, A. (1956). *Themodynamics and Statistical Mechanics: Lectures on Theoretical Physics, Vol. V*. New York: Academic Press.

Thomson, W. (1853). On the dynamical theory of heat, with numerical results deduced from Mr Joule's equivalent of a thermal unit, and M. Regnault's observations on steam. *Transactions of the Royal Society of Edinburgh 20*, 261–288. Reprinted in Thomson (1882, 174-210).

Thomson, W. (1857). On the dynamical theory of heat. Part VI: Thermo-electric currents. *Transactions of the Royal Society of Edinburgh 21*, 123–171. Reprinted in Thomson (1882, 232–291).

Thomson, W. (1882). *Mathematical and Physical Papers*, Volume I. Cambridge: Cambridge University Press.

Tolman, R. C. (1938). *The Principles of Statistical Mechanics*. Oxford: Clarendon Press.

Uhlenbeck, G. E. and G. W. Ford (1963)). *Lectures in Statisical Mechanics*. Providence, R.I.: American Mathematical Society.

Wallace, D. (2016). Thermodynamics as control theory. *Entropy 16*, 699–725.

Wallace, D. (forthcoming). The logic of the past hypothesis. In B. Loewer, E. Winsberg, and B. Weslake (Eds.), *Time's Arrows and the Probability Structure of the World*. Harvard: Harvard University Press. Available at `http://philsci-archive.pitt.edu/8894/`.