Thermodynamics as control theory

David Wallace

July 26, 2013

Abstract

I explore the reduction of thermodynamics to statistical mechanics by treating the former as a control theory: a theory of which transitions between states can be induced on a system (assumed to obey some known underlying dynamics) by means of operations from a fixed list. I recover the results of standard thermodynamics in this framework on the assumption that the available operations do not include measurements which affect subsequent choices of operations. I then relax this assumption and use the framework to consider the vexed questions of Maxwell's demon and Landauer's principle. Throughout I assume rather than prove the basic irreversibility features of statistical mechanics, taking care to distinguish them from the conceptually distinct assumptions of thermodynamics proper.

1 Introduction

Thermodynamics is misnamed. The name implies that it stands alongside the panoply of other "X-dynamics" theories in physics: classical dynamics, quantum dynamics, electrodynamics, hydrodynamics, chromodynamics and so forth.[1] But what makes these theories *dynamical* is that they tell us how systems of a certain kind — classical or quantum systems in the abstract, or charged matter and fields, or fluids, or quarks and gluons, or whatever — evolve if left to themselves. The paradigm of a dynamical theory is a state space, giving us the possible states of the system in question at an instant, and a dynamical equation, giving us a trajectory (or, perhaps, a family of trajectories indexed by probabilities) through each state that tells us how that state will evolve under the dynamics.

Thermodynamics basically delivers on the state space part of the recipe: its state space is the space of systems at equilibrium. But it is not in the business of telling us how those equilibrium states evolve if left to themselves, except in the trivial sense that they do not evolve at all: that is what equilibrium means, after all. When the states of thermodynamical systems change, it is because

[1] In fact, the etymology of "thermodynamics", according to the Oxford English Dictionary, is just that it is the study of heat (thermo) and work (dynamics) and their interaction. (I am grateful to Jos Uffink for this observation.

we do things *to* them: we put them in thermal contact with other systems, we insert or remove partitions, we squeeze or stretch or shake or stir them. And the laws of thermodynamics are not dynamical laws like Newton's: they concern what we can and cannot bring about through these various interventions.

There is a general name for the study of how a system can be manipulated through external intervention: *control theory.* Here again a system is characterised by its possible states, but instead of a dynamics being specified once and for all, a range of possible control actions is given. The name of the game is to investigate, for a given set of possible control actions, the extent to which the system can *be* controlled: that is, the extent to which it can be induced to transition from one specified state to another. The range of available transitions will be dependent on the forms of control available; the more liberal a notion of control, the more freedom we would expect to have to induce arbitrary transitions.

This conception of thermodynamics is perfectly applicable to the theory understood phenomenologically: that is, without any consideration of its microphysical foundations. However, my purpose in this paper is instead to use the control-theory paradigm to explicate the relation between thermodynamics and statistical mechanics. That is: I will begin by assuming the main results of non-equilibrium statistical mechanics and then consider what forms of control theory they can underpin. In doing so I hope to clarify both the control-theory perspective itself and the reduction of thermodynamics to statistical mechanics, as well as providing some new ways to get insight into some puzzles in the literature: notably, those surrounding Maxwell's Demon and Landauer's Principle.

In sections 2 and 3, I review the core results of statistical mechanics (making no attempt to justify them). In sections 4 and 5 I introduce the general idea of a control theory and describe two simple examples: adibatic manipulation of a system and the placing of systems in and out of thermal contact. In sections 6–8, I apply these ideas to construct a general account of classical thermodynamics as a control theory, and demonstrate that a rather minimal form of thermodynamics possesses the full control strength of much more general theories; I also explicate the notion of a one-molecule gas from the control- (and statistical-mechanical) perspective). In the remainder of the paper I extend the notion of control theory to consider systems with feedback, and demonstrate in what senses this does and does not increase the scope of thermodynamics.

I develop the quantum and classical versions of the theory in parallel, and fairly deliberately flit between quantum and classical examples. When I use classical examples, in each case (I believe) the discussion transfers straightforwardly to the quantum case unless noted otherwise. The same is *probably* true in the other direction; if not, no matter, given that classical mechanics is of (non-historical) interest in statistical physics only insofar as it offers a good approximation to quantum mechanics.

2 Statistical-mechanical preliminaries

Statistical mechanics, as I will understand it in this paper, *is* a theory of dynamics in the conventional sense: it is in the business of specifying how a given system will evolve *spontaneously.* For the sake of definiteness, I lay out here exactly what I assume to be delivered by statistical mechanics.

1. The *systems* are classical or quantum systems, characterised *inter alia* by a classical phase space or quantum-mechanical Hilbert space Hamiltonian $H[V_I]$ which may depend on one or more external parameters V_I (in the paradigm case of a gas in a box, the parameter is volume). In the quantum case I assume the spectrum of the Hamiltonian to be discrete; in either case I assume that the possible values of the parameters comprise a connected subset of R^N and that the Hamiltonian depends smoothly on them.

2. The *states* are probability distributions over phase space, or mixed states in Hilbert space. (Here I adopt what is sometimes called a *Gibbsian* approach to statistical mechanics; in Wallace (2013a), I defend the claim that this is compatible with a view of statistical mechanics as entirely objective.)

3. Given two systems, their *composite* is specified by the Cartesian product of the phase spaces (classical case) or by the tensor product of the Hilbert spaces (quantum case), and by the sum of the Hamiltonians (either case).

4. The *Gibbs entropy* is a real function of the state, defined in the classical case as
$$S_G(\rho) = -\int \mathrm{d}x\, \rho(x) \ln \rho(x) \tag{1}$$
and in the quantum case as
$$S_G(\rho) = -\mathsf{Tr}\left(\rho \ln \rho\right). \tag{2}$$

5. The *dynamics* are given by some flow on the space of states. In Hamiltonian dynamics this would be the flow generated by Hamilton's equation from the Hamiltonian $H[V_I]$, under which the Gibbs entropy is a constant of the motion; in statistical mechanics, however, we assume only that the flow (a) is entropy-non-decreasing, and (b) conserves energy, in the sense that the probability given by the state to any given energy is invariant under the flow.

6. For any given system there is some time, the *equilibration timescale*, after which the system has evolved to that state which maximises the Gibbs entropy subject to the conservation constraint above.

Now, to be sure, it is controversial at best *how* statistical mechanics delivers all this. In particular, we have good reason to suppose that isolated (classical or quantum) systems ought really to evolve by Hamiltonian dynamics, according to which the Gibbs entropy is constant and equilibrium is never achieved; more generally, the statistical-mechanical recipe I give here is explicitly time-reversal-noninvariant, whereas the underlying dynamics of the systems in question have a time reversal symmetry.

There are a variety of responses to offer to this problem, among them:

- Perhaps no system can be treated as isolated, and interaction with an external environment somehow makes the dynamics of any realistic system non-Hamiltonian.
- Perhaps the probability distribution (or mixed state) needs to be understood not as a property of the physical system but as somehow tracking our ignorance about the system's true state, and the increase in Gibbs entropy represents an increase in our level of ignorance.
- Perhaps the true dynamics is not, after all, Hamiltonian, but incorporates some time-asymmetric correction.

My own preferred solution to the problem (and the one that I believe most naturally incorporates the insights of the "Boltzmannian" approach to statistical mechanics) is that the state ρ should not be interpreted as the true probability distribution over microstates, but as a coarse-grained version of it, correctly predicting the probabilities relevant to any macroscopically manageable process but not correctly tracking the fine details of the microdynamics, and that the true signature of statistical mechanics is the possibility of defining (in appropriate regimes, under appropriate conditions, and for appropriate timescales) autonomous dynamics for this coarse-grained distribution that abstract away from the fine-grained details.

But from the point of view of understanding the reduction of thermodynamics to *statistical* mechanics, all this is beside the point. The most important thing to realise about the statistical-mechanical results I give above is that *manifestly they are correct*: the entire edifice of statistical mechanics (a) rests upon them, and (b) is abundantly supported by empirical data. (See Wallace (2013b) for more on this point.) There is a foundational division of labour here: the question of how this machinery is justified given the underlying mechanics is profoundly important, but it can be distinguished from the question of how thermodynamics relates to statistical mechanics. Statistical mechanics is a thoroughly successful discipline in its own right, and not merely a foundational project to shore up thermodynamics.

3 Characterising statistical-mechanical equilibrium

The "state which maximises the Gibbs entropy" can be evaluated explicitly. If the initial state ρ has a definite energy U, it will evolve to the distribution with the largest Gibbs entropy for that energy, and it is easy to see that (up to normalisation) in the classical case this is the uniform distribution on the hypersurface $H[V_I](x) = U$, and that in the quantum case it is the projection onto the eigensubspace of $\widehat{H}[V_I]$ with energy U. Writing ρ_U to denote this state, it follows that in general the equilibrium state achieved by a general initial ρ will be that statistical mixture of ρ_U that gives the same probability to each energy as ρ did. In the classical case this is

$$\rho \longrightarrow \int \mathrm{d}U \, \Pr(U) \rho_U \tag{3}$$

where

$$Pr(U) = \int \rho\delta(H - U);$$

in the quantum case, it is

$$\rho \longrightarrow \sum_i \Pr(U_i) \rho_U \tag{4}$$

where the sum is over the distinct eigenvalues U_i of the Hamiltonian, $\Pr(U_i) = \mathsf{Tr}(\rho\Pi_i)$, and Π_i projects onto the energy U_i subspace. I will refer to states of this form (quantum or classical) as called *generalised equilibrium* states.

We can define the *density of states* $\mathcal{V}(U)$ at energy U for a given Hamiltonian H in the classical case as follows: we take $\mathcal{V}(U)\delta U$ to be the phase-space volume of states with energies between U and $U + \delta U$. We can use the density of states to write the Gibbs entropy of a generalised equilibrium state explicitly as

$$S_G(\rho) = \int \mathrm{d}U \, \Pr(U) \ln \mathcal{V}(U) + \left(- \int \mathrm{d}U \, \Pr(U) \ln \Pr(U) \right). \tag{5}$$

In the quantum case it is instead

$$S_G(\rho) = \sum_i \Pr(U_i) \ln(\mathrm{Dim}\ U_i) + \left(- \sum_i \Pr(U_i) \ln \Pr(U_i) \right) \tag{6}$$

where $\mathrm{Dim}(U_i)$ is the dimension of the energy-U_i subspace. Normally, I will assume that the quantum systems we are studying have sufficiently close-spaced energy eigenstates and sufficiently well-behaved states that we can approximate this expression by the classical one (defining $\mathcal{V}\delta U$ as the total dimension of eigensubspaces with energies between U and $U + \delta U$, and $\Pr(U)\delta U$ as the probability that the system has one of the energies in the range $(U, U + \delta U)$).

Now, suppose that the effective spread ΔU over energies of a generalised equilibrium state around its expected energy U_0 is narrow enough that the Gibbs entropy can be accurately approximated simply as the logarithm of $\mathcal{V}(U_0)$. States of this kind are called *microcanonical equilibrium states*, or *microcanonical distributions* (though the term is sometimes reserved for the ideal limit, where $\Pr(U)$ is a delta function at U_0, so that $\rho(x) = (1/\mathcal{V}(U_0))\delta(H(x) - U_0))$. A generalised equilibrium state can usefully be thought of as a statistical mixture of microcanonical distributions.

If ρ is a microcanonical ensemble with respect to $H[V_I]$ for particular values of the parameters V_I, in general it will not be even a generalised equilibrium state for different values of those parameters. However, if close-spaced eigenvalues of the Hamiltonian remain close-spaced even when the parameters are changed, ρ will *equilibrate* into the microcanonical distribution. In this case, I will say that the system is *parameter-stable*; I will assume parameter stability for most of the systems I discuss.

A microcanonical distribution is completely characterised (up to details of the precise energy width δU and the spread over that width) by its energy U and the external parameters V_I. On the assumption that $\mathcal{V}(U)$ is monotonically increasing with U for any values of the parameters (and, in the quantum case, that the system is large enough that we can approximate $\mathcal{V}(U)$ as continuous) we can invert this and regard U as a function of Gibbs entropy S and the parameters. This function is (one form of) the *equation of state* of the system: for the ideal monatomic gas with N mass-m particles, for instance, we can readily calculate that

$$\mathcal{V}(U,V) \propto V^N (2mU)^{3N/2-1} \tag{7}$$

and hence (for $N \gg 1$)

$$S \simeq S_0 + N \ln V + (3N/2) \ln U, \tag{8}$$

which can be inverted to get U in terms of V and S.

The *microcanonical temperature* is then defined as

$$T = \left(\frac{\partial U}{\partial S}\right)_{V_I} \tag{9}$$

(for the ideal monatomic gas, it is $2U/3N$).

At the risk of repetition, it is not (or should not be!) controversial that these probability distributions are empirically correct as regards predictions of measurements made on equilibrated systems, both in terms of statistical averages and of fluctuations around those averages. It is an important and urgent question *why* they are correct, but it is not our question.

4 Adiabatic control theory

Given this understanding of statistical mechanics, we can proceed to the control theory of systems governed by it. We will develop several different control theories, but each will have the same general form, being specified by:

- A *controlled object*, the physical system being controlled.
- A set of *control operations* that can be performed on the controlled object.
- A set of *feedback measurements* that can be made on the controlled object.

- A set of *control processes*, which are sequences of control operations and feedback measurements, possibly subject to additional constraints and where the control operation performed at a given point may depend on the outcomes of feedback measurements made before that point.

Our goal is to understand the range of transitions between states of the controlled object that can be induced. In this section and the next I develop two extremely basic control theories intended to serve as components for thermodynamics proper in section 6.

The first such theory, *adiabatic control theory*, is specified as follows:

- The controlled object is a statistical-mechanical system which is parameter-stable and initially at microcanonical equilibrium.
- The control operations consist of (a) smooth modifications to the external parameters of the controlled object over some finite interval of time; (b) leaving the controlled object alone for a time long compared to its equilibration timescale.
- There are no feedback measurements: the control operations are applied without any feedback as to the results of previous operations.
- The control processes are sequences of control operations ending with a leave-alone operation.

Because of parameter stability, the end state is guaranteed to be not just at generalised equilibrium but at microcanonical equilibrium. The control processes therefore consist of moving the system's state around in the space of microcanonical equilibrium states. Since for any value of the parameters the controlled object's evolution is entropy-nondecreasing, one result is immediate: the only possible transitions are between states x, y with $S_G(y) \geq S_G(x)$. The remaining question is: which such transitions are possible?

To answer this, consider the following special control processes: a process is *quasi-static* if any variations of the external parameters are carried out so slowly that the systems can be approximated to any desired degree of accuracy as being at or extremely close to equilibrium throughout the process.

Now, consider some very small segment δt of an quasi-static control process (and suppose for simplicity that there is only one external parameter V). At the beginning of the segment, if the energy of the system is U, then the system's state ρ is (or is very near to) the unique equilibrium state specified by U and V. During the segment, V changes to $V + \delta V$ and ρ transitions to some new state $\rho + \delta\rho$, which has energy $U + \delta U$. We will attempt to find $\delta\rho$ and δU to first order in δV; taking δt sufficiently small and summing will then give us an expression for the total change in each.

To find $\delta\rho$, we can consider the transition to occur in two steps: first V is changed suddenly to $V + \delta V$, and then ρ evolves under the new dynamics for time δt. The energy change can therefore be attributed entirely to the change in δV, and the entropy change entirely to the subsequent evolution of the system.

(Of course, at a sufficiently high level of accuracy the pattern of imposition of the change in V will affect the evolution, but we can make this effect arbitrarily small by a sufficiently small choice of δt.)

However, since the process is quasi-static, $\rho + \delta\rho$ is the equilibrium state at the new values of U and V, and as such lies at an entropy maximum among such states. So to first order in δV, ρ has the same entropy as $\rho + \delta\rho$ (the derivative of entropy in any direction along the expected-energy hypersurface from an equilibrium state is zero).

To summarise: *quasi-static adiabatic processes are isentropic*: they do not induce changes in system entropy. What about non-quasi-static adiabatic processes? Well, if at any point in the process the system is not at (or very close to) equilibrium, by the baseline assumptions of statistical mechanics it follows that its entropy will increase as it evolves. So an adiabatic control process is isentropic if quasi-static, entropy-increasing otherwise.

In at least some cases, the result that quasi-static adiabatic processes are isentropic does not rely on any explicit equlibration assumption. To be specific: if the Hamiltonian has the form

$$\widehat{H}[V_I] = \sum_i U_i(\Lambda_I) \left|\psi_i(\Lambda_I)\right\rangle \left\langle\psi_i(\Lambda_I)\right| \tag{10}$$

then the *adiabatic theorem* of quantum mechanics[2] tells us that if the parameters are changed sufficiently slowly from λ_I^0 to λ_I^1 then (up to phase, and to an arbitrarily high degree of accuracy) the Hamiltonian dynamics will cause $\left|\psi_i(\lambda_I^0)\right\rangle$ to evolve to $\left|\psi_i(\lambda_I^1)\right\rangle$; hence, in this regime the dynamics takes microcanonical states to microcanonical states of the same energy.

In any case, we now have a complete solution to the control problem. By quasi-static processes we can move the controlled object's state around arbitrarily on a given constant-entropy hypersurface; by applying a *non*-quasi-static process we can move it from one such hypersurface to a higher-entropy hypersurface. So the condition that the final state's entropy is not lower than the initial state's is sufficient as well as necessary: adiabatic control theory allows a transition between equilibrium states iff it is entropy-nondecreasing.

A little terminology: the *work done* on the controlled object under a given adiabatic control process is just the change in its energy, and is thus the same for any two control processes that induce the same transition, and it has an obvious physical interpretation: the work done is the energy cost of inducing the transition by any physical implementation of the control theory. (In phenomenological treatments of thermodynamics it is usual to assume some independent understanding of "work done", so that the observation that adiabatic transitions from x to y require the same amount of work however they are performed becomes contentful, and is one form of the First Law of Thermodynamics; from our perspective, though, it is just an application of conservation of energy.)

Following the conventions of thermodynamics, we write $đW$ for a very small quantity of work done during some part of a quasi-static control process. We

[2] See, e. g. , Messiah (1962, ch.XVII sections 10-14) or Weinberg (2013, pp.193-6).

have

$$đW = \mathrm{d}U|_{\delta S=0} = \sum_I \left(\frac{\partial U}{\partial V_I}\right)_{V_J,S} \mathrm{d}V_I \equiv -\sum_I P^I \mathrm{d}V_I \tag{11}$$

where the derivative is taken with all values of V_J other than V_I held constant and the last step implicitly defines the *generalised pressures.* (In the case where V_I just is the volume, P^I is the ordinary pressure.)

5 Thermal contact theory

Our second control theory, *thermal contact theory*, is again intended largely as a tool for the development of more interesting theories. To develop it, suppose that we have two systems initially dynamically isolated from one another, and that we introduce a weak interaction Hamiltonian between the two systems. Doing so, to a good approximation, will leave the internal dynamics of each system largely unchanged but will allow energy to be transferred between the systems. Given our statistical-mechanical assumptions, this will cause the two systems (which are now one system with two almost-but-not-quite-isolated parts) to proceed, on some timescale, to a *joint* equilibrium state. When two systems are coupled in this way, we say that they are in *thermal contact.* Given our assumption that the interaction Hamiltonian is small, we will assume that the equilibration timescales of each system separately are very short compared to the joint equilibration timescale, so that the interaction is always between systems which separately have states extremely close to the equilibrium state.

The result of this joint equilibration can be calculated explicitly. If two systems each confined to a narrow energy band are allowed to jointly equilibrate, the energies of one or other may end up spread across a wide range. For instance, if one system consists of a single atom initially with a definite energy E and it is brought in contact with a system of a great many such atoms, its post-equilibration energy distribution will be spread across a large number of states. However, for the most part we will assume that the microcanonical systems we consider are not induced to transition out of microcanonical equilibrium as a consequence of joint equilibration; systems with this property I call *thermally stable.*

There is a well-known result that characterises systems that equilibrate with thermally stable systems which is worth rehearsing here. Suppose two systems have density-of-state functions $\mathcal{V}_1$, $\mathcal{V}_2$ and are initially in microcanonical equilibrium with total energy U. The probability of the two systems having energies U_1, U_2 is then

$$\Pr(U_1, U_2) \propto \mathcal{V}(U_1)\mathcal{V}(U_2)\delta(U_1 + U_2 - U) \tag{12}$$

and so the probability of the first system having energy U_1 is

$$\Pr(U_1) \propto \mathcal{V}(U_1)\mathcal{V}(U - U_1). \tag{13}$$

Assuming that the second system is thermally stable, we express the second term on the right hand side in terms of its Gibbs entropy and expand to first order around U (the assumption that the second system's energy distribution is narrow tells us that higher terms in the expansion will be negligible):

$$\mathcal{V}(U-U_1)=\exp(S(U-U_1))\simeq\exp\left\{S(U)-\left(\frac{\partial S}{\partial U}\right)_{V_I}U_1\right\}. \tag{14}$$

Since the partial derivative here is just the inverse of the microcanonical temperature T of the second system, the conclusion is that

$$\Pr(U_1)\propto\mathcal{V}(U_1)\mathrm{e}^{-U_1/T}, \tag{15}$$

which is recognisable as the *canonical* distribution at canonical temperature T.

In any case, so long as we assume thermal stability then systems placed into thermal contact may be treated as remaining separately at equilibrium as they evolve towards a joint state of higher entropy.

We can now state thermal contact theory:

- The controlled object is a fixed, finite collection of mutually isolated thermally stable statistical mechanical systems.
- The available control operations are (i) placing two systems in thermal contact; (ii) breaking thermal contact between two systems; (iii) waiting for some period of time.
- There are no feedback measurements.
- The control processes are arbitrary sequences of control operations.

Given the previous discussion, thermal contact theory shares with adiabatic control theory the feature of inducing transitions between systems at equilibrium, and we can characterise the evolution of the systems during the control process entirely in terms of the energy flow between systems. The energy flow between two bodies in thermal contact is called *heat.* (A reminder: strictly speaking, the actual amount of heat flow is a probabilistic quantity very sharply peaked around a certain value.)

The quantitative *rate* of heat flow between two systems in thermal contact will of course depend *inter alia* on the precise details of the coupling Hamiltonian between the two systems. But in fact the *direction* of heat flow is independent of these details. For the total entropy change (in either the microcanonical or canonical framework) when a small quantity of heat $\text{đ}Q$ flows from system A to system B is

$$\delta S=\delta S_A+\delta S_B=\left\{-\left(\frac{\partial S_A}{\partial U_A}\right)_{V_i}+\left(\frac{\partial S_A}{\partial U_A}\right)_{V_i}\right\}\text{đ}Q. \tag{16}$$

But since the thermodynamical temperature T is just the rate of change of energy with entropy while external parameters are held constant, this can be rewritten as

$$\delta S=(1/T_B-1/T_A)\text{đ}Q. \tag{17}$$

So heat will flow from A to B only if the inverse thermodynamical temperature of A is lower than that of B. In most cases (there are exotic counter-examples, notably in quantum systems with bounded energy) thermodynamical temperature is positive, so that this can be restated as: heat will flow from A to B only if the thermodynamical temperature of A is greater than that of B. For simplicity I confine attention to this case.

If we define two systems as being in *thermal equilibrium* when placing them in thermal contact does not lead to any heat flow between them, then we have the following thermodynamical results:

1. Two systems each in thermal equilibrium with a third system are at thermal equilibrium with one another; hence, thermal equilibrium is an equivalence relation. (The *Zeroth Law of Thermodynamics*).

2. There exist real-valued *empirical temperature functions* which assign to each equilibrium system X a temperature $t(X)$ such that heat flows from X to Y when they are in thermal contact iff $t(X) > t(Y)$.

(2) trivially implies (1); in phenomenological approaches to thermodynamics the converse is often asserted to be true, but of course various additional assumptions are required to make this inference. For our purposes, though, both are corollaries of statistical mechanics, and "empirical temperatures" are just monotonically increasing functions of thermodyamical temperature.

Returning to control theory, we can now see just what transitions can and cannot be achieved via thermal contact theory. Specifically, the only transitions that can be induced are the heating and cooling of systems, and a system can be heated only if there is another system available at a higher temperature. The exact range of transitions thus achievable will depend on the size of the systems (if I have bodies at temperatures $300K$ and $400K$, I can induce *some* temperature increase in the first, but how much will depend on how quickly the second is cooled).

A useful extreme case involves *heat baths*: systems at equilibrium assumed to be so large that no amount of thermal contact with other systems will appreciably change their temperature (and which are also assumed to have no controllable parameters, not that this matters for thermal control theory). The control transitions available via thermal contact theory with heat baths are easy to state: any system can be cooled if its temperature is higher than some available heat bath, or heated if it is cooler than some such bath.

6 Thermodynamics

We are now in a position to do some non-trivial thermodynamics. In fact, we can consider two different thermodynamic theories that can thought of as two extremes. To be precise: *maximal no-feedback thermodynamics* is specified like this:

- The controlled object is a fixed, finite collection of mutually isolated statistical mechanical systems, assumed to be both thermally and parameter stable.
- The control operations are (i) arbitrary entropy-non-decreasing transition maps on the combined states of the system; (ii) leaving the systems alone for a time longer than the equilibration timescale of each system.
- There are no feedback measurements.
- The control processes are arbitrary sequences of control operations terminating in operation (ii) (that is, arbitrary sequences after which the systems are allowed to reach equilibrium).

The only constraints on this control theory are that control operations do not actually *decrease* phase-space volume, and that the control operations to apply are chosen once-and-for-all and not changed on the basis of feedback.

By contrast, here is *minimal thermodynamics*, obtained simply by conjoining thermal contact theory and adiabatic control theory:

- The controlled object is a fixed, finite collection of mutually isolated statistical mechanical systems, assumed to be both thermally and parameter stable.
- The control operations are (i) moving two systems into or out of thermal contact; (ii) making smooth changes in the parameters determining the Hamiltonians of one or more system over some finite interval of time; (iii) leaving the systems alone for a time longer than the equilibration timescale of each system.
- There are no feedback measurements.
- The control processes are arbitrary sequences of control operations terminating in operation (iii) (that is, arbitrary sequences after which the systems are allowed to reach equilibrium).

The control theory for maximal thermodynamics is straightforward. The theory induces transitions between equilibrium states; no such transition can decrease entropy; transitions are otherwise totally arbitrary. So we can induce a transition $x \rightarrow y$ between two equilibrium states x, y iff $S(x) \leq S(y)$. It is a striking feature of thermodynamics that under weak assumptions minimal thermodynamics has exactly the same control theory, so that the apparently much greater strength of maximal no-feedback thermodynamics is illusory.

To begin a demonstration, recall that in the previous sections we defined the *heat flow* into a system as the change in its energy due to thermal contact, and the *work done* on a system as the change in its energy due to modification of the parameters. By decomposing any control process into periods of arbitrarily short length — in each of which we can linearise the total energy change as the change that would have occurred due to parameter change while treating each system as isolated plus the change that would have occurred due to entropy-increasing evolution while holding the dynamics fixed — and summing the results, we can preserve these concepts in minimal thermodynamics. For any system, we then have

$$\Delta U = Q + W, \tag{18}$$

where U is the expected energy, Q is the expected heat flow into the system, and W is the expected work done on the system. This result also holds for any *collection* of systems, up to and including the entire controlled object; in the latter case, Q is zero and W can again be interpreted as the energy cost of performing the control process.

The reader will probably recognise this result as another form of the First Law of Thermodynamics. In this context, it is a fairly trivial result: its content, insofar as it has any, is just that there is a useful decomposition of energy changes by their various causes. In phenomenological treatments of thermodynamics the First Law gets physical content via some independent understanding of what "work done" is (in the axiomatic treatment of Lieb and Yngvason (1999), for instance, it is understood in terms of the potential energy of some background weight). But the real content of the First Law from that perspective is that there is a thermodynamical quantity called energy which is conserved. In our microphysical-based framework the conservation of (expected) energy is a baseline assumption and does not need to be so derived.

The concept of a quasi-static transition also generalises from adiabatic control theory to minimal thermodynamics. If $\mathrm{d}U$ is the change in system energy during an extremely small step of such a control process, we have

$$\mathrm{d}U = \sum_I \left(\frac{\partial U}{\partial V_I}\right)_{V_J,S} \mathrm{d}V_I + \left(\frac{\partial U}{\partial S}\right)_{V_I} \mathrm{d}S \tag{19}$$

and, given that quasi-static adiabatic processes are entropy-conserving, we can identify the first term as the expected work done on the system in this small step and the second as the expected heat flow into the system. Using our existing definitions we can rewrite this as

$$\mathrm{d}U = -\sum_I P^I \mathrm{d}V_I + T\mathrm{d}S, \tag{20}$$

yet another form of the First Law, but it is important to recognise that from our perspective, the expression itself has no physical content and is just a result of partial differentiation. The content comes in the identification of the first term as work and the second as heat.

Putting our results so far together, we know that

1. Any given system can be induced to make any entropy-nondecreasing transition between states.

2. Any given system's entropy may be reduced by allowing it to exchange heat with a system at a lower temperature, at the cost of increasing that system's temperature by a greater amount.

3. The total entropy of the controlled object may not decrease.

The only remaining question is then: which transitions between collections of systems that do not decrease the total entropy can be induced by a combination of (1) and (2)? So far as I know there is no *general* answer to the question. However, we can answer it fully if we assume that one of the systems is what I will call a *Carnot* system: a system such that for any value of S, $(\frac{\partial U}{\partial S})_{V_I}$ takes all positive values on the constant-S hypersurface. The operational content of this claim is that a Carnot system in any initial equilibrium state can be controlled so as to take on any temperature by an adiabatic quasi-static process.

The ideal gas is an example of a Carnot system: informally, it is clear that its temperature can be arbitrarily increased or decreased by adiabatically changing its volume. More formally, from its equation of state (8) we have

$$0 = \frac{N}{V}\mathrm{d}V + \frac{3N}{2U}\mathrm{d}U|_{\delta S=0}, \tag{21}$$

so that the energy can be changed arbitrarily through adiabatic processes, and the temperature is proportional to the energy. Of course, no gas is ideal for all temperatures and in reality the most we can hope for is a system that behaves as a Carnot system across the relevant range of temperatures.

In any case, given a Carnot system we can transfer entropy between systems with arbitrarily little net entropy increase. For given two systems at temperatures T_A, T_B with $T_A > T_B$, we can (i) adiabatically change the temperature of the Carnot system to just below T_A; (ii) place it in thermal contact with the hotter system, so that heat flows into the Carnot system with arbitrarily little net entropy increase; (iii) adiabatically lower the Carnot system to a temperature just above T_B; (iv) place it in thermal contact with the colder system, so that (if we wait the right period of time) heat flows out of the Carnot system with again arbitrarily little net entropy increase. (In the thermodynamics literature this kind of process is called a *Carnot cycle*: hence my name for Carnot systems.)

We then have a complete solution to the control problem for minimal thermodynamics: the possible transitions of the controlled object are exactly those which do not decrease the total entropy of all of the components. So “minimal” thermodynamics is, indeed not actually that minimal.

The major loophole in all this — feedback — will be discussed from section 9 onwards. Firstly, though, it will be useful to make a connection with the Second Law of Thermodynamics in its more phenomenological form.

7 The Second Law of Thermodynamics

While “the Second Law of Thermodynamics” is often read simply as synonymous with “entropy cannot decrease”, in phenomenal thermodynamics it has more directly empirical statements, each of which translates straightforwardly into our framework. Here’s the first:

The Second Law (Clausius statement): No sequence of control processes can induce heat flow Q from one system with an inverse temperature $1/T_A$, heat flow Q into a second system with a lower inverse temperature $1/T_B$, while leaving the states of all other systems unchanged.

This is a generalisation of the basic result of thermal contact theory, and the argument is essentially the same: any such process decreases the entropy of the first system by more than it increases the entropy of the second. Since the entropy of the remaining systems is unchanged (they start and end the process in the same equilibrium states), the process is overall entropy-decreasing and thus forbidden by the statistical-mechanical dynamics. If both temperatures are positive, the condition becomes the more familiar one that T_B cannot be higher than T_A.

And the second:

The Second Law (Kelvin statement): No sequence of control processes can induce heat flow Q from any one system with positive temperature while leaving the states of all other systems unchanged.

By the conservation of energy, any such process must result in net work Q being generated; an alternative way to give the Kelvin version is therefore "no process can extract heat Q from one system and turn it into work while leaving the states of all other systems unchanged". In any case, the Kelvin version is again an almost immediate consequence of the principle that Gibbs entropy is non-decreasing: since temperature is the rate of change of energy with entropy at constant parameter value, heat flow from a positive-temperature system must decrease its entropy, which (since the other systems are left unchanged) is again forbidden by the statistical-mechanical dynamics.

In both cases the "leaving the states of all other systems unchanged" clause is crucial. It is trivial to move heat from system A to system B with no net work cost if, for instance, system C, a box of gas, is allowed to expand in the process and generate enough work to pay for the work cost of the transition. Thermodynamics textbooks often use the phrase "operating in a cycle" to describe this constraint, and it will be useful to cast that notion more explicitly in our framework.

Specifically, let's define *heat bath thermodynamics* (without feedback) as follows:

- The controlled object consists of (a) a collection of heat baths at various initial temperatures; (b) another finite collection of statistical-mechanical systems, the *auxiliary object*, containing at least one Carnot system, and whose initial states are unconstrained.

- The control operations are (a) moving one or more systems in the auxiliary object into or out of thermal contact with other auxiliary-object systems and/or with one or more heat baths; (b) applying any desired smooth change to the parameters of the systems in the auxiliary object over some finite period of time; (c) inducing one or more systems in the auxiliary object to evolve in an arbitrary entropy-nondecreasing way.

- There are no feedback measurements.

- A control process is an arbitrary sequence of control operations.

In this framework, a control process is *cyclic* if it leaves the state of the auxiliary object unchanged. The Clausius and Kelvin statements are then, respectively, that no cyclic process can have as its sole effect on the heat baths (a) that net heat Q flows from one bath to one with a higher temperature at no cost in work, and (b) that net heat Q from one bath is converted into work. And again, these are fairly immediate consequences of the fact that entropy is nondecreasing.

But perhaps we don't care about *cyclic* processes? What does it matter what the actual final state of the auxiliary system is, provided the process works? We can make this intuition more precise like this: a control process delivers a given outcome *repeatably* if (i) we can perform it arbitrarily often using the final state of each process as the initial state of the next, and (ii) the Hamiltonian of the auxiliary object is the same at the end of each process as at the beginning. The Clausius statement, for instance, is now that no process can repeatably cause any quantity Q of heat to flow from one heat bath to another of higher temperature at no cost in work and with no heat flow between other heat baths.

This offers no real improvement, though. In the Clausius case, any such heat flow is entropy-decreasing on the heat baths: specifically, if they have temperatures T_A and T_B with $T_A > T_B$, a transfer of heat Q between them leads to an entropy increase of $Q/(T_A - T_B)$. So the entropy of the auxiliary object must increase by at least this much. By conservation of energy the auxiliary object's expected energy must be constant in this process. But the entropy of the auxiliary object has a maximum for given expected energy[3] and so this can be carried out only finitely many times. A similar argument can readily be given for the Kelvin statement.

I pause to note that we can turn these entirely negative constraints on heat and work into quantitative limits in a familiar way by using our existing control theory results. (Here I largely recapitulate textbook thermodynamics.) Given two heat baths having temperatures T_A, T_B with $T_A > T_B$, and a Carnot system initially at temperature T_A, the Carnot cycle to transfer heat from the colder system to the hotter is:

1. Adiabatically transition the Carnot system to the lower temperature T_B.

2. Place the Carnot system in thermal contact with the lower-temperature heat bath, and modify its parameters quasi-statically so as to cause heat to flow from the heat bath to the system. (That is, carry out modifications

[3] The canonical distribution can be characterised as the distribution which maximises Gibbs entropy for given expected energy, so this maximum is just the entropy of that canonical distribution.

which if done adiabatically would decrease the system's temperature.) Do so until heat Q_B has been transferred to the system.

3. Adiabatically transition the Carnot system to temperature T_A.

4. Place the Carnot system in thermal contact with the higher-temperature heat bath, and return its parameters quasi-statically to their initial values.

At the end of this process the Carnot system has the same temperature and parameter values as at the beginning and so will be in the same equilibrium state; the process is therefore cyclic, and the entropy and energy of the Carnot system will be unchanged. But the entropy of the system is changed only by the heat flow in steps 2 and 4. If the heat flow out of the system in step 4 is Q_A, then the entropy changes in those steps are respectively $+Q_B/T_B$ and $-Q_A/T_A$, so that $Q_A/Q_B = T_A/T_B$. By conservation of energy the net work done on the Carnot system in the cycle is $W = Q_A - Q_B$, and we have the familiar result that

$$W = \left(\frac{T_A}{T_B}\right) Q_B \tag{22}$$

for the amount of work required by a Carnot cycle-based heat pump to move a quantity of heat from a lower- to a higher-temperature heat bath.

Since the process consists entirely of quasi-static modifications of parameters (and the making and breaking of thermal contact), it can as readily be run in reverse, giving us the equally-familiar formula for the efficiency of a heat engine: T_B/T_A. And since (on pain of violating the Kelvin statement) all reversible heat engines have the same efficiency (and all irreversible ones a lower efficiency), this result is general and not restricted to Carnot cycles.

8 The one-molecule Carnot system

The Carnot systems used in our analysis so far have been assumed to be parameter-stable, thermally stable systems that can be treated via the microcanonical ensemble (and thus, in effect, to be macroscopically large). But in fact, this is an overly restrictive conception of a Carnot system, and it will be useful to relax it. All we require of such a system is that for any temperature T it possesses states which will transfer heat to and from temperature-T heat baths with arbitrarily low entropy gain, and that it can be adiabatically and quasi-statically transitioned between any two such states.

As I noted in section 5, it is a standard result in statistical mechanics that a system of any size in equilibrium with a heat bath of temperature T is described by the canonical distribution for that temperature, having probability density at energy U proportional to $\mathrm{e}^{-U/T}$. There is no guarantee that adiabatic, quasi-static transitions preserve the form of the canonical ensemble, but any system where this *is* the case will satisfy the criteria required for Carnot systems. I call such systems *canonical* Carnot systems; from here on, Carnot systems will be allowed to be either canonical or microcanonical.

To get some insight into which systems are canonical Carnot systems, assume for simplicity that there is only one parameter V and that the Hamiltonian can be written in the form required by the adiabatic theorem:

$$\widehat{H}[V] = \sum_i U_i(V) \left|\psi_i(V)\right\rangle \left\langle\psi_i(V)\right| . \tag{23}$$

Then if the system begins in canonical equilibrium, its initial state is

$$\rho(V) = \frac{1}{Z} \sum_i \mathrm{e}^{-\beta U_i(V)} \left|\psi_i(V)\right\rangle \left\langle\psi_i(V)\right| . \tag{24}$$

By the adiabatic theorem, if V is altered sufficiently slowly to V' while the system continues to evolve under Hamiltonian dynamics, it will evolve to

$$\rho(V') = \sum_i \mathrm{e}^{-\beta U_i(V)} \left|\psi_i(V')\right\rangle \left\langle\psi_i(V')\right| . \tag{25}$$

This will itself be in adiabatic form if we can find β' and Z' such that

$$\frac{\mathrm{e}^{-\beta U_i(V)}}{Z} = \frac{\mathrm{e}^{-\beta U_i(V')}}{Z'} \tag{26}$$

for which a necessary and sufficient condition is that

$$U_i(V') - U_j(V') = f(V, V')(U_i(V) - U_j(V)), \tag{27}$$

or equivalently that $U_i(V) = f(V) + g(i)h(V)$.

For an ideal gas, elementary quantum mechanics tells us that the energy of a given mode is inversely proportional to the volume of the box in which the gas is confined:[4]

$$U_i(V) = \frac{g(i)}{V}. \tag{28}$$

So an ideal gas is a canonical Carnot system. This result is independent of the number of particles in the gas and independent of any assumption that the gas spontaneously equilibrates. So in principle, even a gas with a single particle — the famous one-molecule gas introduced by Sizilard (1929) — is sufficient to function as a Carnot system. Any repeatable transfer of heat between heat baths via arbitrary entropy-non-decreasing operations on auxiliary systems can in principle be duplicated using only quasi-static operations on a one-molecule gas.[5]

For the rest of the paper, I will consider how the account developed is modified when feedback is introduced. The one-molecule gas was introduced into thermodynamics for just this purpose, and will function as a useful illustration.

[4] Quick proof sketch: increasing the size of the box by a factor K decreases the gradient by that factor, and hence decreases the kinetic energy density by a factor K^2. Energy is energy density x volume.

[5] The name "one-molecule" is a little unfortunate: the "molecule" here is monatomic and lacks internal degrees of freedom.

9 Feedback

What happens to the Gibbs entropy when a system with state ρ is measured? The classical case is easiest to analyse: suppose phase space is decomposed into disjoint regions Γ_i and that

$$\int_{\Gamma_i} \rho = p_i. \tag{29}$$

Then p_i is the probability that a measurement of which phase-space region the system lies in will give result i. The state can be rewritten in the form

$$\rho = \sum_i p_i \rho_i, \tag{30}$$

where

$$\rho_i(x) = \frac{1}{p_i}\rho(x) \tag{31}$$

if $x \in \Gamma_i$ and is zero otherwise. and by probabilistic conditionalisation, ρ_i is the state of the system after the measurement if result i is obtained.

The *expected* value of the Gibbs entropy after the measurement ('p-m') is then

$$\langle S_G \rangle_{p-m} = \sum_i S_G(\rho_i). \tag{32}$$

But we have

$$S_G(\rho) = -\int \left(\sum_i p_i \rho_i\right) \ln \left(\sum_i p_i \rho_i\right) \tag{33}$$

which, since the ρ_i are mutually disjoint, reduces to

$$S_G(\rho) = -\sum_i p_i \int \rho_i \ln(p_i \rho_i) = -\sum_i p_i \ln p_i \int \rho_i - \sum_i p_i \int \rho_i \ln \rho_i. \tag{34}$$

But the integral in the first term is just 1 (since the ρ_i are normalised) and the integral in the second term is $-S_G(\rho_i)$. So we have

$$\langle S_G \rangle_{p-m} = S_G(\rho) - \left(-\sum_i p_i \ln p_i\right). \tag{35}$$

That is, measurement may decrease entropy for two reasons. Firstly, pure chance may mean that the measurement happens to yield a post-measurement state with low Gibbs entropy. But even the *average* value of the post-measurement entropy decreases, and the level of the decrease is equal to the Shannon entropy of the probability distribution of measurement outcomes. A measurement process which has a sufficiently dramatic level of randomness could, in principle, lead to a very sharp decrease in average Gibbs entropy.

In the quantum case, the situation is slightly more complicated. We can represent the measurement by a collection of mutually orthogonal projectors $\widehat{\Pi}_i$ summing to unity, and define measurement probabilities

$$p_i = \mathsf{Tr}(\widehat{\Pi}_i \rho) \tag{36}$$

and post-measurement states

$$\rho_i = \frac{1}{p_i}\widehat{\Pi}_i \rho \widehat{\Pi}_i, \tag{37}$$

but ρ is not necessarily equal to a weighted some of these states. We can think of the measurement process, however, as consisting of two steps: a diagonalisation of ρ so that it does have this form (a non-selective measurement, or Luders projection, in foundations-of-QM jargon) followed by a random selection of the state. Mathematically the first process increases Gibbs (i. e., von Neumann) entropy, and the second mathematically has the same form as the classical analysis, so that in the quantum case (35) holds as an inequality rather than as a strict equality. (Of course, how this process of measurement is to be interpreted — and even if it can really be thought of as *measuring* anything — is a controversial question and depends on one's preferred solution to the quantum measurement problem.)

Insofar as 'the Second Law of Thermodynamics' is taken just to mean 'entropy never decreases', then, measurement is a straightforward counter-example, as has been widely recognised (see, for instance, Sizilard (1929), Zurek (1989), Albert (2000, ch.5),or Hemmo and Shenker (2012).). From the control-theory perspective, though, the interesting content of thermodynamics is which transitions it allows and which it forbids, and the interesting question about feedback measurements is whether they permit transitions which feedback-free thermodynamics does not. Here the answer is again unambiguous: it does.

To be precise: define *heat bath thermodynamics with feedback* as follows:

- The controlled object consists of (a) a collection of heat baths at various initial temperatures; (b) another finite collection of statistical-mechanical systems, the *auxiliary object*, containing at least one Carnot system, and whose initial states are unconstrained.

- The control operations are (a) moving one or more systems in the auxiliary object into or out of thermal contact with other auxiliary-object systems and/or with one or more heat baths; (b) applying any desired smooth change to the parameters of the systems in the auxiliary object over some finite period of time; (c) inducing one or more systems in the auxiliary object to evolve in an arbitrary entropy-nondecreasing way.

- Arbitrary feedback measurements may be made.

- A control process is an arbitrary sequence of control operations.

In this framework, the auxiliary object can straightforwardly be induced (with high probability) to transition from equilibrium state x to equilibrium state y with $S_G(y) < S_G(x)$. Firstly, pick a measurement such that performing it transitions x to x_i with probability p_i, such that

$$-\sum_i p_i \ln p_i \gg S_G(x) - S_G(y). \tag{38}$$

The expected value of the entropy of the post-measurement state will be much less than that of y; for an appropriate choice of measurement, with high probability the actually-obtained post-measurement state x_i will satisfy $S_G(x_i) < S_G(y)$. Now perform an entropy-increasing transformation from x_i to y. (For instance, perform a Hamiltonian transformation of x_i to some equilibrium state, then use standard methods of equilibrium thermodynamics to change that state to y).

As such, the scope of controlled transitions of the auxiliary object is total: it can be transitioned between any two states. As a corollary, the Clausius and Carnot versions of the Second Law do not apply to this control theory: energy can be arbitrarily transferred from one heat bath to another, or converted from a heat bath into work.

In fact, the full power of the arbitrary transformations available on the auxiliary system is not needed to produce these radical results. Following Szilard's classic method, let us assume that the auxiliary system is a one-molecule gas confined to a cylindrical container by a movable piston at each end, so that the Hamiltonian of the gas is parametrised by the position of the pistons. Now suppose that the position of the gas atom can be measured. If it is found to be closer to one piston than the other, the second piston can rapidly be moved at zero energy cost to the mid-point between the two. As a result, the volume of the gas has been halved without any change in its internal energy (and so its entropy has been decreased by $\ln 2$; cf equation (8).) If we now quasi-statically and adiabatically expand the gas to its original volume, its energy will decrease and so work will have been extracted from it.

Now suppose we take a heat bath at temperature T and a one-atom gas at equilibrium also at temperature T. The above process allows us to reduce the energy of the box and extract some amount of work δW from it. Placing it back in thermal contact with the heat bath will return it to its initial state and so, by conservation of energy, extracts heat $\delta Q = \delta W$ from the bath. This is a straightforward violation of the Kelvin version of the Second Law. If we use the extracted work to heat a heat bath which is hotter than the original bath, we generate a violation of the Clausius version also.

To make this explicit, let's define *Szilard theory* as follows:

- The controlled object consists of (a) a collection of heat baths at various initial temperatures; (b) a one-atom gas as defined above.
- The control operations are (a) moving the one-atom gas into or out of thermal contact with one or more heat baths; (b) applying any desired smooth change in the positions of the pistons confining the one-atom gas.
- The only possible feeback measurement is a measurement of the position of the atom in the one-atom gas.
- A control process is an arbitrary sequence of control operations.

Then the control operations available in Szilard theory include arbitrary cyclic transfers of heat between heat baths and conversion of heat into work.

The use of a *one-atom* gas in this algorithm is not essential. Suppose that we measure instead the particle density in each half of a many-atom gas at equilibrium Random fluctuations ensure that one side of the gas is at a slightly higher density than the other; compressing the gas slightly using the piston on the low-density side will reduce its volume at a slightly lower cost in work than would be possible on average without feedback; iterating such processes will again allow heat to be converted into work. (The actual numbers in play here are utterly negligible, of course — as for the one-atom gas — but we are interested here in in-principle possibility, not practicality.[6])

The most famous example of measurement-based entropy decrease, of course, is *Maxwell's demon*: a partition is placed between two boxes of gas initially at equilibrium at the same temperature. A flap, which can be opened or closed, is placed in the partition, and at short time intervals δt the boxes are measured to ascertain if, in the next period of time δt any particles will collide with the flap from (a) the left or (b) the right. If (a) holds but (b) does not, the flap is opened for the next δt seconds. Applying this alternation of feedback measurement and control operation for a sufficiently long time will reliably cause the density of the gas on the left to be much lower than on the right. Quasi-statically moving the partition to the left will then allow work to be extracted. The partition can then be removed, and reinserted in the middle; the temperature of the box will have been reduced. Placing the box in thermal contact with a heat bath will then extract heat from the bath equal to the work done; the Kelvin version of the Second Law is again violated. I will refrain from formally stating the "demonic control theory" into which these results could be embedded, but it is fairly clear that such a theory could be formulated

10 Landauer's Principle and the physical implementation of control processes

Szilard control theory, and demonic control theory, allow thermodynamically forbidden transitions. Big deal, one might reasonably think: so does abracadabra control theory, where the allowed control operations include completely arbitrary shifts in a system's state. We don't care about abracadabra control theory because we have no reason to think that it is physically possible; we only have reason to care about entropy-decreasing control theories based on measurement if we have reason to think that *they* are physically possible.

Of course, answering the general question of what is physically possible isn't easy. Is it physically possible to build mile-long relativistic starships? The answer turns on rather detailed questions of material science and the like. But no *general* physical principle forbids it. Similarly, detailed problems of implementation might make it impossible to build a scalable quantum computer, but the theory of fault-tolerant quantum computation (Shor 1996; Preskill 1998) gives

[6] For forceful defence of the idea that the practicalities are what prevents Second Law violation in these cases, see Norton (2012).

us strong reasons to think that such computers are not ruled out in principle. On the other hand, we do have reason to think that *faster-than-light* starships, or computers that can compute Turing-non-computable functions *are* in principle ruled out. It is this 'in-principle' question of implementability that is of interest here.

To answer that question, consider again heat-bath control theory. The action takes place mostly with respect to the auxiliary object: the heat baths are not manipulated in any way beyond moving into or out of contact with that object. We can then imagine treating the auxiliary object, and the control machinery, as a single larger system: we set the system going, and then simply allow it to run. It churns away, from time to time establishing or breaking physical contact with a heat bath or perhaps drawing on or topping up an external energy reservoir, and in due course completes the control process it was required to implement.

This imagined treatment of the system can be readily incorporated into our system: we can take the auxiliary object of heat-bath theory with feedback together with its controlling mechanisms, draw a box around both together, and treat the result as a single auxiliary object for a heat-bath theory *without* feedback. Put anoher way, if the feedback-based control processes we are considering are physically possible, we ought to be able to treat the machinery that makes the measurement as physical, and the machinery that decides what operation to perform based on a given feedback result as likewise physical, and treat all that physical apparatus as part of the larger auxiliary object. Let's call the assumption that this is possible the *automation constraint*; to violate it is to assume that some aspects of computation or of measurement cannot be analysed as physical processes, an assumption I will reject here without further discussion.

But we already know that heat bath theory without feedback does not permit any *repeatable* transfer of heat into work, or of a given quantity of heat from a cold body to a hotter body. Such transfers are possible, but only if the auxiliary object increases in Gibbs entropy. And gven that the auxiliary object breaks into controlling sub-object and controlled sub-object and that *ex hypothesi* the control processes we are considering leave the controlled sub-object's state unchanged, we can conclude that the Gibbs entropy of the controlling sub-object must have increased.

This raises an interesting question. From the perspective of the controlling system, control theory with feedback looks like a reasonable idealisation, but from the external perspective, we know that something must go wrong with that idealisation. The resolution of this problem lies in the effects of the measurement process on the controlling system itself: the process of iterated measurement is radically indeterministic from the perspective of the controlling object, and it can have only a finite number of relevantly distinct states, so eventually it runs out of states to use.

This point (though controversial; cf Earman and Norton (1999), Maroney (2009), and references therein) has been widely appreciated in the physics literature and can be studied from a variety of perspectives; in this rest of this section, I briefly describe the most commonly discussed one. Keep in mind in the sequel that we already know that *somehow* the controlling system's strategy must fail (at least given the automation constraint): the task is not to show *that* it does but to understand *how* it does.

The perspective we will discuss uses what might be called a *computational* model of feedback: it is most conveniently described within quantum mechanics. We assume that the controlling object consists, at least in part, of some collection of N systems - *bits* — each of whose Hilbert space is the direct sum of two *memory subspaces* 0 and 1 and each of which begins with its state somewhere in the 0 subspace. A measurement with two outcomes is then a dynamical transition which leaves the measured system alone and causes some so-far-unused bit to transition into the 1 subspace if one outcome is obtained and to remain in the 0 subspace if the other is obtained. That is, if $\widehat{T}$ is some unitary transformation of the bit's Hilbert space that maps the 0 subspace into the 1 subspace, the measurement is represented by some unitary transformation

$$\widehat{V} = \widehat{P} \otimes \widehat{T} + (\widehat{1} - \widehat{P}) \otimes \widehat{1} \tag{39}$$

on the joint system of controlled object and bit (with $(\widehat{P}, 1 - \widehat{P})$ being the projectors defining the measurement. A feedback-based control processes based on the result of this measurement is then represented by a unitary transformation of the form

$$\widehat{U} = \widehat{U}_0 \otimes \widehat{P}_0 + \widehat{U}_1 \otimes \widehat{P}_1 \tag{40}$$

where $\widehat{P}_0$, $\widehat{P}_1$ project onto the 0 and 1 subspaces and $\widehat{U}_0$ and $\widehat{U}_1$ are unitary operations on the controlled system. The combined process of $\widehat{V}$ followed by $\widehat{U}$ represents the process of measuring the controlled object and then performing $\widehat{U}_0$ on it if one result is obtained and $\widehat{U}_1$ if the other is. Measurements with 2^N outcomes, and control operations based on the results of such measurements, can likewise be represented through the use of N bits. The classical case is essentially identical (but the formalism of quantum theory makes the description simpler in the quantum case).

The problem with this process is that eventually, the system runs out of unused bits. (Note that the procedure described above only works if the bit is guaranteed to be in the 0 subspace initially. To operate repeatably, the system will then have to reset some bits to the initial state. But *Landauer's Principle* states that such resetting carries an entropy cost. Since the principle is controversial (at least in the philosophy literature!) I will work through the details here from a control-theory perspective.

Specifically, let's define a *computational process* as follows: it consists of N bits (the *memory*) together with a finite system (the *computer*) and another system (the *environment*). A *computation* is a transition which is deterministic at the level of bits: that is, if the N bits begin, collectively, in subspaces that encode the binary form of some natural number n, after the transition they are found, collectively, in subspaces encoding $f(n)$ for some fixed function f.[7] The

[7] Maroney (2005) is a highly insightful discussion which *inter alia* considers the case of indeterministic computation.

control processes are arbitrary unitary (quantum) or Hamiltonian (classical) evolutions on the combined system of memory, computer, and environment; the question of interest is what constraints on the transitions of computer and environment are required for given computational transitions to be implemented. For the sake of continuity with the literature I work in the classical framework (the quantum generalisation is straightforward); for simplicity I assume that the bits have equal phase space V assigned to 0 and 1.

If the function f is one-to-one, the solution to the problem is straightforward. The combined phase space of the memory can be partitioned into 2^N subspaces each of equal volume and each labelled with the natural number they represent. There is then a phase-space-preserving map from n to $f(n)$ for each n, and these maps can be combined into a single map from the memory to itself. One-to-one ('reversible') computations can then be carried out without any implications for the states of computer or environment.

But now suppose that the function f takes values only between 1 and 2^M $(M < N)$, so that any map implementing f must map the bits $M+1, \ldots N$ into their zero subspaces independent of input. Any such map would map the uniform distribution over the memory (which has entropy $N \ln 2V$) to one with support in a region of volume $(2V)^M \times V^{N-M}$ (and so with maximum entropy $M \ln 2V + (N-M) \ln V$). Since the map as a whole is by assumption entropy-preserving, it must increase the joint entropy of system plus environment by $(N-M) \ln 2$. In the limiting case of reset, $M = 0$ ($f(n) = 0$ for all n) and so the computer and environment must jointly increase in entropy by at least $N \ln 2$. This is Landauer's principle: each bit that is reset generates at least $\ln 2$ entropy.

If the computer is to carry out the reset operation repeatably, its own entropy cannot increase without limit. So a *repeatable* reset process dumps at least entropy $\ln 2$ per bit into the environment. In the special case where the environment is a heat bath at temperature T, Landauer's principle becomes the requirement that reset generates $T \ln 2$ heat per bit.

A more realistic feedback-based control theory, then, might incorporate Landauer's Principle explicitly, as in the following (call it *computation heat-bath thermodynamics*:

- The controlled object consists of (a) a collection of heat baths at various initial temperatures; (b) another finite collection of statistical-mechanical systems, the *auxiliary object*, containing at least one Carnot system, and whose initial states are unconstrained; (c) a finite number N of 2-state systems ('bits'), the *computational memory*, each of which begins in some fixed ('zero') initial state with probability 1.
- The control operations are (a) moving one or more systems in the auxiliary object into or out of thermal contact with other auxiliary-object systems and/or with one or more heat baths ; (b) applying any desired smooth change to the parameters of the systems in the auxiliary object over some finite period of time; (c) inducing one or more systems in the auxiliary object to evolve in an arbitrary entropy-nondecreasing way; (d) erasing

M bits of the memory — that is, restoring them to their zero states — and at the same time transferring heat $M \ln 2/T$ to some heat bath at temperature T; (e) applying any $1-1$ map to the computational memory.

- Arbitrary feedback measurements may be made (including of the memory bits) provided that (a) they have finitely many results; (b) the result of the measurement is faithfully recorded in the state of some collection of bits which initially each have probability 1 of being in the 0 state.

- A control process is an arbitrary sequence of control operations.

At first sight, measurement in this framework is in the long run entropy-*increasing*: a measurement with 2^M outcomes having probabilities $p_1, \ldots p_{2^M}$ will reduce the entropy by $\Delta S = -\sum_i p_i \ln p_i$, but the maximum value of this is $M \ln 2$, which is the entropy increase required to erase the M bits required to record the result. But as Zurek (1989) has pointed out, Shannon's noiseless coding theorem allows us to compress those M bits to, on average, $-\sum_i p_i \ln p_i$ bits, so that the overall process can be made entropy-neutral.

This strategy of using Landauer's principle to explain why Maxwell demons cannot repeatably violate the Second Law has a long history (see Leff and Rex (2002) and references therein). It has recently come under sharp criticism by John Earman and John Norton (Earman and Norton 1999; Norton 2005) as either trivial or question-begging: they argue that any such defences ('exorcisms') rely on arguments for Landauer's Principle that are either Sound (that is, start off by assuming the Second Law), or Profound (that is, do not so start off). Exorcisms relying on Sound arguments are question-begging; those relying on Profound exorcisms leave us no good reason to accept Landauer's principle in the first place.

Responses to Earman and Norton (see, e. g., Bennett (2003), Ladyman, Presnell, and Short (2008)) have generally embraced the first horn of the dilemma, accepting that Landauer's Principle does assume the Second Law but arguing that use of it can still be pedagogically illuminating. (See Norton (2005, 2011) for responses to this move.) But I believe the dialectic here fails to distinguish between statistical mechanics and thermodynamics. The argument here for Landauer's Principle does indeed assume that the underlying dynamics are entropy-non-decreasing, and from that perspective appeal to Landauer's principle is merely of pedagogical value: it helps us to make sense of how feedback processes can be entropy-decreasing despite the fact that any black-box process, even if it involves internal measurement of subsystems, cannot repeatedly turn heat into work. But (this is one central message of this paper) that dynamical assumption within statistical mechanics should not simply be identified with the phenomenological Second Law. In Earman and Norton's terminology, the argument for Landauer's Principle is Sound with respect to statistical mechanics, but Profound with respect to phenomenological thermodynamics.

11 Conclusion

The results of my exploration of control theory can be summarised as follows:

1. In the absence of feedback, physically possible control process are limited to inducing transitions that do not lower Gibbs entropy.

2. That limit can be reached with access to very minimal control resources: specifically, a single Carnot system and the ability to adiabatically control it and to put it in thermal contact with other systems.

3. Introducing feedback allows arbitrary transitions.

4. If we try to model the feedback process as an internal dynamical process in a larger system we find that feedback does not increase the power of the control process.

5. (3) and (4) can be reconciled by considering the physical changes to the controlling system during feedback processes. In particular, on a computation model of control and feedback, the entropy cost of resetting the memory used to record the result of measurement at least cancels out the entropy reduction induced by the measurement.

I will end with a more general moral. As a rule, and partly for pedagogical reasons, foundational discussions of thermal physics tend to begin with thermodynamics and continue to statistical mechanics. The task of recovering thermodynamics from a successfully grounded statistical mechanics is generally not cleanly separated from the task of understanding statistical mechanics itself, and the distinctive requirements of thermodynamics blur into the general problem of understanding statistical-mechanical irreversibility. Conversely, foundational work on thermodynamics proper is often focussed on thermodynamics understood phenomenologically: a well-motivated and worthwhile pursuit, but not one that obviates the need to understand thermodynamics from a statistical-mechanical perspective.

The advantage of the control-theory way of seeing thermodynamics is that it permits a clean separation between the foundational problems of statistical mechanics itself and the reduction problem of grounding thermodynamics in statistical mechanics. I hope to have demonstrated that (a) these really are distinct problems, so that an understanding of (e.g.) why systems spontaneously aproach equilibrium does not in itself suffice to give an understanding of thermodynamics, but also (b) that such an understanding, via the interpretation of thermodynamics as the control theory of statistical mechanics, can indeed be obtained, and can shed light on a number of extant problems at the statistical-mechanics/thermodynamics boundary.

Acknowledgements

In writing this paper I benefitted greatly from conversations with David Albert, Harvey Brown, Wayne Myrvold, John Norton, Jos Uffink, and in particular Owen Maroney.

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.

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.

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

Ladyman, J., S. Presnell, and A. Short (2008). The use of the information-theoretic entropy in thermodynamics. *Studies in History and Philosophy of Modern Physics 39*(2), 315–324.

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

Lieb, E. H. and J. Yngvason (1999). The physics and mathematics of the second law of thermodynamics. *Physics Reports 310*, 1–96.

Maroney, O. (2009). Information processing and thermodynamic entropy. In the Stanford Encyclopedia of Philosophy (Fall 2009 edition), Edward N. Zalta (ed.), available online at http://plato.stanford.edu/archives/fall2009/entries/information-entropy/ .

Maroney, O. J. E. (2005). The (absence of a) relationship between thermodynamic and logical reversibility. *Studies in the History and Philosophy of Modern Physics 36*, 355–374.

Messiah, A. (1962). *Quantum Mechanics*, Volume II. North.

Norton, J. (2011). Waiting for Landauer. Available online at http://philsci-archive.pitt.edu/8635/.

Norton, J. (2012). The end of the thermodynamics of computation: a no go result. To appear in the Philosophy of Science Association 23rd Biennial Meeting Collected Papers; available online at http://philsci-archive.pitt.edu/9658/.

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.

Preskill, J. (1998). Fault-tolerant quantum computation. *Introduction to quantum computation and information 213.*

Shor, P. W. (1996). Fault-tolerant quantum computation. In *Foundations of Computer Science, 1996. Proceedings., 37th Annual Symposium on*, pp. 56–65. IEEE.

Sizilard, L. (1929). ber die entropieverminderung in einem thermodynamischen system bei eingriffen intelligenter wesen. *Zeitschrift fur Physik 53*, 840–856. English translation, "On the Decrease of Entropy in a Thermodynamic System by the Intervention of Intelligent Beings", available in B.T. Feld and G. Weiss Szilard (eds.), *The Collected Works of Leo Szilard: Scientific Papers* (Cambridge, Massachusetts: MIT Press, 1972), pp. 103-129.

Wallace, D. (2013a). The non-problem of Gibbs vs. Boltzmann entropy. Forthcoming.

Wallace, D. (2013b). What statistical mechanics actually does. Forthcoming.

Weinberg, S. (2013). *Quantum Mechanics.* Cambridge: Cambridge University Press.

Zurek, W. H. (1989). Algorithmic randomness and physical entropy. *Physical Review A 40*, 4731–4751.