Thermodynamics with and without irreversibility

David Wallace[*]

July 1, 2023

Abstract

Working inside the control-theoretic framework for understanding thermodynamics, I develop a systematic way to characterize thermodynamic theories via their compatibility with various notions of coarse-graining, which can be thought of as parametrizing an agent's degree of control of a system's degrees of freedom, and explore the features of those theories. Phenomenological thermodynamics is reconstructed via the 'equilibration' coarse-graining where a system is coarse-grained to a canonical distribution; finer-grained forms of thermodynamics differ from phenomenological thermodynamics only in that some states of a system possess a free energy that can be extracted by reversibly transforming the system (as close as possible) to a canonical distribution. Exceeding the limits of phenomenological thermodynamics thus requires both finer-grained control of a system and finer-grained information about its state. I consider the status of the Second Law in this framework, and distinguish two versions: the principle that entropy does not decrease, and the Kelvin/Clausius statements about the impossibility of transforming heat to work, or moving heat from a cold body to a hotter body, in a cyclic process. The former should be understood as relative to a coarse-graining, and can be violated given finer control than that coarse-graining permits; the latter is absolute, and binds any thermodynamic theory compatible with the laws of physics, even the entirely reversible limit where no coarse-graining is appealed to at all. I illustrate these points via a discussion of Maxwell's demon.

1 Introduction

Dynamical theories are concerned with what physical systems spontaneously do: they provide dynamical equations and dynamical laws relating a state of a system at one time to a state at another, all independent of outside intervention. Thermodynamics, the misleading name notwithstanding, is *not* a dynamical theory in this sense: if it has a dynamical law, it is simply that systems left to themselves go to equilibrium and stay there. Most of its content involves

[*] Department of History and Philosophy of Science / Department of Philosophy, University of Pittsburgh; email `david.wallace@pitt.edu`

transitions between equilibrium states, and since equilibrium states by definition do not carry out any transitions at all if left to themselves, those transitions must be understood as externally induced. Thus understood, thermodynamics is a *control theory*, concerned not with what transitions spontaneously happen but with what externally-induced transitions are and are not possible, or put another way, what *control operations* can or cannot be performed.[1]

Viewed as a control theory, thermodynamics is predominantly concerned with a somewhat more specific problem. Given some additive, conserved quantity — usually energy, but it could be angular momentum, or charge, or any other such quantity — how much of that quantity can be *extracted* from a given system by a given class of control operations?

An immediate answer might be: none of it. These are *conserved* quantities, after all: the quantity might flow from one part of the system to another, but it cannot increase or decrease. But of course among our control operations might be couplings of the collection to other systems — batteries, flywheels, lifted weights, or other storage devices — so that the collection's share of the quantity decreases even while the amount of it in the larger world remains fixed. Indeed, that is just what 'extract' means for an overall-conserved quantity: the extractor has acquired some of it, so less of it remains.

That suggests a second answer: all of it. If the only constraint on extraction is compliance with the conservation laws, then the most we can extract is the difference between the *actual* amount of the quantity, and the *minimum* amount — for energy, for instance, the difference between the actual energy of the system and the energy of its ground state.

Phenomenological thermodynamics tells us that at least in some circumstances, the constraints are more stringent. If our system consists of a box of gas at uniform temperature, and in our control operations we are obliged to return the gas to the same volume at which it begins, then *none* of the gas's energy can be extracted. If the system instead consists of two such gases at different temperatures, we may run a heat engine between the two gases and extract some of the energy that thus flows — but not all of it, and only until the temperatures equalize. Indeed, the lesson of thermodynamics is that all such constraints can be systematized into two principles: that all control processes conserve energy (the First Law), and that no control process decreases thermodynamic entropy (the Second Law). And the irreversibility of equilibrium thermodynamics is often said to lie in the fact that the Second Law's inequality need not be saturated: that there although there are no control processes that decrease entropy, there are some which increase it, and so their are some operations which, once done, cannot be undone.

The First Law sits on a solid microphysical foundation: thermodynamic energy is identified with dynamical energy (that is: the conserved quantity associated with time translation symmetry) and the principle that it is conserved is inviolate, no matter how sophisticated our control processes might be: any

[1] The control-theory way of understanding thermodynamics is defended by Wallace (2014) and Myrvold (2020); it has been widely discussed in the recent physics literature under the title of 'resource theory' (for a review, see, e. g. , Gour et al (2015).)

increase in energy in some external storage device we operate must be compensated, exactly, by a decrease of energy elsewhere, either in another external device or in the system on which we operate. (And the same applies *mutatis mutandis* to the extraction of other conserved quantities, albeit this is seldom a focus of equilibrium thermodynamics.)

The Second Law, on the other hand, is often attributed to the macroscopic nature of equilibrium thermodynamics, and/or to the crude limitations of the control operations that clumsy creatures like us have access to. The point was made vividly by Maxwell (1867; 1871, ch.12) at the dawn of statistical mechanics: his infamous 'demon' is a microscopic being, capable of tracking the movements of individual molecules in a gas and steering them, ever so delicately, so as to reduce the gas's entropy and allow work to be extracted. And ever since Maxwell's proposal the possibility has been raised and reraised that a subtler control theory, appropriate to a subtler agent, might transcend its limitations. This conception of the Second Law is closely tied to the idea that thermodynamics is in some way a higher-level, emergent, approximate theory, useful for large-scale systems no doubt but only statistically true even then, breaking down entirely on sufficiently small scales.

This way of thinking is in tension with another tradition, just as old, of seeing thermodynamics as exact and universal. Einstein (1949, p.33) famously described thermodynamics as "the only physical theory of universal content concerning which I am convinced that, within the framework of applicability of its basic concepts, it will never be overthrown"; more recently, the physics consensus[2] about Maxwell's demon is that reasons of principle and not just practice make it impossible, and the defenders of that consensus consider microscopic systems in states far from equilibrium; meanwhile, recent advances in statistical mechanics[3] have seen physicists apply thermodynamics to extremely small systems, not just to macroscopically large ones, and to obtain theoretical results which are then confirmed empirically.

This paper has several goals. Firstly, I aim to disentangle these two aspects of thermodynamics: the emergent, high-level, large-system, approximate features, from the exact, microscopically derivable features. I do so by introducing the idea of a coarse-graining (a formal transformation of a system's state which washes out certain fine details of its microstructure) and of a control theory defined by that coarse-graining, not in the sense that any control operation actually performs the coarse-graining but that control operations are indifferent to details that the coarse-graining washes out. Each coarse-graining then specifies a class of control theories, and the finer the coarse-graining, the more powerful the control theory. The coarsest grain, which defines *equilibrium thermodynamics*, just replaces every system with its equilibrium state; the finest, which defines *reversible thermodynamics*, does not coarse-grain at all. I offer this framework as a general method both to the derivation of phenomenological thermodynamics from microphysics, and to generalize to finer-grained control theories where

[2] See section 11 for details and references.

[3] See, e. g. , (Jarzynski 1997; Crooks 1998; Collin *et al* 2005; Bustamante, Liphardt, and Ritort 2005); for a conceptual discussion, see (Wallace 2020).

agents have control options not present in phenomenological thermodynamics.

Secondly, I explore the microphysical origins of irreversibility in thermodynamic control theory (to be distinguished from the irreversibility of non-equilibrium statistical mechanics, which I do not consider here). This (I will argue) is closely tied to the limitations of our control of a system, and hence to our choice of coarse-graining.

Thirdly, I attempt to provide some clarity on the vexed question of how information connects to thermodynamic entropy. In some corners of physics, it is a commonplace that information and entropy are one and the same, but the idea has been met with skepticism in other corners and even more so in philosophy. In my account, gaining information about a system decreases its entropy, and allows us more control over it, only when that information tells us about the coarse-grained features of a system: coarse-grained, that is, with respect to the coarse-graining that characterizes an agent's particular control theory.

Finally, I argue that there are hard limits on what can be done to thermodynamic systems even by an agent with total, microscopic, control of the system, limitations that apply not because of our limited ability to influence systems at the microscopic level but because of microscopic physics itself. Specifically, the most important phenomenological versions of the Second Law continue to apply, even in the case of a fully reversible thermodynamics.

The structure of the paper is as follows. In sections 2 and 3 I review first the structure of phenomenological thermodynamics (with an emphasis on the central role of the equation of state) and then the 'canonical recipe' by which the equation of state is derived from a system's microphysics (which at this stage I simply state without defense). In section 4 I provide a general framework for discussing thermodynamic control theories, and in section 5 I introduce the idea of coarse-graining in general and the equilibration coarse-graining that justifies the canonical recipe in particular. In sections 6–8 I further develop the microphysical basis for phenomenological thermodynamics, in the process introducing the important concepts of 'free energy of equilibration' (section 7) and 'heat baths' (section 8) which apply in a much more general context. In sections 9–10 I consider some more powerful thermodynamic control theories, firstly one inspired by Boltzmannian statistical mechanics and secondly a fully reversible thermodynamics in which the controlling agent can apply arbitrary unitary (or, in a classical context, Hamiltonian) transformations. In section 11 I consider the circumstances in which collecting information about a system can be thermodynamically useful; in sections 12–13 I return to the Second Law in the context of the very general notions of thermodynamics I am considering, and then distinguish ways in which Maxwell demons are or are not possible. Section 14 is the conclusion.

This paper builds on previous work by myself and others. Wallace (2014) defends the idea of thermodynamics as control theory, and develops the microscopic details in the specific context of equilibrium thermodynamics. Myrvold (2011, 2020) offers a similar defense, but introduces the important idea that there are different forms of thermodynamics corresponding to different levels of control an agent might have (or rather, reintroduces it, since he provides strong historical evidence that the idea was well understood by many of the physicists who developed thermodynamics and statistical mechanics); he too works in the context of equilibrium thermodynamics. Maroney (2007) develops a fully reversible form of thermodynamics, separated entirely from considerations of equilibrium, and my development here of 'reversible thermodynamics' is indebted to his version. The physics I discuss is very standard and I do not attempt to provide original references; for textbook discussions see, e. g. , (Tolman 1938; Landau and Lifshitz 1980; Kittel and Kroemer 1980; Blundell and Blundell 2010; Throne and Blandford 2017).

Four technical notes. Firstly, I work in natural units in which Boltzmann's constant k_B is set to 1. Secondly, I develop my account for the most part neutrally as to whether the underlying dynamical theory is classical or quantum, though for certain technical proofs I specialize to quantum mechanics. (In most cases there is an analogous classical proof, but in any case classical mechanics is valid only insofar as it successfully approximates quantum mechanics.) Thirdly, when discussing quantum mechanics I assume (i) that the normal quantum formalism is complete; (ii) that closed-system dynamics are always unitary; (iii) that processes of observation and measurement can in principle be modelled mechanically within the quantum formalism. In my view (Wallace 2012) this more or less amounts to assuming the Everett (many-worlds) interpretation, but I make no use of Everett-specific ideas and most of what I say here can probably be taken over to other approaches to quantum mechanics with little change. Finally, by 'phenomenological thermodynamics' I mean the macroscopically developed theory of thermodynamics, understood in isolation from its microphysical roots; I will later construct a microscopic control theory, (partial) equilibrium thermodynamics, which provides a microphysical justification of phenomenological thermodynamics.

2 Phenomenological thermodynamics

Phenomenological thermodynamics is concerned with those transitions which can be induced by an external agent (put another way, those control operations) that move the a physical system or collection of systems from one equilibrium state to another), where 'equilibrium' is understood operationally as the state to which an undisturbed system settles down to after some period of time. Different possible states of a thermodynamic system are labeled by the system's total energy U, by some (possibly empty) set of parameters describing externally-controllable features of the system, and by some (possibly empty) set of conserved quantities other than energy. For instance:

- The thermodynamic state of a piece of rock of fixed size and shape is characterized by its energy alone.
- The thermodynamic state of a fixed quantity of gas in a box is characterized by the energy of the gas and the volume of the box. If the amount of

gas in the box is also adjustable, the thermodynamic state is in addition characterized by the total number of particles (or, equivalently, moles) of the gas.

- The thermodynamic state of black-body radiation in a box is characterized by its energy and the box volume (but not by the number of photons, since that is not a conserved quantity).
- The thermodynamic state of a ferromagnet in an external magnetic field is characterized by its energy and the field strength.
- The thermodynamic state of a black hole, described in its rest frame, is characterized by its mass (=energy), charge, and angular momentum.

For simplicity I will usually write as if there was one parameter, V, and one conserved quantity, N; everything that follows generalizes straightforwardly to the case of none or several parameters and/or conserved quantities.

The quantitative features of a thermodynamic system are determined — at least as far as phenomenological thermodynamics is concerned — entirely by one function, the *state function* (or *equation of state*), which is a real-valued function $S(U, V, N)$ of U, V and N whose value for a particular state is called the *thermodynamic entropy* of the state. It is normally assumed that (for fixed V, N) S is a strictly increasing function of U, so that this may be inverted to write $U = U(S, V, N)$ ('state function' is often used to describe this function also). Given a system that can be decomposed into non- or weakly-interacting subsystems, each has its own equation of state and the entropy of the total system is the sum of the entropies of its subsystem; energy and the other conserved quantities are likewise assumed to be additive. (The approximate additivity of the energy is more or less constitutive of two systems being weakly interacting.)

The operational significance of this machinery is then given by the two *laws of phenomenological thermodynamics* and by a secondary *accessibility principle*:

The First Law: For any transition $(U, V, N) \rightarrow (U', V', N')$ induced on an otherwise isolated system by an agent, the net energy required by the agent to perform the transition equals $(U' - U)$, the change in energy of the system. That is: the sum of system energy and controller energy is conserved. (As is standard in phenomenological thermodynamics, this assumes that we have an antecedent understanding of the energy of a mechanical system.)

The Second Law: Any transition $(U, V, N) \rightarrow (U', V', N')$ induced on an otherwise isolated system satisfies $S(U', V', N') \geq S(U, V, N)$.

The Accessibility Principle: If two states have entropies S, S' with $S' > S$, there is an transition from the first to the second.

Other thermodynamic quantities are definable from the equation of state. We can define the temperature $T(S, V, N)$, the generalized pressure $P(S, V, N)$, and the generalized potential $\mu(S,V,N)$, via the differential expression

$$\mathrm{d}U(S,V,N) = T(S,V,N)\mathrm{d}S - P(S,V,N)\mathrm{d}V + \mu(S,V,N)\mathrm{d}N \qquad (1)$$

so that (for instance),

$$T = \left(\frac{\partial U}{\partial S}\right)_{V,N}. \qquad (2)$$

We define the *work done* to a thermodynamic system in a control process as the increase in the system's energy caused by the process. If the work done is negative, we speak of *work extracted.* If we think of a control process as ultimately physically realized in a larger system in which total energy is conserved, the work done by a control operation is the energy that the controller must provide in order to perform it, and the work extracted, conversely, is the energy gain to the controller in carrying out the control operation (which is often operationalized via some energy-storage device such as a raised weight, flywheel, or battery).

If, as is often the case, a thermodynamic system can be broken into some number of (for simplicity, let's say two) approximately-isolated subsystems, it is generally possible to consider the total work W done as the sum of the work W_1 done on system 1 and W_2 done on system 2. If the energy changes in the two systems are ΔU_1, ΔU_2, then we define the *heat flow* Q_i into the ith system as

$$Q_i = \Delta U_i - W_i. \qquad (3)$$

We have $Q_1 = -Q_2$, so that heat can be thought of as flowing from one system to the other; definitionally, the total heat flow into our system is zero. (It is possible to construct versions of thermodynamics involving heat baths in which net heat flow can be nonzero — and I discuss these later in section 8 — but for our purposes we can think of these as effective theories, reducible to phenomenological thermodynamics by including the heat bath as a subsystem of the thermodynamic system.)

It is worth noting that (except in the special case of an overall quasistatic, reversible process, which I will not need to consider here) the machinery of thermodynamics does not itself have the resources to cleanly define work and heat for subsystems, since it does not seek to characterize control processes directly, but only to constrain them via the Second Law.

As an illustration of how this framework has phenomenological consequences, let us define a transition as *cyclic* if it leaves parameters and conserved quantities (other than energy) unchanged. Then a cyclic transition operating on a single system which extracts a small amount of net work from that system satisfies

$$T\delta S = \delta U \leq 0 \qquad (4)$$

and so if T is positive no such transition is possible. A small cyclic transition operating instead on a pair of systems (1 and 2) in which zero net work is done will have as its only effect the transfer of a small amount of heat between the systems; it satisfies

$$0 = \delta U = \delta U_1 + \delta U_2 \qquad (5)$$

so that $T_1\delta S_1 + T_2\delta S_2 = 0$. The Second Law requires that $\delta S_1 + \delta S_2 \geq 0$, so that if $\delta S_1 < 0$, $T_1 \geq T_2$. That is: no cyclic process can induce heat flow from a cold body to a hotter body without some expenditure of external energy or other change in the systems. These two results can be recognized as, respectively, the Kelvin and Clausius statements of the Second Law (I discuss them further in section 12.)

If we instead consider allowing heat $đQ$ to flow from a higher-T to a lower-T body, while siphoning some quantity $-đW$ of it out altogether as extracted work, we have (assuming $T_2 > 0$)

$$\delta S_1 = -đQ/T_1 \tag{6}$$

and

$$\delta S_2 = +(đQ + đW)/T_2 \tag{7}$$

so that

$$0 \leq T_2(\delta S_1 + \delta S_2) = đQ\left(1 - \frac{T_1}{T_2}\right) + đW \tag{8}$$

or

$$-đW \leq đQ\left(1 - \frac{T_1}{T_2}\right) \tag{9}$$

which is the well-known Carnot limit on the efficiency of a heat engine. The accessibility principle tells us that this limit can be approached to arbitrary accuracy; we can define *reversible* processes (aka *Carnot* processes) as idealized limits of real processes which generate less and less entropy increase, and which in that limit obtain the Carnot efficiency. Carnot efficiency.

3 Deriving the equation of state

While it is contested to what extent equilibrium statistical mechanics offers a *conceptual* underpinning of phenomenological thermodynamics, there is no room to doubt that it provides methods to *quantitatively calculate* the thermodynamic description of a system from its microphysics. The recipe for doing so — let's call it the *canonical recipe*[4] (it would be premature to call it a 'derivation') is well known, and can be used both in quantum and classical mechanics: it will be helpful to adopt as far as possible a notation neutral between the two. For these purposes, let's define a *dynamical state space* as either a phase space, i. e. a symplectic manifold (classical) or a separable Hilbert space (quantum), and a *quantity* for a dynamical state space as either a smooth real function on the space (classical) or a self-adjoint operator on the space (quantum). Any such quantity induces a time-indexed flow on the dynamical state space, either under Hamilton's equations (classical) or the Schrödinger equation (quantum), in either case using the quantity as a Hamiltonian. And a *distribution* on a

[4] There is also an *microcanonical* recipe, which I will not discuss here; in practice the canonical recipe is almost always the one used.

dynamical state space is either a probability distribution on the space (classical) or a density operator on the space (quantum). For any quantity X, the symbol

$$\int X$$

represents either the integral of X over phase space (classical) or the trace of X (quantum), and by definition the expectation value of a quantity X with respect to a distribution ρ is

$$\langle X \rangle_\rho = \int X\rho. \tag{10}$$

The canonical recipe can now be summarized as follows:

1. An isolated thermodynamic system is represented by a dynamical state space and a parameterized family $H(V)$ of Hamiltonians for the space: each Hamiltonian is a quantity, and $V \to H(V)$ is a smooth map from the space of parameters into the space of quantities. Any other conserved quantity is represented by another such parameterized family $N(V)$ of quantities, which are conserved in the usual sense: for each V, $N(V)$ is invariant under the dynamical flow induced by $H(V)$. (Which is to say that the Poisson bracket or commutator, as appropriate, vanishes for each V between $H(V)$ and $N(V)$.)

   For future purposes, I take this as the *definition* within statistical mechanics of a thermodynamic system.

2. For given values U, V, N of the energy, parameter, and conserved quantity, the *canonical distribution* $\rho_c(U, V, N)$ is the distribution

   $$\rho_c(U, V, N) = \frac{1}{Z(T,\mu)} \exp\left(-\frac{H - \mu N}{T}\right) \tag{11}$$

   where the *partition function*

   $$Z(T,\mu) = \int \exp\left(-\frac{H - \mu N}{T}\right) \tag{12}$$

   is a normalizing factor, and T and μ are functions of U, V, N defined implicitly by the requirements that

   $$U = \langle H(V) \rangle_\rho \tag{13}$$

   and

   $$N = \langle N(V) \rangle_\rho. \tag{14}$$

   (The generalization to several or no conserved quantities is trivial; it is common to call this expression the *grand* canonical distribution where conserved quantities other than energy are present, but I eschew this as adding unneeded complexity.)

3. The equation of state for a thermodynamic system is given by
$$S(U,V,N) = S_G[\rho_c(U,V,N)] \tag{15}$$
where $S_G[\rho]$ is the Gibbs entropy,
$$S_G[\rho] = -\int \rho \ln \rho. \tag{16}$$
(The notation continues to be systematically ambiguous between classical and quantum: in the quantum case this is the von Neumann entropy.)

4. Explicit calculation from (15) gives
$$S = U/T + \ln Z \tag{17}$$
and hence
$$\left(\frac{\partial S}{\partial T}\right)_{V,N} = \frac{1}{T}\left(\frac{\partial U}{\partial T}\right)_{V,N}. \tag{18}$$
By comparison with (1), T is indeed the thermodynamic temperature, as my notation suggests; a similar calculation shows that μ is likewise the generalized potential.

Why the canonical recipe works is contested. *That* it works is amply demonstrated by the last century of work in equilibrium statistical mechanics.

4 Characterizing control theory

To explain why the canonical recipe works, we need to consider the control theory of a thermodynamic system in microphysical terms. That theory ought to consist of a set of allowable transformations, the control operations, on microphysically characterized states of the thermodynamic system; the hope would be to find some independently-justified way of specifying the allowable control operations such that the resultant control theory explains the recipe's success.

What *are* the 'microphysically characterized states' of a thermodynamic system? A natural guess (working for the moment specifically in classical mechanics) would be that they are the points of phase space, each of which describes exactly the properties of the system (say, the positions and momenta of all its composite particles). A slightly more careful guess would be that the states are ordered pairs (x, V), where x is a phase space point and V is a parameter value, since we ought to include the parameter values as part of the specification of the state.[5]

[5] In the case of volume in particular, it may seem strange to think of the volume as a parameter in the Hamiltonian, rather than just part of the specification of the phase space. But ultimately, the particles are confined to the box not by *a priori* kinematics but by their dynamical interaction with the walls, and we can model this conveniently by treating the box as a potential, zero within the box and then increasing extremely rapidly to a value much larger than any relevant particle kinetic energy outside the box. Any particle position is then kinematically possible, but positions outside the box will be dynamically prohibited. Thanks to Katie Robertson for pressing me on this point.

This is not in fact the normal practice in (at least contemporary) statistical mechanics. Here the standard move is instead to treat the state as a probability distribution, giving the probability density for the system to be represented by any given phase-space point (again, in our framework the more careful move would be to take a state as an ordered pair (ρ, V) where ρ is a distribution and V a parameter value). Often, especially in textbooks and foundational presentations, this distribution is given an epistemic gloss: it represents the controlling agent's partial knowledge of the system's true state.

This move to a probabilistic characterization of the state has been heavily criticized by philosophers. David Albert, for instance,[6] writes:

> Can anyone seriously think that our merely being ignorant of the exact microconditions of thermodynamics plays some part in bringing it about, in making it the case, that (say) milk [mixes into[7]] coffee? (Albert 2000, p.64)

However:

1. The epistemic interpretation of the probabilities is for the most part[8] optional. There are numerous ways to interpret them as objective — as long-run relative frequencies (Tolman 1938); as the result of a Hume-Lewis best-systems analyis (Lewis 1980; Loewer 2002; Ismael 2009); as the consequence of primordial chance events (Demerast 2016); as the classical limit of quantum indeterminacy (Albert 2000; Wallace 2016).

2. In any case, Albert's criticism applies to statistical-mechanical explanations of what systems do *by themselves*. In the present paper, I am concerned with the control theory of mechanical systems, not their spontaneous behavior — and there is nothing inherently paradoxical about the fact that my having more information about a system gives me better ability to control it. (Of course, whether a piece of information about a system's state is useful to me depends on whether I have access to control operations that can exploit that particular piece of information; this is a theme which will recur several times in my analysis.)

3. Use of distributions in at least the formal sense is unavoidable once we start doing quantum rather than classical mechanics. The classical limit of a quantum state (even a pure state) is a classical probability distribution, not a classical microstate; and even if a system's state is pure, its subsystems' states are in general mixed; even if a composite system begins in a product state, dynamical evolution will generically entangle those states. (For more on this point, in the context of non-equilibrium statistical mechanics, see (Wallace 2016).)

[6] For other examples, see, e. g. , (Goldstein 2001; Callender 2001).

[7] Albert writes 'dissolves in' here, but milk is not water-soluble.

[8] In statistical mechanics (as distinct from thermodynamics) it is entirely optional. In thermodynamics the existence of an agent who intervenes on a system on the basis of their information about the system means that some notion of the state as partly representing an agent's information about the system cannot entirely be dispensed with.

For these reasons (and also given the urgent goal of providing a justification for actually-used methods in physics) I continue to follow physics orthodoxy: by definition a *state* of a thermodynamic system is an ordered pair (ρ, V) of a distribution and a parameter value (where we have returned to our systematic ambiguity between classical and quantum mechanics: in the quantum case states are ordered pairs of density operators and parameter values).

We can now give a formal definition of a control theory for a thermodynamic system (or a *thermodynamic control theory*):

- A *control operation* for the system is a triple $C = \langle \Pi_C, V, V^C \rangle$ where Π_C, the *control map* of C, is a map from distributions to distributions and V, V^C are parameter values. It defines a partial map on states: C maps the state (ρ, V) to $C(\Pi_C, V, V^C) = (\Pi_C \rho, V^C)$. (I adopt the systematic notation that Π_C and V^C are respectively the distribution map and final parameter value associated with C.)

  Given two control operations $C_1 = \langle \Pi_{C_1}, V_1, V^{C_1} \rangle$ and $C_2 = \langle \Pi_{C_2}, V_2, V^{C_2} \rangle$ with $V^{C_1} = V_2$, their composition is

  $$C = C_2 \cdot C_1 = \langle \Pi_{C_2} \Pi_{C_1}, V_1, V^{C_2} \rangle \tag{19}$$

  and the compositions of three or more operations can be defined inductively in the obvious way: $C_3 \cdot C_2 \cdot C_1 = C_3 \cdot (C_2 \cdot C_1)$, etc. .

- A *control theory* for the system consists of

  (i) A set of *allowable control operations* for that theory, containing at least the identity operations $\langle \mathrm{id}, V, V \rangle$ for all V.

  (ii) A set of *allowable initial states* for that theory.

  (Note that I do not assume the allowable control operations are closed under composition; indeed, this will generally prove impossible.)

In phenomenological terms, the control operations for thermodynamics include (but are not limited to) operations like 'vary the parameters over some period of time' and 'place two isolated subsystems in dynamical contact and allow them to come partially to equilibrium, then decouple them'.[9] Microphysically, we can characterize the control map of the former operation by dynamical flow under the time-dependent Hamiltonian $H(V(t))$, where $V(t)$ is a path through parameter space, and of the latter by dynamical flow under a time-dependent Hamiltonian that begins and ends at the total Hamiltonian of the isolated subsystems but passes through a region in quantity space representing Hamiltonians which dynamically couple the subsystems. More generally, any time-dependent path $h(t)$ through the space of quantities satisfying $h(T) = H(V)$, $h(t') = H(V')$ for some $t' > t$ determines a control operation in which the state flows under the

[9] In some treatments (e. g. , (Myrvold 2020)) these are taken to exhaust the range of available operations, but normally operations like vigorously stirring a fluid are included in the range of operations available even in phenomenological thermodynamics.

time-dependent Hamitonian $h(t)$, applied between times T and T'. Standard results of classical and quantum dynamics tell us that a control operation like this — where the Hamiltonian is varied along some smooth path from $H(V)$ to $H(V^C)$ and the system evolves under the resultant time-dependent dynamics — determines a dynamical map which is the push-forward of a symplectomorphism (i. e. , a canonical transformations) in the classical case, and the adjoint action of a unitary operator in the quantum case; I call maps of this kind *canonical* (continuing with our systematically ambiguous notation). Conversely, any canonical map that is topologically connected to the identity can be generated (non-uniquely) by some choice of time-dependent Hamiltonian. We call a control operation *Hamiltonian* if its control map is canonical, and a control theory Hamiltonian if all its control operations are Hamiltonian.

Are Hamiltonian control theories the most general that we can consider? To some extent it is a matter of definition. There are certainly things we could do to the systems that are not Hamiltonian control operations: for instance, we can bring in a new system and let it dynamically interact with the old ones so that not only does it induce a change in the Hamiltonian of the latter, but it allows new and old system's microstates to become correlated. Or we can measure some properties of the systems and choose our control operation accordingly. (These are actually quite closely related, as we will see in section 11.) But in each case we can absorb the operation in our overall framework by zooming out. In the first case, we can simply choose to include the 'new' system as a subsystem of a larger thermodynamic system, and consider the control theory of that system. In the second, we can automate the process of measuring and applying the measurement-dependent operation, and include the machine that so automates it as one more system in the collection; the externally-applied control operation just consists of turning the machine on, which can be modelled just fine in the Hamiltonian class of control operations. So for the most part we will avail ourselves of these tricks to restrict attention to the Hamiltonian operations; I develop the details in sections 8 and 11. (It is, however, sometimes useful to consider certain formal extensions of the class of Hamiltonian control theories, such as idealized limits of Hamiltonian control operations which are not Hamiltonian.)

A key feature of a Hamiltonian control map Π_C is that it conserves the Gibbs entropy: $S_G[\Pi_C\rho] = S_G[\rho]$. The converse is not true; we will call a control operation *volume-preserving* if it conserves Gibbs entropy, whether or not it is Hamiltonian (since the underlying mathematical reason that canonical transformations leave S_G invariant is that they conserve phase-space volume or its analog, Hilbert space dimension), and a control theory as volume-preserving iff its control operations are all volume-preserving.

5 Justifying the canonical recipe

So much for the *general* framework for thermodynamic control theory. What *specific* assumptions about the theory are needed to recover phenomenological thermodynamics? We can get insight by asking why the canonical distribution plays the central role in the recipe for calculating the equation of state of a thermodynamic system. It has two crucial properties, related but distinct, that explain this. Firstly, $\rho_c(U,V,N)$ is the unique maximum-Gibbs entropy distribution for given U,V,N: any other distribution defined for the same parameter value V and with the same expectation values of energy and other conserved quantities has a strictly lower Gibbs entropy. (This is a standard result; for completeness I prove it in the Appendix in the quantum case, under mild simplifying assumptions.) Secondly, near-universal practice in statistical mechanics treats $\rho_c(U,V,N)$ as the unique equilibrium distribution which an isolated system with those expectation values and parameters *spontaneously* approaches — approaches in the sense that it is this distribution that is used to calculate all observable features of a system at equilibrium, including not only thermodynamic features but statistical properties like multi-time correlation functions, fluctuations, and the expected fraction of particles with a given energy.

This suggests a rather straightforward (indeed, as we will see, naive) control theory for phenomenological thermodynamics. Suppose that the controlling agent can manipulate the bulk features of a system but cannot prevent that system from spontaneously approaching equilibrium. And suppose the approach to equilibrium is *literally* modeled by the evolution of a state with given U,V,N to the *canonical state* $(\rho_c(U,V,N),V)$. Then (i) the allowable initial states for the control theory should comprise the canonical states, and (ii) the control operations should consist of some Hamiltonian operation applied to an initial state, followed by equilibration, so that the state after a control operation is also a canonical state.

At this point, we can derive phenomenological thermodynamics fairly directly. Firstly, if the actual equilibrium state at (U,V,N) is a canonical distribution, then it is no wonder that the expectation values of that canonical distribution correspond to their phenomenological values (at least, assuming a system large enough that fluctuations are small), and the First Law follows straightforwardly from energy conservation. Secondly, since the canonical distribution maximizes Gibbs entropy for given expectation values, the equilibration process is Gibbs entropy-non-decreasing. And since Hamiltonian flow is Gibbs entropy-conserving, the overall process of Hamiltonian flow plus equilibration is Gibbs-entropy-nondecreasing. The Second Law then also follows via the definition (15). The Accessibility Principle is somewhat harder to establish (and depends rather more on the details of the control theory) but if we assume that we can vary the parameters of the Hamiltonian arbitrarily slowly, the very fact that Gibbs entropy is maximized by the canonical distribution and so small perturbations from that distribution have lower entropy only at second and higher order in the perturbation establishes that we can carry out transitions with arbitrarily low entropy increase (see (Wallace 2014, section 4) for the details.)

Alas, it's (probably) not that simple. Precisely because equilibration thus defined is Gibbs entropy-increasing, equilibration cannot be a Hamiltonian process. And if equilibration is just a playing out of the ordinary, Hamiltonian (or Schrödinger) dynamics of a system in isolation, then equilibration is a Hamiltonian process. Only if equilibration requires fundamentally new dynamics, and/or relies essentially on the interaction with a system with its environment, could our naive control theory be correct. Serious cases have been made for both,[10] but the general consensus in foundations of statistical mechanics, with which I concur, is that it ought to be possible to understand equilibration without either. If so, this naive control theory cannot be the true underpinning of phenomenological thermodynamics.

But it can suggest a better control theory. In orthodox accounts of the approach to equilibrium the idea is not that an equilibrated system *really and truly* is described by the canonical distribution, but that *for all intents and purposes* it is, at least in macroscopically large systems: any remotely-realistically measurable dynamical quantity, on any remotely-realistic timescale, will have the same expectation values on the canonical distribution as on the true distribution. Only if *per impossibile* we were to measure something like an N-particle correlation function for N of the order of the entire system size could we detect any difference between the two. If so, it seems reasonable to suppose that remotely-realistic control operations likewise fail to distinguish between the true post-equilibration state and a canonical state with the same expectation values and parameters.

To express this precisely, it will be helpful to be somewhat more general. A *coarse-graining map* J is an assignment to each parameter value V of a projection J_V on the space of distributions (that is, a map of distributions to distributions satisfying $J_V^2 = J_V$) with the additional properties that

1. The expectation value of the Hamiltonian at parameter V, and any other conserved quantity N, is conserved under J_V: $\langle H(V)\rangle_{J_V\rho} = \langle H(V)\rangle_\rho$ and $\langle N(V)\rangle_{J_V\rho} = \langle N(V)\rangle_\rho$.

2. The Gibbs entropy satisfies $S_G(J_V\rho) \geq S_G(\rho)$, with equality only if $J_V\rho = \rho$; that is, if $J_V\rho \neq \rho$, $S_G(J_V\rho) > S_G(\rho)$.

The coarse-graining map induces a projection on the set of states: $J(\rho, V) = (J_V\rho, V)$, and defines a function, the *thermodynamic entropy with respect to J* ('J-entropy' for short) on states:

$$S^J(\rho, V) = S_G[J_V\rho]. \tag{20}$$

The most important example of coarse-graining is *equilibration coarse-graining*, defined by

$$J_V^{eq}\rho = \rho_c(\langle H(V)\rangle_\rho, V, \langle N(V)\rangle_\rho) \tag{21}$$

but many others will be useful in due course.

A control theory is *forward compatible with J*, or J-compatible for short, if for any allowable control operations $C_1, C_2, \ldots C_N$ whose composition is defined and allowable, and for any allowable initial state (ρ, V),

$$JC_NJC_{N-1}\cdots JC_1J(\rho, V) = J(C_N\cdots C_1)(\rho, V). \tag{22}$$

[10] See, e. g. , (Sklar 1993, 246–254) and references therein; see also (Albert 2000, ch.7).

It is *forward complete with respect to* J, or J-complete for short, if

(i) For any distribution ρ and parameter V, $(J\rho, V)$ is an allowable initial state.

(ii) For any allowable initial state (ρ_1, V), and any state (ρ_2, V_2) with

$$S^J(\rho_2, V_2) > S^J(\rho_1, V_1)$$

there is an allowable control operation that takes (ρ_1, V_1) to some state (ρ_2', V_2) with $J_{V_2}\rho_2' = J_{V_2}\rho_2$.

It is *strictly* J*-complete* if in addition

(iii) For any allowable initial state (ρ_1, V), and any state (ρ_2, V_2) with

$$S^J(\rho_2, V_2) = S^J(\rho_1, V_1)$$

there is an allowable control operation that takes (ρ_1, V_1) to some state (ρ_2', V_2) with $J_{V_2}\rho_2' = J_{V_2}\rho_2$

(in other words, if (ii) holds when the entropy is nondecreasing and not just strictly increasing.)

This will require some unpacking. Forward compatibility with J means that control operations give the same result, up to coarse graining, whether they are applied to the actual initial state or a coarse-graining of it, and furthermore that this continues to be the case if control operations are composed (provided that the composition remains allowed). This is to say that the control operation can be understood as acting autonomously on the coarse-grained features of the system, irrespective of the fine-grained details. A control theory forward compatible with equilibration coarse-graining, for instance, takes effectively-canonical distributions to effectively-canonical distributions in a fashion that does not depend on the in-practice-inaccessible ways in which those distributions differ from the exact canonical distribution; this (I suggest) accurately captures the intuitive idea, stated above, that 'remotely-realistic control operations ... fail to distinguish between the true post-equilibration state and a canonical state with the same expectation values and parameters'.

The J-entropy is non-decreasing, and in general strictly increasing, under volume-preserving (and in particular Hamiltonian) control operations:

$$\begin{aligned} S^J(C(\rho, V)) &\equiv S_G[J_{V^C}\Pi_C\rho] \\ &= S_G[J_{V^C}\Pi_C J_V\rho] \\ &\geq S_G[\Pi_C J_V\rho] \\ &= S_G[J_V\rho] \\ &\equiv S^J(\rho, V). \end{aligned} \tag{23}$$

J-completeness expresses, up to coarse-graining, the idea that this limit is the *only* limit on the control theory: in a J-complete control theory, given any state there is another allowable initial state with the same coarse-grained description, and given any pair of states where the second has higher J-entropy than the first, there is a control operation that transforms the first into one identical to the second up to coarse-graining. (And strict J-completeness means that this continues to hold even when the two states have the same J-entropy.)

Let's now specialize to the specific case of the equilibration coarse-graining ma J^{eq}, which replaces any state with the canonical state at the same expected values of conserved quantities. I define an *equilibrium thermodynamics* of a thermodynamic system as any control theory which is forward compatible and forward complete with respect to J^{eq}. (We have seen that there are good (if heuristic) reasons to think that any control theory that allows arbitrarily slow variation of the parameters of a system's Hamiltonian will define an equilibrium thermodynamics; we have also seen that any two control theories forward-compatible and forward-complete with respect to a given coarse-graining are essentially equivalent, so that it will rarely matter which equilibrium thermodynamics we have in mind.)

The thermodynamic entropy, with respect to J^{eq}, is determined only by the expected value of the conserved quantities and the external parameters:

$$S^{J^{eq}}(\rho, V) = S_G[\rho_c(\rho_c(\langle H(V)\rangle_\rho, V, \langle N(V)\rangle_\rho)] \tag{24}$$

and indeed any state can be characterized completely by parameters and conserved-quantity expectations. At this point it should be apparent that equilibrium thermodynamics simply reproduces the canonical recipe.

6 Partial equilibrium

To unpick this further, suppose that we have a thermodynamic system which can be decomposed into subsystems 1,2 (with dynamical state spaces S_1, S_2), and that each subsystem i has a parameter V_i such that the total Hamiltonian $H(V_1, V_2)$ can be written at least to a good approximation as

$$H(V_1, V_2) \simeq H_1(V_1) \times \mathrm{id}_2 + \mathrm{id}_1 \times H_2(V_2) \tag{25}$$

i. e. the two systems are very weakly interacting. (For the moment suppose no conserved quantities are present other than energy.) Then the equilibration coarse-graining map is

$$J^{eq}_{V_1,V_2}\rho = \rho_c(\langle H(V_1, V_2)\rangle_\rho, V_1, V_2) \tag{26}$$

or more explicitly (and writing $A \otimes B$ as a neutral expression for either the function $A \otimes B(x, y) = A(x)B(y)$ (classical) or for the tensor product of A and B (quantum)),

$$J^{eq}_{V_1,V_2}\rho = \frac{\mathrm{e}^{-(H_1(V_1)\times \mathrm{id}_2 + \mathrm{id}_1 \times H_2(V_2))/T}}{Z(T)} \simeq \frac{\mathrm{e}^{-H_1(V_1)/T}}{Z_1(T)} \otimes \frac{\mathrm{e}^{-H_2(V_2)/T}}{Z_2(T)} \tag{27}$$

where $T = T(\langle H(V)\rangle_\rho, V_1, V_2)$ is the thermodynamic temperature given the expected energy and parameters of ρ and

$$Z_i = \int_i \mathrm{e}^{-H_i/T} \tag{28}$$

and the approximation becomes exact in the limit where the total Hamiltonian factorizes exactly. That is: equilibration coarse-graining transforms a distribution to the product of two canonical distributions at the same thermodynamic temperature.

This coarse-graining map is appropriate for the control theory of an agent who is unable to prevent interactions between the two systems, and who acts on timescales slow compared to the equilibration timescale of the joint system: by the time they are able to intervene on the system, the two subsystems have (effectively) reached their joint equilibrium state. But we can also consider an agent who can intervene more rapidly, and who can control whether or not interactions occur between the two systems (but who cannot intervene rapidly compared to the equilibration timescales of the two systems considered separately, and lacks fine-grained control of those separate systems). This latter agent is better modeled via the *partial equilibration coarse-graining*, where the coarse-graining map takes each subsystem to its own equilibrium state, as determined by its own expected energy.

More precisely: for a given distribution ρ for the system, we define

$$\rho|_1 = \int_{S_2} \rho \tag{29}$$

as the marginal distribution of ρ over S_1: again this is a neutral notation, denoting either the marginal probability distribution (classical) or the partial trace (quantum). Then partial equilibration coarse-graining is defined by

$$J^{pe}_{V_1,V_2}\rho = J^{eq,1}_{V_1}\rho|_1 \otimes J^{eq,2}_{V_2}\rho|_2 \tag{30}$$

where $J^{eq,i}_{V_i}$ is the usual equilibration coarse-graining map for the ith system. Explicitly, we have

$$J^{pe}_{V_1,V_2}\rho = \frac{\mathrm{e}^{-H_1(V_1)/T_1}}{Z_1(T_1)} \otimes \frac{\mathrm{e}^{-H_2(V_2)/T_2}}{Z_2(T_2)} \tag{31}$$

where $T_i = T_i(\langle H_i(V_i)\rangle_{\rho|_i}, V_i)$ is the thermodynamic temperature of system i. So partial equilibration coarse-graining again transforms a distribution into a product of canonical distributions, but now they may be at different temperatures.

Since the canonical distribution of the whole system is the lowest-energy state for given energy and parameters, it follows that if we have access to a control theory that is complete and compatible with respect to *partial* equilibration — to *partial-equilibrium thermodynamics*, that is — we can extract energy in this situation in a cyclic process, just by transforming (31) reversibly to the full (same-temperature) canonical distribution at the same parameter values. Indeed, the explicit process for doing so is familiar: we simply run a Carnot cycle between the two systems and extract the energy, continuing until they are at the same thermodynamic temperature.

This is our first example of how a more powerful control theory may allow us to extract more energy from a system than is possible in equilibrium thermodynamics. Note that it requires two things: a control theory (approximately) forward complete with respect to a finer-grained coarse-graining, but also an initial state whose coarse-graining under the new theory's coarse-graining map is not canonical. If the initial state of the system is the product of two canonical distributions at the same temperature, the additional power of partial-equilibrium thermodynamics is useless; conversely, if we have access only to (full) equilibrium thermodynamics then it is useless to us to know that the two subsystems are initially at different temperatures.

To illustrate further, suppose for simplicity that the two subsystems are identical and have no adjustable parameters, and define $\rho(U)$ as the canonical distribution for either of the subsystems at expected energy U. Let $U_C < U_H$ be two energies (for a 'cold' and 'hot' system), and let the initial distribution be

$$\rho = \frac{1}{2}\left(\rho(U_C)\otimes\rho(U_H) + \rho(U_H)\otimes\rho(U_C)\right). \tag{32}$$

Equilibration coarse-graining this system gives

$$J^{eq}\rho = \rho((U_C+U_H)/2)\otimes\rho((U_C+U_H)/2). \tag{33}$$

Partial-equilibration coarse-graining gives the same result, since the marginal distribution of each system is $1/2(\rho(U_C)+\rho(U_H))$. The thermodynamic entropies of ρ are the same in either full-equilibrium or partial-equilibrium thermodynamics, and in both cases are maximal for the given energy, reflecting the fact that neither control theory can extract energy from the system. In the full-equilibrium case this is just because the controller cannot intervene on the separate systems, at any rate not quickly enough; in the latter case, if only the controller knew which system was hotter they could run a heat engine between them, but they do not.

But now suppose that the controller learns (say, from a helpful third party who had already measured it) that in fact the system's state is

$$\rho' = \rho(U_C)\otimes\rho(U_H). \tag{34}$$

This makes no difference to the effect of equilibration coarse-graining, and hence no difference to the equilibrium-thermodynamics entropy: the system's total energy is still U_C+U_H, and that alone determines the global equilibrium distribution. But ρ' is invariant under *partial* equilibration coarse-graining, and so has a lower thermodynamic entropy, reflecting the fact that energy can be extracted from the system: now that the controller knows which is the hotter system, they can run a reversible heat engine between the two and extract the resultant energy.

This is a special case of a general principle which will recur. Gaining information about a system always lowers the fine-grained Gibbs entropy, but it lowers the *thermodynamic entropy* only with respect to a control theory able to exploit that information.

(There is, to be sure, a certain arbitrariness as to whether we regard a given control problem as *partial*-equilibrium thermodynamics, or as *full*-equilibrium thermodynamics with additional conserved quantities. We could have set up our problem *ab initio* as being defined by the same Hamiltonian and parameters but by an additional conserved quantity $H_1 \otimes \mathrm{id}_2$, at least in the idealized limit where the two systems are exactly isolated from one another. The full-equilibrium control theory of the latter system is identical to the partial-equilibrium control theory of the former, though the definition of 'cyclic process' is narrower for the latter.)

As a second illustration,[11] suppose our system contains two boxes of gas, and that the gas is a mixture of two difficult-to-distinguish isotopes (for definiteness let the gas be chlorine, and the isotopes be ^{35}Cl and ^{37}Cl). The numbers of the two isotopes are separately conserved, and we assume that both boxes begin at the same thermodynamic temperature. Now consider two controllers, Alice and Bob, both of whom hope to extract energy through cyclic processes. Alice has access only to the full-equilibrium control theory for the system; Bob possesses semipermeable membranes that allow him, e.g., to allow ^{35}Cl but not ^{37}Cl isotopes to flow between the two boxes.

If the initial state of the system has a 50/50 mixture[12] of both isotopes present in each box separately, Bob's control theory grants him no advantage over Alice's (and both agree on the thermodynamic entropy). Likewise, if the initial state is an equally-weighted mixture of (a) all the ^{35}Cl in the left hand box, all the ^{37}Cl in the right hand box, and (b) vice versa, Bob and Alice agree on the coarse-grained state, on the thermodynamic entropy, and on the impossibility of cyclicly extracting energy. But suppose they learn that in fact all the ^{35}Cl is on the left and all the ^{37}Cl is on the right. The information is useless to Alice — but it allows Bob, using his semipermeable membranes, to allow the ^{35}Cl to reversibly expand into both boxes and then to do likewise with the ^{37}Cl, in each case extracting energy from the pressure exerted on the membrane as it is slowly moved. (As always, genuine reversibility occurs only in the unphysical — and useless — limit of infinitely slow movement, but can be approximated arbitrarily well.)

If before either agent can act the partition between the two boxes is removed and the gases are allowed to mingle freely, Alice and Bob will disagree about its significance. To Alice this makes no thermodynamically-relevant difference to the state: coarse-graining before or after has the same effect. To Bob, it is an irreversible process that increases the thermodynamic entropy and wastes energy that could otherwise have been extracted.

[11] Readers familiar with the Gibbs paradox (see, e. g. , (Saunders 2018) and references therein) will recognize its close connection to this example; for reasons of space I do not explore them here.

[12] Ordinary chlorine has a 75/25 mixture of the two isotopes; I use 50/50 for simplicity.

7 Free energy

The common theme of these partial-equilibrium processes is that they involve extracting energy from a system in a cyclic process by reversibly guiding it to the canonical state. In fact this is a perfectly general feature of thermodynamic control theories, as we can see by returning to the general coarse-graining framework. Suppose we are operating with a control theory forward-compatible with a coarse-graining J; then for any state (ρ, V), the *free energy of equilibration* of that state relative to J is defined by

$$E^J(\rho, V) = \langle H(V) \rangle_\rho - U(S^J(\rho, V), V, \langle N(V) \rangle_\rho). \tag{35}$$

Since $U(S^J(\rho, V), V, \langle N(V) \rangle_\rho)$ is the minimum expected energy of any state with the same J-entropy and $\langle N(V) \rangle$ as (ρ, V), and since any J-compatible control operation is J-entropy-nondecreasing, the free energy of equilibration is the maximum work extractable from (ρ, V) by J-compatible cyclic control operations. That maximum can be approached arbitrarily closely in a J-complete control theory (and reached in a strictly J-complete control theory).

Similarly, suppose $(\rho, V) \rightarrow (\rho', V')$ is the result of a *non*-cyclic (but J-compatible) control operation, and suppose the J-entropy and conserved-quantity expectation values before and after are respectively S, N and S', N'. Then the work extracted by that process is

$$\begin{aligned} W &= \langle H(V) \rangle_\rho - \langle H(V') \rangle_{\rho'} \\ &= (E^J(\rho, V) + U(S, V, N)) - (E^J(\rho', V') + U(S', V',' N')) \\ &= (E^J(\rho, V) - E^J(\rho', V')) + (U(S, V, N) - U(S', V', N'). \end{aligned} \tag{36}$$

That is: extracted work equals decrease in free energy of equilibration, plus decrease in canonical energy. And the difference between any control theory and equilibrium control theory is just given by the fact that states may have free energy of equilibration according to the first control theory, whereas no state in equilibrium control theory has free energy of equilibration.

We can use the free-energy concept to analyze the partial-equilibrium examples in the previous section. Given two systems at different temperatures, and taking J to be partial equilibration, the free energy of equilibration is just the work that is extractable by letting the two systems reversibly equilibrate, extracting work as we do so. If we let the systems *irreversibly* equilibrate, not all the free energy will be extracted as work; some will simply be lost. In the limit when we just put the systems in thermal contact and leave them to equilibrate, all the free energy will be lost. Quantitatively, we have (writing S_1 and S_2 as the thermodynamic entropies for the two systems separately and $U_i(S_i)$ as the canonical energy of the ith system)

$$E(S_1, S_2) = U_1(S_1) + U_2(S_2) - U(S_1 + S_2) \tag{37}$$

($(U_1(S_1) + U_2(S_2))$ is the actual energy of the system at entropies S_1, S_2; $U(S_1 + S_2)$ is the minimum energy it can have for that total entropy.) Since U is an increasing function of entropy, we can see directly that the work extracted from the system (= the decrease in $(U_1 + U_2)$) is less than or equal to the decrease in free energy, with equality only when the process is reversible, i. e. entropy-conserving.

If the heat flow between the two systems is reversible, so that overall entropy is conserved, the change in the free energy when a small amount of entropy δS flows from system 1 to system 2 is just minus the total change in energy, so we have

$$\delta E = -(T_1 - T_2)\delta S. \tag{38}$$

Similarly, for the case of the two boxes of chlorine gas, if the two systems have chemical potentials μ_1, μ_2 with respect to number of ^{35}Cl atoms and are both at temperature T, then if some small quantity δN of atoms flows from box 1 to box 2 and the overall process causes entropy increases δS_1, δS_2 in the two boxes, the change in energy of the boxes will be

$$\delta U_1 + \delta U_2 = (\mu_2 - \mu_1)\delta N + T(\delta S_1 + \delta S + 2) \tag{39}$$

and since the last term here is just the change in the canonical energy of the combined system,

$$\delta E = -(\mu_1 - \mu_2)\delta N \tag{40}$$

which is the familiar result that work can be extracted from combining two samples of a fluid iff their chemical potentials differ. (Note that since N is conserved even in irreversible processes this expression holds whether or not the process is reversible; if it is irreversible, though, it will heat up the gas and so the energy extracted will be less than the decrease in free energy.)

In the partial-equilibrium case, we can express the free-energy idea another way: the largest amount of work we can extract from a system through partial-equilibrium control processes is the difference between their total energy and the minimum value of that energy at constant entropy (and other conserved parameters). Since the minimum-energy state is also the overall equilibrium state, we can also use this to characterize overall equilibrium for a collection of subsystems: at equilibrium, energy is minimized with respect to entropy-conserving processes and any process that transfers conserved quantities from one system to another.

8 Free energy and heat baths

(This section lies somewhat outside the main line of development of this paper and may be skipped on a first reading, other than the definition of a heat bath; its primary purpose is to connect the use of 'free energy' in this paper with other uses in the physics literature.)

Suppose we have a system consisting of two subsystems, one of which is a *heat bath*: a system with no adjustable parameters, no conserved quantities other than energy, and whose initial state is[13] the canonical state, and so large that energy flows into and out of it negligibly affect its temperature, so that we can in practice treat it as permanently at thermodynamic temperature T. And suppose that we are doing equilibrium thermodynamics with respect to this system, so that the primary (i. e. , non-heat-bath) system begins and ends every control process at (effective) thermal equilibrium with the heat bath. (We could generalize this to partial-equilibrium thermodynamics, with respect to various subsystems of the primary system, but with the additional constraint that all the subsystems remain in *thermal* equilibrium with one another and with the heat bath, even if they do not equilibrate with respect to other conserved quantities.) Then we can derive an effective control theory for the primary system alone, treated now not as isolated but as in thermal contact with the heat bath.

To do so, consider a control operation that takes the primary system from an initial state with entropy, parameter values, and conserved quantities (S, V, N), to another where these quantities are (S', V', N'). Since the primary system must begin and end at temperature T, there will in general be some energy flow between the heat bath and the primary system during the course of the control operation. If this flow changes the heat-bath energy by ΔU_{HB}, we must have $\Delta U = T\Delta S_{HB}$, where ΔS_{HB} is the entropy change of the heat bath. The work W extracted by the control operation is then $-\Delta U_{HB}$ plus the decrease in energy of the system itself:

$$W = -U(S', V', N') + U(S, V, N) - T\Delta S_{HB}. \tag{41}$$

But since $\Delta S_{HB} + (S' - S')$ is the total change in the entropy, which must be ≥ 0, this can be rewritten as

$$W \leq -(U(S', V', N') - TS') + (U(S, V, N) - TS) \tag{42}$$

If we invert the equation of state so that S may be written as a function of T, V, N, we can define the quantity

$$F(T, V, N) = U(S(T, V, N), V, N) - TS(T, V, N) \tag{43}$$

which is called the *Helmholtz free energy.* Up to an additive constant, it equals the free energy of equilibration of the total system. For a system in contact with a heat bath, the Helmholtz free energy plays a similar role as the energy does for a thermally isolated system:

- The maximum work extractable from a system by any transformation in equilibrium thermodynamics equals the decrease in Helmholtz free energy.
- The free energy of equilibration of a collection of subsystems in a partial-equilibrium thermodynamics (but with all systems in thermal contact with the heat bath) is the sum of the Helmholtz free energies of the subsystems, minus the value of that sum at its lowest value (for given T, V, N for the combined system).

[13] For the purposes of this section it suffices to require that the heat bath's state coarse-grains to the canonical state.

- The maximum work extractable from a system by any transformation in partial-equilibrium thermodynamics equals the decrease in the total Helmholtz free energies of the subsystems under the transformation.
- The equilibrium state of a collection of subsystems can be characterized by the fact that it minimizes Helmholtz free energy under any transformation that leaves parameters and non-energy conserved quantities unchanged.

Other effective control theories can be defined similarly:

- By considering a primary system separated by an adiabatic barrier from a very large 'pressure bath' system at generalized pressure P, so that the barrier can adjust to equalize pressure between the two systems but the systems cannot otherwise interact, we find that the role of energy is played by the *enthalpy*,

$$H(S,P,N) = U(S,V(S,P,N),N) + PV(S,P,N). \qquad (44)$$

- By combining heat and pressure baths together, so that the external path constrains the primary system to have both fixed pressure and fixed temperature, we find that the role of energy is played by the *Gibbs free energy*,

$$G(T,P,N) = U(S(T,P,N),V(T,P,N),N) + PV(T,P,N) - TS(T,P,N). \qquad (45)$$

9 Beyond equilibration: Boltzmannian thermodynamics

In this section I want to consider a class of control theories not built on the assumption of even partial equilibration: control theories based on Boltzmann's idea of a partition of a system 's state space. (These are of limited practical use but are often conceptually helpful.) To describe that class, suppose for simplicity[14] that we are dealing with a classical system with phase space $\mathcal{P}$, and that its phase space has been partitioned into *macrostates*: subsets of the phase space, of positive measure, such that intuitively two phase-space points in the same macrostate are indistinguishable to an external observer. (For instance, in a dilute gas the macrostates can be defined by coarse-graining the phase space of a single particle into cells of equal Liouville volume, and then individuating the macrostates by the number of particles in each cell.) More precisely, the partition may depend on any external parameter V: write $\mathcal{M}(V)$ for the partition for parameter V.

In the thermodynamic context, 'macroscopically indistinguishable' has an operational meaning: it implies that a controlling agent can move a system from macrostate to macrostate but has no control of a system on a grain finer

[14] For the quantum version, replace the partition with a collection of mutually-orthogonal finite-dimensional subspaces whose direct sum is the full Hilbert space.

than that defined by the partition. We can incorporate this by defining a *Boltzmannian coarse-graining* which smears out these fine-grained details. To state it, first define for any macrostate M the projection operator P_M, which acts on distributions as

$$P_M\rho(x) = \begin{cases} \rho(x) & \text{if } x \in M. \\ 0 & \text{if } x \notin M. \end{cases} \tag{46}$$

The probability of a macrostate given a distribution ρ, $\Pr(M|\rho)$, is then

$$\Pr(M|\rho) = \int P_M\rho. \tag{47}$$

And now we can define our coarse-graining map:

$$J_V^B\rho = \sum_{M\in\mathcal{M}(V)} \Pr(M|\rho)\frac{\chi_M}{\mathcal{V}(M)} \tag{48}$$

where χ_M is the characteristic distribution of M, given by

$$\chi_M(x) = \begin{cases} 1 & \text{if } x \in M. \\ 0 & \text{if } x \notin M \end{cases} \tag{49}$$

and $\mathcal{V}(M)$ is the Liouville volume of M.[15] In other words, $J^B\rho$ is the distribution that assigns the same probability to each macrostate as ρ itself, but which is uniform across each macrostate.

The thermodynamic entropy defined by this coarse-graining is the *generalized Boltzmann entropy.* In the special case where ρ has support entirely on one macrostate M, it is a function of M alone and reduces to the *Boltzmann entropy*:

$$S_B(M) = \ln \mathcal{V}(M). \tag{50}$$

In the more general case it is the sum of two terms: the expected value of the Boltzmann entropy with respect to the probability distribution across macrostates, and the Shannon entropy of that distribution:

$$S^{J_B}(\rho, V) = \left(\sum_{M\in\mathcal{M}(V)} \Pr(M|\rho)S_B(M)\right) + \left(-\sum_{M\in\mathcal{M}(V)} \Pr(M|\rho)\ln\Pr(M|\rho)\right). \tag{51}$$

In the special case where (i) ρ has support only on one macrostate M, and (ii) the control operation maps it to another distribution with support only on another macrostate M' (call such a control operation *macrodeterministic* for M), it is immediate from volume preservation that $\mathcal{V}(M)' \geq \mathcal{V}(M)$, i. e. Boltzmann entropy is nondecreasing. But of course the generalized Boltzmann entropy is nondecreasing under any control map that is forward compatible with Boltzmannian coarse-graining, as a consequence of our general framework.

[15] Quantum mechanically: take χ_M to be the projector onto macrostate M, define $P_M\rho = \chi_M\rho\chi_M$, and take $\mathcal{V}(M)$ to be the dimension of M (equivalently, $\mathcal{V}(M) = \mathsf{Tr}\chi_M$).

Boltzmannian thermodynamics offers an agent significantly more control over a system than equilibrium thermodynamics, and so significantly more opportunity to extract work, given an appropriate initial state. For instance, suppose that our system consists of a gas of N particles in a box of volume V, and that in fact the initial distribution has all the particles localized in a smaller region of the box, with volume $V' < V$. This additional information is useless to a controller who only has control operations forward commuting with equilibration, for whom the situation is not interestingly different from one where the gas is at full thermal equilibrium. But a Boltzmannian agent can, e.g., slam a partition into the box so that the particles remain confined to the region of volume V', and then slowly allow the gas to expand, doing work against the partition, until it returns to volume V. The work extracted is the free energy of equilibration of the state relative to Boltzmannian coarse-graining. Once it is extracted, though, the greater control power of Boltzmannian thermodynamics is of no further use, and we effectively return to equilibrium thermodynamics.

This difference between the control theories also shows up in the different entropies they assign. The equilibrium-thermodynamics entropy is as usual set only by the volume of the box and by the total energy and number of particles; the Boltzmannian entropy is set by the size of the volume in which the particles actually are (and so is lower than the equilibrium-thermodynamics entropy by $N \ln(V/V')$). And so being told that the particles are in fact on one side of the box makes no difference to equilibrium-thermodynamic entropy, but decreases Boltzmannian entropy. There is no fact of the matter as to which entropy is *correct*, but the *appropriate* choice is set by the control theory being used.

10 Reversible thermodynamics

We have seen that coarse-grained thermodynamics has a very general form. The coarse-graining, together with the initial state, determines the free energy of equilibration. In evolving the system to a state which coarse-grains to the canonical state, it may be possible to extract some of that free energy as work (it can all be extracted only in the ideal limit of a control theory strictly complete with respect to J. Once it is extracted, we can extract any more energy only by varying the system's control parameters and/or the values of other conserved quantities, and the amount of energy thus extractable does not depend on the coarse-graining operation.

The limiting case of all of this is *reversible thermodynamics*, where the control operations include *every Hamiltonian operation.* Reversible thermodynamics is forward-compatible with only the trivial coarse-graining, defined by the identity map; its thermodynamic entropy is just the Gibbs entropy. The inequality that thermodynamic entropy increases under control operations continues to hold, but is always saturated, and simply reflects conservation of Gibbs entropy under dynamical flow. Reversible thermodynamics is reversible both in the literal sense that the time-reverse of any allowed control operation is also an allowed control operation, and in the formal sense that — unlike our various coarse-grained versions of thermodynamics — the entropy is not only nondecreasing but is strictly conserved.

Insofar as thermodynamics just reflects the macroscopic limitations of what we can do to a system, we would expect that in reversible thermodynamics — where there are no such limitations — there are no constraints on how much energy we can extract beyond those set by the conservation laws, so that it should be possible to extract all of a system's energy through a cyclic operation. But this is not the case. Assume for simplicity that the energy has been scaled so that the system's ground state energy is zero. The free energy of equilibration of a state (ρ, V) with respect to the trivial coarse graining is just the difference between its actual expected energy and its canonical energy with respect to the trivial coarse-graining:

$$E(\rho, V) = \langle H(V)\rangle_\rho - U(S_G(\rho), V, \langle N(V)\rangle_\rho). \tag{52}$$

Only if the latter equals zero will the free energy equal the total system energy. And the canonical energy can be zero, in general, only if the initial distribution is known *exactly*. More precisely: working in quantum mechanics, and assuming the nondegeneracy of the ground state, only if ρ is pure will the canonical energy equal the ground-state energy. In this case a control operation can steer the system's state reversibly into the ground state and extract all the energy. But if the initial distribution is mixed then no unitary operation can so steer it, and the lowest-energy state will still assign some probability to excited states. Similarly, in classical mechanics, and assuming that the gradient of the energy function nowhere vanishes, the subset of lowest-energy states will have Liouville measure zero; only if the initial state likewise has support on a measure-zero region will it be possible to steer it into the lowest-energy region by a Hamiltonian control operation.

(Can all the free energy actually be extracted in reversible thermodynamics? In general, no: extracting it requires us to steer the system's distribution into a *canonical* distribution, and in general that will not be possible. Specializing for simplicity to quantum mechanics, and following Pusz and Woronowicz (1978) (see also the very helpful discussion in section 4 of (Maroney 2007)) we can define a quantum distribution as *passive* iff no unitary control operation decreases its expected energy; it can be shown that a distribution is passive iff it is a mixture of energy eigenstates where the probability of an energy eigenstate is nondecreasing with energy, but there are passive distributions with strictly higher energy than the canonical distribution at the same Gibbs entropy. For small systems it is an open question exactly how to characterize the extractable energy (the associated research project is sometimes called 'single-shot thermodynamics'; see (Yunger Halpern and Renes 2016) for an introduction). However, for sufficiently large systems the difference is generally expected to be negligible, albeit I am not aware of rigorous theorems to this effect: such systems can generally be expected to factorize into a large number of approximately-isolated subsystems so that the overall density matrix is approximately a product of identical density matrices, and Pusz and Woronowicz prove that the canonical distribution is the only distribution whose N-fold product is passive for all N.)

Reversible thermodynamics is the limiting case of the successively more fine-grained thermodynamic control theories we have been considering. In those theories, we have seen that the finer our control over a system, the more valuable it is to have fine-grained information about the system's state. In reversible thermodynamics, *any such information* can be used to extract work: in the limiting case, if we are told the exact microstate of a system (or, in quantum mechanics, if the system's state is pure and we know that state), we can extract all of its energy.[16] So to an agent with access to reversible thermodynamics, any information about the state is entropy-decreasing. But the most any such agent can do is extract the free energy from the system. Once that is done, any further work extraction has to be done by varying parameters and conserved quantities, and all the power of reversible thermodynamics avails us nothing beyond what equilibrium thermodynamics already permitted.

Free energy also provides a third way in which reversible thermodynamics is indeed reversible (and our various coarse-grained thermodynamic theories are irreversible). In coarse-grained thermodynamics, the change in free energy in a process is an upper limit on the energy extractable by that process, but it is possible to dissipate free energy without extracting any work from the system. In partial-equiilibrium thermodynamics, for instance, if we just let two systems at different temperatures equilibrate without running a heat pump between them, the free energy is simply and irreversibly lost. This never happens in reversible thermodynamics, where the decrease in free energy always equals the energy extracted. (In reversible thermodynamics, we could simply run the equilibration process backwards and return the systems to their distinct temperatures.)

11 Collecting information

We have seen that the more fine-grained an agent's control over a system, the more important are the fine-grained details of the system's initial state, and the more the entropy they assign can drop if they receive information about the system. On the face of it, an obvious strategy presents itself: instead of just carrying out a Hamiltonian control operation, the agent should first gain more information about the system (causing them to update the state) and only then apply a control operation, tailored to the information they collect.

It is easy to see that this cannot actually work, at least if the information-collection process conforms to the laws of physics. As I noted in section 4, any information-collection process can be mechanized, and the system can be expanded to include both the original system and that mechanism; we can then include 'turning on the mechanism' as just another Hamiltonian control operation. Or put another way, the agent themselves can be mechanized and included in the system. From that agent's point of view, they are collecting information and carrying out control operations conditional on the result, but

[16] In classical physics, if we literally know the *exact* microstate of a system we can use it as a resource to extract all the energy from any system: its entropy is $-\infty$. Of course this is an artifact of classical mechanics; it has no quantum equivalent.

from the point of view of the *ur*-agent outside the box in which the agent and their system sit, telling the inside agent to do their thing is a control operation that involves no information-gathering and no conditional control.

Still, it is instructive to ask what actually goes wrong with the collect-information strategy. The answer is well known[17]: any such process involves an ancilla system that records the result of the information-collection; that ancilla is effectively a finite resource which can become exhausted, so that no further information-collection is possible without resetting the ancilla; the resource cost of the reset process undoes the gain of collecting the information in the first place. For completeness, and to connect with this paper's approach to thermodynamics, I outline it here.

Suppose, specifically, that we have some system S with no adjustable parameters and which begins in the canonical distribution for some given temperature T, so that (assuming quantum mechanics for definiteness) the specific form of the distribution is

$$\rho_S = \frac{1}{Z}\sum_n \mathrm{e}^{-E_n/T} \left|n, S\right\rangle \left\langle n, S\right| \tag{53}$$

where $|n, S\rangle$ is an eigenstate of S's Hamiltonian H_S with eigenvalue E_n. Since ρ_S minimizes expected energy for given Gibbs entropy, no unitary process can extract energy from the system.

The collect-information strategy would have us measure the system in the energy basis, and then if the result is that the system's state is $|n, S\rangle$, apply a unitary transformation U_n which implements $U_n |n, S\rangle = |n, 0\rangle$, so that the system ends up determinately in the ground state. But of course no unitary process acting on S alone can have this effect, since it would have to map orthogonal states to the same state. The only way to implement the procedure is to have some ancilla system A whose state after the process records the result. Concretely, if the ancilla system starts in some fixed state $|0, A\rangle$, we can find a unitary process U with the effect

$$U \left|n; S\right\rangle \otimes \left|0; A\right\rangle = \left|0; S\right\rangle \otimes \left|\psi_N; A\right\rangle \tag{54}$$

with $\langle \psi_n, A|\psi_m, A\rangle = \delta_{m,n}$. (Often this is illustrated by supposing that (i) the state of the system is measured and recorded in the ancilla, and then subsequently (ii) an operation is performed on the system conditional on the state of the ancilla. But we do not need to make this separation: the crucial point is that the ancilla must end up in a state that records the original state of the system.) At the level of distributions, this unitary transformation enacts

$$U(\rho_S \otimes \left|0, A\right\rangle \left\langle 0, A\right|)U^\dagger = \left|0; S\right\rangle \left\langle 0; S\right| \otimes \rho_A \tag{55}$$

[17] See, e. g. , (Bennett 2003) or the overview and articles in (Leff and Rex 2002). The topic has been rather controversial in recent philosophy of physics (Earman and Norton 1999; Bub 2002; Norton 2005; Ladyman, Presnell, Short, and Groisman 2007; Norton 2011; Ladyman and Robertson 2013; Norton 2013a; Norton 2013b; Ladyman and Robertson 2014; Myrvold 2021); I do not attempt to engage with these controversies here.

where

$$\rho_A = \frac{1}{Z}\sum_n \mathrm{e}^{-E_n/T}\,|n;A\rangle\,\langle n;A|\,. \tag{56}$$

Does this *conditional process* extract net energy (measured as always with respect to the self-Hamiltonian of the system before or after the control intervention, i. e. in the case the sum of the Hamiltonians H_S and H_A of system and ancilla)? It depends on the details of the record basis $|n;A\rangle$ and the ancilla's Hamiltonian; certainly there is no systematic reason that it will not. (If the ancilla has enough states that each $|\psi_n;A\rangle$ has negligible energy compared to the original energy of the system, energy can certainly be extracted.) But whether or not energy is extracted, the Gibbs entropy of the ancilla post-process now equals the Gibbs entropy of the system pre-process, say S.

If ρ_A is not the canonical distribution for the ancilla at entropy S, then there is some free energy available by reversibly steering it to that distribution; we may as well suppose this done, so that we extract that free energy together with the energy extracted by the conditional process. (Of course we may not be able to extract *all* the free energy, but if so that only reduces the total work extractable.) The end result is that we reversibly transformed a system whose state was initially a product of a canonical distribution at entropy S and a pure state, to another system whose state is a product of a pure state and a canonical distribution at entropy S.

But we could have done that anyway, without the need for anything so complicated. Suppose we had instead (i) extracted the ancilla's initial free energy by reversibly transforming it into its ground state, and then (ii) run a Carnot cycle between system and ancilla to reversibly cool the system into its ground state and heat up the ancilla. The result is exactly the same, and so extracts exactly the same energy — and clearly is in general strictly less efficient than just running the Carnot cycle until system and ancilla reached the same temperature. Other than the possibility of extracting the ancilla's initial free energy, our information-collection and conditional-operation process gains us nothing that we could not gain just by letting the system and ancilla reversibly equilibrate.

(If instead we wanted to return the ancilla to its initial state so as to allow us to use it again to extract energy from another system, we would have to dump its Gibbs entropy into still another system. If, for instance, we have available a heat bath at temperature T, doing so would have energy cost TS.)

The example is simple, but the point is perfectly general. An ancilla system is a potential thermodynamic resource. Its free energy of equilibration can be extracted by reversibly transforming it to a canonical distribution; having done so, if the resultant distribution has a thermodynamic temperature different from other systems then further energy can be extracted through a reversible heat engine. Using the ancilla to record the result of measurements on another system and then cool that system is just one, complicated, way of running that heat engine: it will extract net energy if the thermodynamic temperature of the ancilla (that is, the temperature of the canonical distribution to which it can be unitarily transformed) is lower than that of the system, and will actually cost energy if it is higher. In the classic example where the ancilla starts off in a known pure state, its thermodynamic temperature is zero, and systems at absolute zero are genuinely useful resources to extract energy from a finite-temperature. But that extraction possibility has nothing really to do with the possibility of gathering information, and works only until the ancilla has been brought up to thermal equilibrium with the system.

I have been assuming reversible thermodynamics throughout this section, but the basic points generalize to a coarse-grained thermodynamics. In the latter theory, some of what reversible thermodynamics treats as free energy may after all not be extractable — may not be free according to the coarse-grained theory — and it might be that a specific thermodynamics can extract some part of the free energy only through processes naturally described as 'measure and then conditionally act'. But it is still the case that any such process can be thought of as extracting some part of the free energy of equilibration of the ancilla, and/or the joint free energy of equilibration of system and ancilla, and that the limits of reversible thermodynamics are upper limits for any thermodynamics.

12 The Second Law again

In my summary of phenomenological thermodynamics, I described the Second Law in fairly formal terms as the principle that no transition decreases thermodynamic entropy. And we have seen a fairly general derivation of that result in thermodynamic control theory, of which equilibrium thermodynamics is only a special case: if J is a coarse-graining, then the thermodynamic entropy with respect to J is nondecreasing under any J-compatible control process.

Historically and pedagogically, though, one often finds two more directly phenomenological versions of the Second Law: the Kelvin and Clausius statements, reviewed in section 2. Both have fairly straightforward translations into control-theoretic terminology. Recall from section 8 that a *heat bath* is a system with no adjustable parameters, no conserved quantities other than energy, and whose initial state is the canonical state, and so large that energy flows into and out of it negligibly affect its temperature, so that we can in practice treat it as permanently at thermodynamic temperature T. Heat baths are supposed to represent large systems which have achieved thermodynamic equilibrium in an uncontrolled way, so that an agent has no initial information about the system other than that it has equilibrated. (On some interpretations of quantum statistical mechanics, using an exact canonical state to describe the heat bath may reflect an agent's epistemic limitations; on others, it may reflect the fact that a heat bath is uncontrollably entangled with the environment and that its actual quantum state is canonical; it matters for my purposes only that we in fact represent heat baths that way.)

Kelvin Statement (precise control theory version): No control process can extract energy from a single heat bath without in doing so decreasing the free energy of some ancilla system or the joint free energy of an ancilla system and the heat bath.

Clausius Statement (precise control theory version): No control process with no overall energy cost can cause energy to be transferred from a heat bath to another heat bath at a higher thermodynamic temperature, without in doing so decreasing the free energy of some ancilla system or the joint free energy of an ancilla system and one of the heat baths.

Thus stated, both statements are direct consequences of our results so far. The most we can do in any thermodynamic control theory is extract all of the free energy. A heat bath in isolation has no free energy (even in reversible thermodynamics) and so any process which decreases the energy of a heat bath must do so by either decreasing the free energy of some other system by itself (say, to extract the energy to operate a heat pump) or by decreasing the free energy of the joint system of ancilla-plus-heat bath ((say, by running a heat engine between the bath and the ancilla). Similarly, any control process run on two heat baths without ancilla will extract negative energy if it decreases the energy of the lower-temperature system, so we an induce energy transfer from lower to higher temperature only by exploiting free energy provided by some ancilla, either in its own right or in combination with one or other heat bath.

And of course, any free energy provided by an ancilla is a finite resource: once extracted, it is gone. An alternative, and somewhat less formal, version of the two statements would be:

Kelvin Statement (informal control theory version): No indefinitely repeatable control process can extract energy from a single heat bath.

Clausius Statement (informal control theory version): No indefinitely repeatable control process with no overall energy cost can cause energy to be transferred from a heat bath to another heat bath at a higher thermodynamic temperature.

Importantly, neither the Kelvin nor the Clausius statements (unlike the statement that thermodynamic entropy is nondecreasing) make any reference to the coarse-graining operation: they apply to any thermodynamic control operation, even fully reversible thermodynamics.

13 Two kinds of Maxwell demon

To illustrate the difference between the nondecreasing-entropy form of the Second Law, and the Kelvin and Clausius statement, let's consider Maxwell's famous demon, which (recall) was originally conceived of as a 'very observant and neat-fingered being'[18] that could respond to the microscopic state of a gas by opening and closing a partition, and thus violate the Second Law.

[18] (Maxwell 1867); see (Leff and Rex 2002, pp.3-6) and (Myrvold 2011) for more on Maxwell's own conception of the Demon.

It is useful to distinguish two kinds of demon: a *Maxwell demon of the first kind* can decrease a system's thermodynamic entropy, while a *Maxwell demon of the second kind* can bring about a violation of the Kelvin or Clausius statements of the Second Law.[19] The difference between the two can be seen starkly by considering the practical uses to which they could be put:

- The main use of a demon of the first kind is to refute someone who claims that thermodynamic entropy is nondecreasing.
- The main use of a demon of the second kind is to run the electrical grid off the ambient air temperature, remove any real constraints human civilization faces from the scarcity of usable energy, and become (should you like that kind of thing) the richest person in the world.

Demons of the first kind are not so hard to come by. They exploit the fact that thermodynamic entropy is defined relative to a choice of coarse-graining, and is guaranteed to be non-decreasing only under control operations forward compatible with that coarse-graining. So, given the thermodynamic entropy defined by a coarse-graining J, all you need to build a demon of the first kind is some J-incompatible control operation (and an initial state non-invariant under J). For instance:

1. Maxwell's original 'neat-fingered being' is a demon of the first kind with respect to equilibrium thermodynamics, provided it has a sufficiently large memory capacity to record the various measurements it needs to make of the system's microstate. Eventually that memory will fill, but until then the demon can certainly lower the temperature of the gas just as Maxwell proposed. (But it could not do so if its memory subsystem was already in a canonical state at the same thermodynamic temperature as the gas.)

2. Given two boxes containing different isotopes of chlorine at the same temperature and pressure, any agent whose control operations can distinguish between isotopes is a demon of the first kind with respect to a control theory that cannot so distinguish them: to an agent using the latter control theory, another agent using the former theory will appear to have achieved a miraculous cooling of the gases. (But they could not achieve it if the two gases were already fully mixed.)

3. In the spin-echo experiment[20] a system of coupled spins apparently relaxes to thermal equilibrium but then returns to its original state when the time reverse of the original operation is applied. This can be thought of as a demon of the first kind relative to ordinary equilibrium thermodynamics: the time-reversed operation is not forward compatible with equilibration coarse-graining. (Some of the researchers on spin-echo (Rhim, Pines, and

[19] This distinction was first made in print by Myrvold (2020, section 8); as he acknowledges, it is drawn from an early version of the present paper.

[20] (Hahn 1953; Rhim, Pines, and Waugh 1971); there is a good conceptually-focused discussion in (Sklar 1993, pp.219–222).

Waugh 1971) described it as a 'Loschmidt demon', referring to Loschmidt's famous time-reversal objection to Boltzmann.)

As a slight variant on these, we can consider a demon that can reduce Boltzmann entropy (here it will be helpful, to avoid issues with the quantum measurement problem, to work in classical physics). The Boltzmann entropy has the useful feature that it is a function of the system's microstate and not of its full probability distribution (taking the Boltzmann entropy of a phase-space point to be the Boltzmann entropy — i. e., the logarithm of the phase-space volume — of the unique macrostate in which that phase-space point is situated). We saw in section 9 that a control operation that is macrodeterministic for a macrostate M will be Boltzmann-entropy-nondecreasing (since all points in M must be mapped into some fixed macrostate M', and so phase space volume conservation means that M' has at least as high a volume as M). But even a control theory compatible with Boltzmannian coarse-graining can decrease Boltzmann entropy, provided that it is not macrodeterministic:

(i) A process that decreases Boltzmann entropy on average (that is: decreases the expected value of Boltzmann entropy) might still, through fluctuations, happen to decrease it in a specific instance.

(ii) A sufficiently macro-indeterministic process could even decrease average Boltzmann entropy. We saw in section 9 that the generalized Boltzmann entropy — which is strictly nondecreasing under Boltzmann-compatible control operations — is the sum of two terms: the expected value of the Boltzmann entropy and the Shannon entropy of the probability distribution over macrostates. By spreading the initial distribution very widely across macrostates, a process could increase the latter term enough that the former term might decrease, perhaps significantly. (Albert (2000, ch.5) and Hemmo and Shenker (2012, ch.13) give detailed constructions of demons of this kind.)

Of course, for such demons the availability of many macrostates is itself a resource, which can be depleted: at least for systems contained in a finite region, the Shannon entropy cannot increase indefinitely, and when it has reached a maximum any further Boltzmann-compatible control operations must increase expected Boltzmann entropy. (And note that if one's system is not confined to a finite region, it is usually possible to extract all its energy even without a demon: adiabatically expanding a cylinder of gas to an arbitrarily large volume will reduce its temperature arbitrarily close to absolute zero.)

Demons of the first kind are curiosities. The really interesting possibility would be a demon of the second kind — but this is not a real possibility as long as the underlying dynamics are unitary (or, in the classical regime, Hamiltonian). We saw in section 12 that the Kelvin and Clausius statements are absolute, not relativized to any notion of coarse-graining, and that they follow from unitarity alone. And so demons of the second kind are in principle impossible. It is not the crude imprecisions of our macroscopic tools that prevents us building the kind of perpetual motion machine that the Kelvin and Clausius statements forbid: it is the fundamental laws of physics.

14 Conclusions

This paper offers a rather general framework in which to discuss various forms of thermodynamics. The notion of a coarse-graining projector, and of a control theory forward compatible and (at least in idealization) forward complete with respect to that projector, provides a powerful method to characterize that control theory without needing to attend to its microphysical details. The framework is broad enough to encompass all the concrete versions of thermodynamic control theory I know, from the various versions of equilibrium thermodynamics, to Boltzmannian thermodynamics, to fully reversible thermodynamics.

The framework allows us to characterize a J-compatible control theory in terms of the thermodynamic entropy and free energy with respect to J, and to establish that the most any control operation can do is extract as work the free energy plus any work extractable in equilibrium thermodynamics. It also gives a criterion for when knowing a system's state more precisely is useful: if and only if the new state has lower thermodynamic entropy with respect to J, which is to say: if and only if the coarse-graining of the new state has lower Gibbs entropy than that of the old state. Information is thus useful only where it concerns degrees of freedom over which we have control. At one extreme, in equilibrium thermodynamics the entropy is set only by parameters and conserved quantities, and fine-grained information is useless; at another extreme, in reversible thermodynamics any information is exploitable.

For a J-compatible control theory, there is a J-dependent specification of the Second Law in its entropy-is-nondecreasing form: the thermodynamic entropy S^J with respect to J is nondecreasing under J-compatible control operations. Assuming that the control theory contains at least some S^J-increasing control operations, this version of the Second Law characterizes the way in which the theory is irreversible: entropy can go up but never down. But a more powerful control theory might contain J-incompatible control operations and might very well be able to decrease S^J. This creates a general recipe for building Maxwell demons of the first kind, which can decrease entropy under a given definition: all they need to do is to be able to carry out control operations not adapted to that definition.

But there are other forms of the Second Law — the Kelvin and Clausius statements — that do not depend on J and which hold in any control theory, with or without irreversibility. It is these versions of the Second Law that best express the hard constraints Nature puts on our ability to get work out from systems in the wild, and no Maxwell demon of the second kind, which would allow us to bypass them, can be built.

As a closing comment, note that the crucial feature of the dynamics that underwrites the microphysical justification of the Second Law (in any of its forms) is that they preserve Hilbert-space dimension or phase space volume. These are the characteristic features of *reversibility*, the natural generalizations to their respective continuum systems of the discrete-system idea that dynamical maps are 1:1. Genuinely irreversible (and deterministic) operations — operations that map many Hilbert-space vectors to the same vector, or phase-space regions to smaller regions — would, at least if combined with a sufficiently large class of reversible operations, make it straightforward to extract all of a system's energy and leave it in its ground state. The Second Law is often held up as the example *par excellence* of irreversibility: it is ironic that it is reversibility that provides its microphysical foundation.

Acknowledgements

I am grateful for conversations on earlier versions of this material with too many people to name over a number of years. Special thanks are due to David Albert, Owen Maroney, and Wayne Myrvold. I am also grateful to John Norton for help locating the authoritative source for Einstein's famous quote about thermodynamics, and to Wayne Myrvold and Katie Robertson for helpful comments on an earlier draft.

Appendix: properties of the canonical distribution

Theorem: Suppose $\mathcal{H}$ is a separable Hilbert space of dimension>1 and H is a self-adjoint operator on $\mathcal{H}$ (the Hamiltonian) with discrete spectrum whose lowest eigenvalue is nondegenerate (and whose expectation value I write as $\langle H \rangle$, or $\langle H \rangle_\rho$ when the specific density operator ρ needs to be stated). For each $\beta > 0$ define the canonical density operator $\rho_c(\beta)$ by

$$\rho_c(\beta) = \frac{\mathrm{e}^{-\beta H}}{Z(\beta)} \tag{57}$$

where

$$Z(\beta) = \mathsf{Tr}\mathrm{e}^{-\beta H}. \tag{58}$$

Let $S(\beta) = S_G[\rho_c(\beta)]$ be the Gibbs entropy of $\rho_c(\beta)$ and $U(\beta) = \langle H \rangle_{\rho_c(\beta)}$. Then:

(i) Both $U(\beta)$ and $S(\beta)$ and are decreasing functions of β.

(ii) The range of $S(\beta)$ is the interval $(0, \ln \dim \mathcal{H})$ (taking $\ln \dim \mathcal{H} = \infty$ if $\mathcal{H}$ is infinite-dimensional) and if $s \in (0, \ln \dim \mathcal{H})$, there is a unique β such that $s = S(\beta)$.

(iii) The range of $U(\beta)$ is some interval (E_0, E_{max}), where E_0 is the lowest eigenvalue of H, E_{max} may or may not be infinite, and for any u in this range, there is a unique β such that $U(\beta) = u$.

(iv) For any $u \in (E_0, E_{max})$, $\rho_c(U^{-1}(u))$ is the unique global maximum-Gibbs entropy density operator with $\langle H\rangle = u$.

(v) For any $s \in (0, \ln \dim \mathcal{H})$, $\rho_c(S^{-1}(s))$ is the unique global minimum-$\langle H\rangle$ density operator with Gibbs entropy s.

Proof:

(i) Differentiate U to obtain

$$U'(\beta) = -(\langle H^2\rangle - \langle H\rangle^2) \tag{59}$$

which is strictly negative (unless H is a constant, which it cannot be, as an operator with nondegenerate lowest eigenvalue on a Hilbert space of dimension>1). Then observe that

$$S(\beta) = \ln Z(\beta) + \beta U(\beta), \tag{60}$$

and differentiate to obtain $S'(\beta) = \beta U'(\beta)$.

(ii),(iii) Let the eigenstates of H, labelled in increasing order of eigenvalue, be $|0\rangle, |1\rangle, \ldots$. Then

$$\lim_{\beta\to\infty} \rho_c(\beta) = |0\rangle\langle 0| \tag{61}$$

so that

$$\lim_{\beta\to\infty} S(\beta) = 0 \tag{62}$$

and

$$\lim_{\beta\to\infty} U(\beta) = E_0. \tag{63}$$

Taking limits in (60) gives

$$\lim_{\beta\to 0} S(\beta) = \ln \dim \mathcal{H} \tag{64}$$

and we define

$$E_{max} = \lim_{\beta\to 0} U(\beta) \tag{65}$$

Uniqueness follows in each case because S and U are strictly decreasing.

(iv) Let $\mathcal{D}$ be the set of density operators on $\mathcal{D}$, and let $\rho \in \mathcal{D}$. Consider an infinitesimal variation $\rho \to \rho + \delta\rho$. The variations of Gibbs entropy, $\langle H\rangle$, and trace, respectively, are, to second order,

$$\begin{aligned} \delta S_G &= -\mathsf{Tr}(\{\ln\rho + \mathrm{id}\}\delta\rho) - \frac{1}{2}\mathsf{Tr}(\delta\rho^2\rho^{-1}) + o(\delta\rho^2) && (66)\\ \delta\langle H\rangle &= \mathsf{Tr}(H\delta\rho) && (67)\\ \delta\mathsf{Tr}\rho &= \mathsf{Tr}\delta\rho. && (68) \end{aligned}$$

Firstly, we establish that the stationary points of the Gibbs entropy under variations that leave $\langle H\rangle$ constant are exactly the canonical states (though possibly with $\beta < 0$). For ρ is a stationary point under variations that conserve trace and $\langle H\rangle$ iff the first-order variation of S_G is a linear combination of the first-order variations of $\langle H\rangle$ and $\mathsf{Tr}\rho$ (this is the method of Lagrange multipliers). This forces

$$\mathsf{Tr}(\{-\ln\rho - \mathrm{id} + \beta H + \alpha\mathrm{id}\}\delta\rho) = 0 \tag{69}$$

for some α, β and all $\delta\rho$, and we can solve to get $\rho = \rho_c(\beta)$ (where again β need not be positive).

Secondly, we establish that ρ_c is a local *maximum* of S_G under these variations. This follows from (66): the first order variation vanishes under the constraint that trace and $\langle H\rangle$ are invariant, and $\mathsf{Tr}(\delta\rho^2\rho^{-1})$ is a positive-definite quadratic function of $\delta\rho$, as can be verified explicitly by expanding in a basis of energy eigenstates: if $\delta\rho = \sum_{n,m}\delta\rho_{nm}$ then (writing $H\,|n\rangle = E_n\,|n\rangle$)

$$\mathsf{Tr}(\delta\rho^2\rho_c^{-1}(\beta)) = \frac{1}{Z(\beta)}\sum_{m,n}\delta\rho_{mn}\mathrm{e}^{\beta E_n}\delta\rho_{mn} = \frac{1}{Z(\beta)}\sum_{m,n}|\delta\rho_{mn}|^2\mathrm{e}^{\beta E_n}. \tag{70}$$

Finally, we establish that $\rho_c(U^{-1}(u))$ is a global maximum of entropy for $\langle H\rangle = u$. For suppose otherwise; then there would be another density operator ρ_0 with $S(\rho_0) \geq \rho_c(U^{-1}(u))$, and then the function

$$\rho(x) = x\rho_c(U^{-1}(u)) + (1-x)\rho_0 \tag{71}$$

would satisfy

$$S_G[\rho(x)] > S_G[\rho_c(U^{-1}(u)) \tag{72}$$

for $0 < x < 1$, which contradicts the claim that $\rho_c(U^{-1}(u))$ is a local maximum.

(v) Fix $\beta = S^{-1}(s)$, and suppose for contradiction that there is a density operator ρ with $\langle H\rangle_\rho \leq \langle H\rangle_{\rho_c(\beta)}$ and $S_G[\rho] = S_G[\rho_c(\beta)]$. Note firstly that by (iv) we must have $\langle H\rangle_\rho < \langle H\rangle_{\rho_c(\beta)}$, since otherwise $\rho_c(\beta)$ would not be the unique maximum-entropy density operator at $\langle H\rangle = \langle H\rangle_{\rho_c(\beta)}$. But then, again by (iv), there must exist β' such that $S_G[\rho_c(\beta')] \geq S_G[\rho]$; by (i) $\beta' > \beta$. But then $S_G[\rho_c(\beta')] \geq S_G[\rho_c(\beta)]$, in contradiction with (i).

References

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

Bennett, C. H. (2003). Notes on Landauer's principle, reversible computation, and Maxwell's Demon. *Studies in the History and Philosophy of Modern Physics 34*, 501–510.

Blundell, S. J. and K. M. Blundell (2010). *Concepts in Thermal Physics* (2nd ed.). Oxford: Oxford University Press.

Bub, J. (2002). Maxwell's demon and the thermodynamics of computation. *Studies in the History and Philosophy of Modern Physics 32*, 569–579.

Bustamante, C., J. Liphardt, and F. Ritort (2005). The nonequilibrium thermodynamics of small systems. https://arxiv.org/abs/cond-mat/0511629.

Callender, C. (2001). Taking thermodynamics too seriously. *Studies in the History and Philosophy of Modern Physics 32*, 539–553.

Collin, D., F. Ritort, C. Jarzynski, S. B. Smith, I. Tinoco, and C. Bustamante (2005). Verification of the Crooks fluctuation theorem and recovery of RNA folding free energies. *Nature 437*, 231–234.

Crooks, G. E. (1998). Nonequilibrium measurements of free energy differences for microscopically reversible Markovian systems. *Journal of Statistical Physics 90*, 1481.

Demerast, H. (2016). The universe had one chance. *Philosophy of Science 83*, 248–264.

Earman, J. and J. Norton (1999). EXORCIST XIV: The wrath of Maxwell's demon. part II. from Szilard to Landauer and beyond. *Studies in the History and Philosophy of Modern Physics 30*, 1–40.

Einstein, A. (1949). Autobiographical notes. In P. A. Schilpp (Ed.), *Albert Einstein: Philosopher-Scientist*, pp. 1–95. New York: MJF Books. English translation by Paul Arthur Schilpp.

Goldstein, S. (2001). Boltzmann's approach to statistical mechanics. In J. Bricmont, D. Dürr, M. Galavotti, F. Petruccione, and N. Zanghí (Eds.), *In: Chance in Physics: Foundations and Perspectives*, Berlin, pp. 39. Springer. Available online at http://arxiv.org/abs/cond-mat/0105242.

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

Hahn, E. (1953). Free nuclear induction. *Physics Today 6*, 4–9.

Hemmo, M. and O. Shenker (2012). *The Road to Maxwell's Demon: Conceptual Foundations of Statistical Mechanics*. Cambridge: Cambridge University Press.

Ismael, J. (2009). Probability in deterministic physics. *Journal of Philosophy 106*, 89–108.

Jarzynski, C. (1997). Nonequilibrium equality for free energy differences. *Physical Review Letters 78*, 2690.

Kittel, C. and H. Kroemer (1980). *Thermal Physics* (2nd ed.). W.H.Freeman.

Ladyman, J., S. Presnell, A. Short, and B. Groisman (2007). The connection between logical and thermodynamic irreversibility. *Studies in History and Philosophy of Modern Physics 38*(1), 58–79.

Ladyman, J. and K. Robertson (2013). Landauer defended: reply to Norton. *Studies in History and Philosophy of Modern Physics 44*, 263–271.

Ladyman, J. and K. Robertson (2014). Going round in circles: Landauer vs. Norton on the thermodynamics of computation. *Entropy 16*, 2278–2290.

Landau, L. and E. Lifshitz (1980). *Statistical Physics* (3rd ed.). Elsevier. English translation by J.B.Sykes and M.J.Kearsley.

Leff, H. and A. F. Rex (2002). *Maxwell's Demon: Entropy,Information, Computing* (2nd ed.). Institute of Physics Publishing.

Lewis, D. (1980). A subjectivist's guide to objective chance. In R. C. Jeffrey (Ed.), *Studies in Inductive Logic and Probability, volume II*. Berkeley: University of California Press. Reprinted, with postscripts, in David Lewis, *Philosophical Papers*, Volume II (Oxford University Press, Oxford, 1986); page numbers refer to this version.

Loewer, B. (2002). Determinism and chance. *Studies in the History and Philosophy of Modern Physics 32*, 609–620.

Maroney, O. (2007). The physical basis of the Gibbs-von Neumann entropy. https://arxiv.org/abs/quant-ph/0701127.

Maxwell, J. C. (1867). Letter to P.G.Tait, 11 December 1867. Reprinted in C.G. Knott, *Life and Scientific Work of Peter Guthrie Tait* (Cambridge University Press, London, 1911), 213–215.

Maxwell, J. C. (1871). *Theory of Heat*. London: Longmans, Green and Co.

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

Myrvold, W. C. (2020). The science of $\theta\Delta^{cs}$. *Foundations of Physics 50*, 1219–1251.

Myrvold, W. C. (2021). Shakin' all over: Proving Landauer's principle without neglect of fluctuations. Forthcoming in BJPS; online at https://www.journals.uchicago.edu/doi/10.1086/716211.

Norton, J. D. (2005). Eaters of the lotus: Landauer's principle and the return of Maxwell's demon. *Studies in the History and Philosophy of Modern Physics 36*, 375–411.

Norton, J. D. (2011). Waiting for Landauer. *Studies in History and Philosophy of Modern Physics 42*, 184–198.

Norton, J. D. (2013a). Author's reply to Landauer defended. *Studies in History and Philosophy of Modern Physics 44*, 272.

Norton, J. D. (2013b). The end of the thermodynamics of computation: a no-go result. *Philosophy of Science 80*, 1182–1192.

Pusz, W. and S. Woronowicz (1978). Passive states and KMS states for general quantum systems. *Communications in Mathematical Physics 1978*, 273–290.

Rhim, W.-K., A. Pines, and J. Waugh (1971). Time-reversal experiments in dipolar-coupled spin systems. *Physical Review B 3*, 684–696.

Saunders, S. (2018). The Gibbs paradox. *Entropy 20*, 552.

Sklar, L. (1993). *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics*. Cambridge: Cambridge University Press.

Throne, K. S. and R. D. Blandford (2017). *Modern Classical Physics: Optics, Fluids, Plasmas, Electricity, Relativity, and Statistical Physics*. Princeton: Princeton University Press.

Tolman, R. C. (1938). *The principles of statistical mechanics*. Oxford University P.

Wallace, D. (2012). *The Emergent Multiverse: Quantum Theory according to the Everett Interpretation*. Oxford: Oxford University Press.

Wallace, D. (2014). Thermodynamics as control theory. *Entropy 16*, 699–725.

Wallace, D. (2016). Probability and irreversibility in modern statistical mechanics: Classical and quantum. To appear in D. Bedingham, O. Maroney and C. Timpson (eds.), *Quantum Foundations of Statistical Mechanics* (Oxford University Press, forthcoming). Preprint at https://arxiv.org/abs/2104.11223.

Wallace, D. (2020). The necessity of Gibbsian statistical mechanics. In V. Allori (Ed.), *Statistical Mechanics and Scientific Explanation: Determinism, Indeterminism and Laws of Nature*. World Scientific.

Yunger Halpern, N. and J. M. Renes (2016). Beyond heat baths: Generalized resource theories forsmall-scale thermodynamics. *Physical Review E 93*, 022126.