ELSEVIER

Studies in History and Philosophy of Modern Physics 36 (2005) 375–411

Studies in History and Philosophy of Modern Physics

www.elsevier.com/locate/shpsb

Eaters of the lotus: Landauer's principle and the return of Maxwell's demon

John D. Norton

Department of History and Philosophy of Science, University of Pittsburgh, Pittsburgh, PA 15260, USA

Received 11 June 2004; received in revised form 13 November 2004; accepted 8 December 2004

Abstract

Landauer's principle is the loosely formulated notion that the erasure of n bits of information must always incur a cost of $k \ln n$ in thermodynamic entropy. It can be formulated as a precise result in statistical mechanics, but for a restricted class of erasure processes that use a thermodynamically irreversible phase space expansion, which is the real origin of the law's entropy cost and whose necessity has not been demonstrated. General arguments that purport to establish the unconditional validity of the law (erasure maps many physical states to one; erasure compresses the phase space) fail. They turn out to depend on the illicit formation of a canonical ensemble from memory devices holding random data. To exorcise Maxwell's demon one must show that all candidate devices—the ordinary and the extraordinary—must fail to reverse the second law of thermodynamics. The theorizing surrounding Landauer's principle is too fragile and too tied to a few specific examples to support such general exorcism. Charles Bennett's recent extension of Landauer's principle to the merging of computational paths fails for the same reasons as trouble the original principle.

Keywords: Entropy; Information; Landauer's principle; Maxwell's demon

1. Introduction

A sizeable literature is based on the claim that Maxwell's demon must fail to produce violations of the second law of thermodynamics because of an inevitable

E-mail address: jdnorton@pitt.edu (J.D. Norton).

1355-2198/$ - see front matter
doi:10.1016/j.shpsb.2004.12.002

entropy cost associated with certain types of information processing. In the second edition of their standard compilation of work on Maxwell's demon, Leff and Rex (2003, p. xii) note that more references have been generated in the 13 years since the volume's first edition than in all years prior to it, extending back over the demon's 120 years of life. A casual review of the literature gives the impression that the demonstrations of the failure of Maxwell's demon depend on the discovery of independent principles concerning the entropy cost of information processing. It looks like a nice example of new discoveries explaining old anomalies. Yet closer inspection suggests that something is seriously amiss. There seems to be no independent basis for the new principles. In typical analyses, it is *assumed* at the outset that the total system has canonical thermal properties so that the second law will be preserved; and the analysis then infers back from that assumption to the entropy costs that it assumes must arise in information processing. In our Earman and Norton (1998, 1999),[1] my colleague John Earman and I encapsulated this concern in a dilemma posed for all proponents of information theoretic exorcisms of Maxwell's demon. Either the combined object system and demon are assumed to form a canonical thermal system or they are not. If not ("profound" horn), then we ask proponents of information theoretic exorcisms to supply the new physical principle needed to assure failure of the demon and give independent grounds for it. Otherwise ("sound" horn), it is clear that the demon will fail; but it will fail only because its failure has been assumed at the outset. Then the exorcism merely argues to a foregone conclusion.

Charles Bennett has been one of the most influential proponents of information theoretic exorcisms of Maxwell's demon. The version he supports seems now to be standard. It urges that a Maxwell demon must at some point in its operation erase information. It then invokes Landauer's principle, which attributes an entropy cost of at least $k \ln n$ to the erasure of n bits of information in a memory device, to supply the missing entropy needed to save the second law (k is Boltzmann's constant). We are grateful for Bennett's (2003, pp. 501, 508–510) candor in responding directly to our dilemma and accepting its sound horn.[2] He acknowledges that his use of Landauer's principle is "in a sense…indeed a straightforward consequence or restatement of the Second Law, [but] it still has considerable pedagogic and explanatory power…" While some hidden entropy cost can be inferred from the presumed correctness of the second law, its location remains open. The power of Landauer's principle, Bennett asserts, resides in locating this cost properly in information erasure and so correcting an earlier literature that mislocated it in information acquisition.

My concern in this paper is to look more closely at Landauer's principle and how it is used to exorcise Maxwell's demon. My conclusion will be that this literature

[1] For another critical approach to this literature, see Shenker (1999, 2000).

[2] In responding to the dilemma, Leff and Rex (2003, p. 34) appear to accept the profound horn. They point out derivations of Landauer's principle that do not explicitly invoke the second law of thermodynamics. These derivations still fall squarely within the sound horn since they all assume that the systems examined exhibit canonical thermal behavior entirely compatible with the second law and, at times, strong enough to entail it.

overreaches. Its basic principles are vaguely formulated; and its justifications are rudimentary and sometimes dependent on misapplications of statistical mechanics. It is a foundation too weak and fragile to support a result as general as the impossibility of Maxwell's demon. I will seek to establish claims belonging to three groups:

A precise formulation of Landauer's principle for the cases in which it applies: In the current literature, the principle appears as a vague slogan associating an entropy cost of $k \ln n$ with the erasure of n bits of information; or as the demonstration that, in some quite specific example systems, this entropy cost does obtain. I present a version of the principle sufficiently general to include the standard examples, and also sufficiently precise for it to be proved as a result in statistical mechanics. It turns out that the result depends essentially on the erasure procedure employing a thermodynamically irreversible expansion of the memory device's phase space. This irreversible step is the real origin of the erasure's entropy cost. Since there is no demonstration that an erasure procedure must employ this thermodynamically irreversible step, this formulation of Landauer's principle does not sustain the general slogan that all erasure is associated with a thermodynamic entropy cost. The presence of the thermodynamically irreversible, entropy creating step has been obscured in the literature by a mischaracterization of reversible processes. The term is applied incorrectly to uncontrolled expansions. In standard thermodynamics, reversible processes are taken as co-extensive with quasi-static processes; that is processes each of whose stages are in thermal equilibrium or removed from it to an arbitrarily small degree. Because of this mischaracterization, thermodynamic entropy as traditionally defined by Clausius, ceases to be a state function.

The misuse of canonical distributions and ensembles: In statistical mechanics, a system in thermal equilibrium is represented by a canonical probability distribution over the microstates in the portions of the phase space that are accessible to the system under its time evolution. These systems possess a thermodynamic entropy that varies with the logarithm of the accessible phase space volume. In the Landauer's principle literature, a memory device recording random data is treated illicitly as if it were a canonically distributed system. This treatment is illicit since the different states possible for the memory device are not the states accessible to the device under its time development. (If they were, the device could not function as a memory device.) So we cannot associate a thermodynamic entropy with the logarithm of the phase volume of possible states using the standard formulae of statistical mechanics; the assumptions needed to deduce these formulae do not obtain. The outcome is that much of the ensuing analysis of the thermodynamics of memory devices is erroneous, including the general argument offered in support of Landauer's principle. It argues that erasure maps many states to one; that is, it compresses the phase space. This compression is then spuriously associated with a change in thermodynamic entropy. A more careful analysis shows that, with natural symmetry assumptions,

memory devices recording random data have the same thermodynamic entropy as the reset devices. Putting the error in its simplest terms, the probability distribution of random data is not the sort of probability distribution that can be associated with thermodynamic entropy; or, more technically, a collection of memory devices recording random data cannot be treated as if it were a canonical ensemble, a standard way of representing a canonically distributed system in statistical mechanics.

Failure of exorcisms of Maxwell's demon: The challenge of exorcising Maxwell's demon is to show that no device, no matter how extraordinary or how ingeniously or intricately contrived, can find a way of accumulating fluctuation phenomena into a macroscopic violation of the second law of thermodynamics. The existing analyses of Landauer's principle are too weak to support such a strong result. The claims to the contrary depend on displaying a few suggestive examples in which the failure of a Maxwell demon of some particular design is deduced from Landauer's principle. We are then to expect, on the basis of an argument from analogy, that every other possible design of a Maxwell demon must fare likewise. I argue that there is no foundation for this expectation because of the fragility of the analogy. There are many ways in which the design of extraordinary Maxwell's demons might differ from the ordinary examples. For example, we can readily find candidate demons that cannot be said to compute or store information or erase a memory, so that Landauer's principle cannot be applied.

In the sections to follow, the precise but restricted version of Landauer's principle is developed and stated in Section 2, along with some thermodynamic and statistical mechanical preliminaries, introduced for later reference. Section 3 identifies how canonical ensembles are illicitly assembled in the Landauer's principle literature and shows how this illicit assembly leads to the failure of the many-to-one mapping argument. Section 4 reviews the challenge presented by Maxwell's demon and argues that the present literature on Landauer's principle is too fragile to support its exorcism. Section 5 reviews Bennett's extension of Landauer's principle to the merging of computational flows. I argue that the extension fails and thereby also fails in its goal of exorcising the no-erasure demon introduced by John Earman and me in our Earman and Norton (1999, pp. 16–17).

2. The physics of Landauer's principle

2.1. Which sense of entropy?

There are several senses for the term entropy. We can affirm quite rapidly that *thermodynamic* entropy is the sense relevant to the literature on Maxwell's demon and Landauer's principle. By thermodynamic entropy, I mean the quantity S that is a function of the state of a thermal system in equilibrium at temperature T and is

defined by the classical Clausius formula

$$\delta S = \frac{\delta Q_{\text{rev}}}{T}, \tag{1}$$

δS represents the rate of gain of entropy during a thermodynamically reversible process by a system at temperature T that gains heat at the rate of δQ_{rev}. A thermodynamically reversible process is one that can proceed in either forward or reverse direction because all its components are at equilibrium or removed from it to an arbitrarily small degree. Clausius formula (1) defines entropy changes during a particular process. If these changes are to be associated with a thermodynamic entropy S that is a property of thermal states, these changes must be path independent; that is, for any closed path in system's state space we must have

$$\oint \frac{\mathrm{d}Q_{\text{rev}}}{T} = 0. \tag{1'}$$

To see that this is the appropriate sense of entropy, first note the effect intended by Maxwell's original demon (Leff & Rex, 2003, p. 4). It was to open and close a hole in a wall separating two compartments containing a kinetic gas so that faster molecules accumulate on one side and the slower on the other. One side would become hotter and the other colder without expenditure of work. That would directly contradict the "Clausius" form of the second law as given by Thomson in its original form:

> It is impossible for a self-acting machine, unaided by any external agency, to convey heat from one body to another at a higher temperature. (Thomson, 1853, p. 14)

A slight modification of Maxwell's original scheme is the addition of a heat engine that would convey heat from the hotter side back to the colder, while converting a portion of it into work. The whole device could be operated so that the net effect would be that heat, drawn from the colder side, is fully converted into work, while further cooling the colder side. This would be a violation of the "Thomson" form of the second law of thermodynamics as given by Thomson:

> It is impossible, by means of inanimate material agency, to derive mechanical effect from any portion of matter by cooling it below the temperature of the coldest of the surrounding objects. (Thomson, 1853, p. 13)

Another standard implementation of Maxwell's demon is the Szilard one-molecule gas engine, described more fully in Section 4.2. Its net effect is intended to be the complete conversion of a quantity of heat extracted from the thermal surroundings into work.

One of the most fundamental results of thermodynamic analysis is that these two versions of the second law of thermodynamics are equivalent and can be re-expressed as the existence of the state property, thermodynamic entropy, defined by (1) that obeys:

> Every physical or chemical process in nature takes place in such a way as to increase the sum of the entropies of all the bodies taking part in the process. In the limit, i.e. for reversible processes, the sum of the entropies remains unchanged.

> This is the most general statement of the second law of thermodynamics. (Planck, 1926, p. 103)

One readily verifies that a Maxwell demon, operating as intended, would reduce the total thermodynamic entropy of a closed system, in violation of this form of the second law. Thus, the burden of an exorcism of Maxwell's demon is to show that there is a hidden increase in thermodynamic entropy associated with the operation of the demon that will protect the second law.

The present orthodoxy is that Landauer's principle successfully locates this hidden increase in the process of memory erasure. According to the principle, erasure of one bit reduces the entropy of the memory device by $k \ln 2$. That entropy is clearly intended to be thermodynamic entropy. It is routinely assumed that a reduction in entropy of the memory device must be accompanied by at least as large an increase in the entropy of its environment. That in turn requires the assumption that the relevant sense of entropy is governed by a law like the second law of thermodynamics that prohibits a net reduction in the entropy of the total system. More directly, Landauer's principle is now often asserted not in terms of entropy but in terms of heat: erasure of one bit of information in a memory device must be accompanied by the passing of at least $kT \ln 2$ of heat to the thermal environment at temperature T.[3] This form of Landauer's principle entails that entropy of erasure is thermodynamic entropy. If the process passes $kT \ln 2$ of heat to the environment in the least dissipative manner, then the heating must be a thermodynamically reversible process. That is, the device must also be at temperature T during the time in which the heat is passed and it must lose $kT \ln 2$ of energy as heat. It now follows from definition (1) that the thermodynamic entropy of the memory device has decreased by $k \ln 2$.

2.2. Canonical distributions and thermodynamic entropy

The memory devices Landauer (1961) and the later literature describe are systems in thermal equilibrium with a much larger thermal environment (at least at essential moments in their history); and the relevant sense of entropy is thermodynamic entropy. Statistical mechanics represents systems in thermal equilibrium with a much larger thermal environment at temperature T by canonical probability distributions over the systems' phase spaces. If a system's possible states form a phase space Γ with canonical position and momentum coordinates $x_1, \ldots, x_n$ (henceforth abbreviated "x"), then the canonical probability distribution for the system is the

[3] Landauer's (1961, p. 152) early statement of the principle immediately relates the entropy of erasure to a heating effect: "[In erasing one bit, t]he entropy therefore has been reduced by $k \log_e 2 = 0.6931$ k per bit. The entropy of a closed system, e.g. a computer with its own batteries, cannot decrease; hence this entropy must appear elsewhere as a heating effect, supplying 0.6931 kT per restored bit to the surroundings. This is of course a minimum heating effect…" Shizume (1995, p. 164) renders the principle as "Landauer argued that the erasure of 1 bit of information stored in a memory device requires a minimal heat generation of $k_B T \ln 2$…"; and Piechocinska (2000, p. 169) asserts: "Landauer's principle states that in erasing one bit of information, on average, at least $k_B T \ln(2)$ energy is dissipated into the environment…"

probability density

$$p(x) = \exp(-E(x)/kT)/Z, \tag{2}$$

where $E(x)$ is the energy of the system at x in its phase space and k is Boltzmann's constant. The regions (of non-zero natural measure) of the phase space everywhere accessible to the system over time must have non-zero probability and so must lie within regions of finite energy $E(x)$. Regions with infinite energy $E(x)$ have zero probability and are inaccessible. If the system visits accessible regions densely over time and we require that the probability of a region coincide with the portion of time spent there, then the accessible regions will coincide with those of finite energy $E(x)$. The partition function is

$$Z = \int_{\Gamma} \exp(-E(x)/kT)\,\mathrm{d}x. \tag{3}$$

A standard calculation (e.g. Thomson, 1972, Section 3.4) allows us to identify which quantity corresponds to the thermodynamic entropy. If such a function exists at all, it must satisfy (1) during a thermodynamically reversible transformation of the system. The reversible process sufficient to fix this function is:

S: *Specification of a thermodynamically reversible process in which the system remains in thermal equilibrium with an environment at temperature T.*

S1: The temperature T of the system and environment may slowly change, so that T should be written as function $T(t)$ of the parameter t that measures degree of completion of the process. To preserve thermodynamic reversibility, the changes must be so slow that the system remains canonically distributed as in (2).

S2: Work may also be performed on the system. To preserve thermodynamic reversibility, the work must be performed so slowly so that the system remains canonically distributed. The work is performed by direct alteration of the energy $E(x)$ of the system at phase space x, so that this energy is now properly represented by $E(x, \lambda)$, where the manipulation variable $\lambda(t)$ is a function of the completion parameter t.

As an illustration of how work is performed on the system according to S2, consider a particle of mass m and velocity v confined in a well of a potential field φ in a one-dimensional space. The energy at each point x in the phase space is given by the familiar $E(\pi, x) = \pi^2/2m + \varphi(x)$, where π is the canonical momentum mv and x the position coordinate. The gas formed by the single molecule can be compressed reversibly by a very slow change in the potential field that restricts the volume of phase space accessible to the particle, as shown in Fig. 1. Another very slow change in the potential field also illustrated in Fig. 1 may merely have the effect of relocating the accessible region of phase space without expending any net work or altering the accessible volume of phase space.

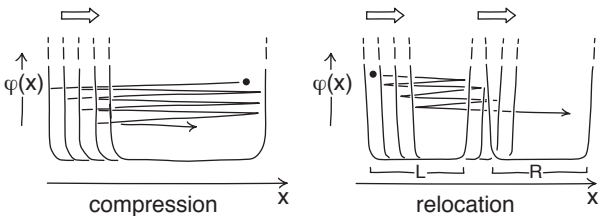

Fig. 1. Thermodynamically reversible processes due to slow change in potential field.

The mean energy of the system at any stage of such a process is

$$\bar{E} = \int_{\Gamma} E(x,\lambda)p(x,t)\,\mathrm{d}x. \tag{4}$$

So the rate of change of the mean energy is

$$\frac{\mathrm{d}\bar{E}}{\mathrm{d}t} = \int_{\Gamma} E(x,\lambda)\,\frac{\mathrm{d}p(x,t)}{\mathrm{d}t}\,\mathrm{d}x + \int_{\Gamma} \frac{\mathrm{d}E(x,\lambda)}{\mathrm{d}t}\,p(x,t)\,\mathrm{d}x. \tag{5}$$

The second term in the sum is the rate at which work W is performed on the system

$$\frac{\mathrm{d}W}{\mathrm{d}t} = \int_{\Gamma} \frac{\mathrm{d}E(x,\lambda)}{\mathrm{d}t} p(x,t)\,\mathrm{d}x. \tag{6}$$

This follows since the rate at which work is performed on the system, if it is at phase point x, is

$$\frac{\partial E(x,\lambda)}{\partial\lambda}\,\frac{\mathrm{d}\lambda}{\mathrm{d}t} = \frac{\mathrm{d}E(x,\lambda)}{\mathrm{d}t}.$$

The rate at which work is performed is just the phase average of this quantity, which is the second term in the sum (5). It is essential for this conclusion that the system explore the phase space accessible to it much more rapidly than the rate at which the process alters the energy function $E(x,\lambda)$. For example, in the thermodynamically reversible compression depicted in Fig. 1, the particle must bounce many times between the walls of the potential well, while the walls move only very slightly. Otherwise (6) will fail to express the rate at which work is performed on the system.

The first law of thermodynamics assures us that

energy change = heat gained + work performed on system.

So, by subtraction, we identify the rate at which heat is gained by the system as

$$\frac{\mathrm{d}Q_{\mathrm{rev}}}{\mathrm{d}t} = \int_{\Gamma} E(x,\lambda)\,\frac{\mathrm{d}p(x,t)}{\mathrm{d}t}\,\mathrm{d}x. \tag{7}$$

Combining this formula with Clausius expression (1) for entropy and expression (2) for a canonical distribution, we recover after some manipulation that

$$\frac{\mathrm{d}S}{\mathrm{d}t} = \frac{1}{T}\,\frac{\mathrm{d}Q_{\mathrm{rev}}}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\left(\frac{\bar{E}}{T} + k\,\ln\int_{\Gamma}\exp(-E/kT)\,\mathrm{d}x\right)$$

so that the thermodynamic entropy of a canonically distributed system is just

$$S = \frac{\bar{E}}{T} + k \ln \int_\Gamma \exp(-E/kT)\,\mathrm{d}x \tag{8}$$

up to an additive constant—or, more cautiously, if any quantity can represent the thermodynamic entropy of a canonically distributed system, it is this one.

This expression for thermodynamic entropy should be compared with another more general expression

$$S = -k \int_\Gamma p(x) \ln p(x)\,\mathrm{d}x \tag{9}$$

that assigns an entropy to *any* probability distribution over a space Γ.[4] If I am as sure as not that an errant asteroid brought the demise of the dinosaurs, then I might assign probability $\frac{1}{2}$ to the hypothesis it did; and probability $\frac{1}{2}$ to the hypothesis it did not. Expression (9) would assign entropy $k \ln 2$ to the resulting probability distribution. If I subsequently become convinced of one of the hypotheses and assign unit probability to it, the entropy assigned to the probability distribution drops to zero. In general, the entropy of (9) has nothing to do with thermodynamic entropy; it is just a property of a probability distribution. The connection arises in a special case. If the probability distribution $p(x)$ is the canonical distribution (2), then, upon substitution, expression (9) reduces to the expression for thermodynamic entropy (8) for a system in thermal equilibrium at temperature T.

2.3. Landauer's principle for the erasure of one bit

What *precisely* does Landauer's principle assert? And why *precisely* should we believe it? These questions prove difficult to answer. Standard sources in the literature express Landauer's principle by example, noting that this or that memory device would incur an entropy cost were it to undergo erasure. The familiar slogan is (e.g. Leff & Rex, 2003, p. 27) that "erasure of one bit of information increases the entropy of the environment by at least $k \ln 2$". One does not so much learn the general principle, as one gets the hang of the sorts of cases to which it can be applied. Landauer's (1961) original article gave several such illustrations. A helpful and revealing one is (p. 152):

> Consider a statistical ensemble of bits in thermal equilibrium. If these are all reset to ONE, the number of states covered in the ensemble has been cut in half. The entropy therefore has been reduced by $k \log_e 2 = 0.6931\,\mathrm{k}$ per bit.

This remark also captures the central assertion of justifications given for the principle. The erasure operation reduces the number of states, or it effects a "many-to-one mapping" (Bennett, 1982, p. 305) or a "compression of the occupied volume of the [device's] phase space" (p. 307).

[4] The constant k and the use of natural logarithms amount to a conventional choice of units that allows compatibility with the corresponding thermodynamic formula.

Matters have improved somewhat with what Leff and Rex (2003, p. 28) describe as new "proofs" of Landauer's principle in Shizume (1995) and Piechocinska (2000). However, neither gives a general statement of the principle beyond the above slogan, thereby precluding the possibility of a real proof of a general principle. Instead they give careful and detailed analysis of the entropy cost of erasure in several more examples, once again leaving us to wonder which of the particular properties assumed for the memory devices and procedures are essential to the elusive general principle.

If Landauer's principle is to supply the basis for a general claim of the failure of all Maxwell's demons, we must have a general statement of the principle and of the grounds that support it. We must know what properties of the memory devices are essential and which incidental; what range of erasure procedures are covered by the principle; which physical laws are needed for the demonstration of the principle; and a demonstration that those laws do entail the principle. While trying to avoid spurious precision and overgeneralization,[5] my best effort to meet these demands follows. The resulting principle is not universally applicable since it requires the use of an erasure procedure with a thermodynamically irreversible step, while the necessity of such a step has not been demonstrated. It is specialized in less important ways. It is limited to the case of erasure of one bit and to the setting of classical physics. The extension to the erasure of n bits is obvious. The extension to quantum systems appears not to involve any matters of principle, as long as quantum entanglement is avoided; rather it is mostly the notational nuisance of replacing integrations by summations.

Landauer's principle for erasure of one bit of information in a memory device:

IF

LP1. The memory device and erasure operation are governed by the physics of statistical mechanics and thermodynamics as outlined in Section 2.2.[6]

LP2. The memory device has a phase space Γ on which energy functions $E(x)$ are defined and, at least at certain times in its operation as indicated below, the system is in thermal equilibrium with a larger environment at T, so it is canonically distributed over the accessible portions of phase space according to (2).

LP3. The phase space contains two disjoint regions "L" and "R", with their union designated "$L+R$". There are different energy functions $E(x)$ available. $E_L(x)$ confines the system to L; $E_R(x)$ to R; and $E_{L+R}(x)$ to $L+R$. If the device is in thermal equilibrium at T and confined to L, R or $L+R$, we shall say it is in state L_T, R_T or $(L+R)_T$. When the device's state is confined to L, it registers a value L; when confined to R, it registers R.

[5] For example, one could weaken the symmetry requirement and try to recover entropy generation of $k \ln 2$ on average, per erasure. That would greatly complicate the analysis for little useful gain.

[6] Through this condition, the demonstration of Landauer's principle presumes that the systems are canonically thermal and thus it falls within the "sound" horn of the dilemma of Earman and Norton (1998, 1999).

LP4. The energy function's two regions L and R are perfectly symmetric in the sense that there is a one-one map of canonical coordinates $x_L \in L \to x_R \in R$ between regions L and R that assures they have equal phase space volume and such that $E_L(x_L) = E_R(x_R)$; and $E_{L+R}(x_L) = E_{L+R}(x_R)$; and $E_{L+R}(x_L) = E_L(x_L)$.

LP5. The erasure process has two steps.

LP5a. ("removal of the partition") The device in state L_T or R_T proceeds, through a thermodynamically *irreversible*, adiabatic expansion, to the state $(L+R)_T$.

LP5b. ("compression of the phase space") The device in state $(L+R)_T$ proceeds through a thermodynamically *reversible* process of any type to the state L_T, which we designate conventionally as the reset state.

THEN The overall effect of the erasure process LP5 is to increase the thermodynamic entropy of the environment by $k \ln 2$. This represents a lower bound that will be exceeded if thermodynamically irreversible processes replace reversible processes.

The proof of the result depends largely on using relation (8) to compute the entropies S_L, S_R and S_{L+R} of the three states L_T, R_T or $(L+R)_T$. We have

$$\begin{aligned} S_L &= \frac{\bar{E}_L}{T} + k \ln \int_L \exp(-E_L/kT)\,\mathrm{d}x \\ &= \frac{\bar{E}_R}{T} + k \ln \int_R \exp(-E_R/kT)\,\mathrm{d}x = S_R, \end{aligned}$$

where the symmetry $E_L(x_L) = E_R(x_R)$ of LP4 assures equality of the above integrals and mean energies. We also have from the remaining symmetries that

$$\begin{aligned} S_{L+R} &= \frac{\bar{E}_{L+R}}{T} + k \ln \int_{L+R} \exp(-E_{L+R}/kT)\,\mathrm{d}x \\ &= \frac{\bar{E}_L}{T} + k \ln \left(2 \int_L \exp(-E_L/kT)\,\mathrm{d}x \right) = S_L + k \ln 2, \end{aligned}$$

where these symmetries also assure us that $\bar{E}_{L+R} = \bar{E}_L = \bar{E}_R$. Hence

$$S_L = S_R, \qquad S_{L+R} = S_L + k \ln 2 = S_R + k \ln 2. \tag{10}$$

Since the expansion LP5a is adiabatic, no heat passes between the device and the environment, so the process does not directly alter the environment's entropy. Since process LP5b is thermodynamically reversible, but its final state entropy S_L is lower than the initial state entropy S_{L+R} by $k \ln 2$, it follows that the process cannot be adiabatic and must pass heat to the environment, increasing the entropy of the environment by $k \ln 2$. This completes the proof.

This version of Landauer's principle is sufficiently general to cover the usual examples. Aside from the selection of the particular erasure procedure LP5, the principal assumptions are that the memory device states form a phase space to which ordinary statistical mechanics applies and that there are two regions L and R in it obeying the indicated symmetries. It is helpful to visualize the states and processes in terms of particles trapped in chambers, as

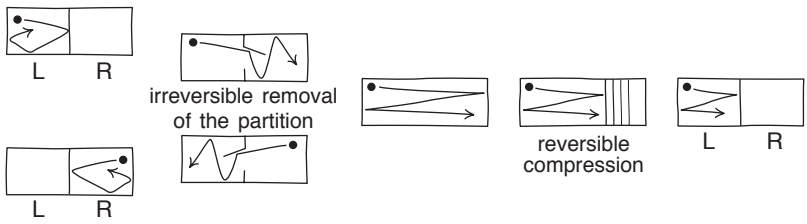

Fig. 2. The erasure process.

Fig. 2 suggests. However that visualization is far more specific than the result described.

The step LP5a is called "removal of the partition" since it is commonly illustrated as the removing of a partition that blocks the access of a single molecule gas to half the chamber. That the expansion of LP5a is thermodynamically irreversible seems unavoidable if the final result is to be attained. A thermodynamically irreversible (and adiabatic) process increases the entropy of the system by $k \ln 2$ in passing from S_L or S_R to S_{L+R}. If the process were thermodynamically reversible so that the total entropy of the device and environment would remain constant, then there would have to be a compensating entropy decrease in the environment of $k \ln 2$. That decrease would negate the entropy increase of LP5b. This expansion is the part of the erasure process that creates entropy. The second step LP5b, the "compression of the phase space", does not create entropy, since it is thermodynamically reversible. It merely moves the entropy created in the first step from the device to the environment. The step is commonly illustrated by such processes as the compression of a one-molecule gas by a piston. No such specific process need be assumed. Any reversible process that takes the state $(L + R)_T$ to L_T is admissible, since all reversible processes conserve the total entropy of the device and environment.

Setting aside unilluminating embellishments, this appears to capture the most general sense in which erasure of information held in memory devices in thermal equilibrium at a temperature T must increase the thermodynamic entropy of the environment. The principle does not license an unqualified entropy cost whenever an erasure occurs. It is limited by the assumption that a particular erasure procedure must be used, with the real entropy cost arising in the first step, the thermodynamically irreversible "removal of the partition". That this first step is the essential entropy generating step has been obscured in the literature by the erroneous assertion that this first step may sometimes be a thermodynamically reversible constant entropy process. As I will show in the following section that assertion depends upon the illicit assembly of many L_T and R_T states into what is incorrectly supposed to be an equivalent canonical ensemble $(L + R)_T$.

It may seem that we can generate Landauer's principle with a much simpler and more general argument that calls directly on expression (9) for entropy. Prior to erasure, we are unsure of whether the memory device is in state L or in state R. So we

assign equal probabilities to them:

$$P(L) = P(R) = 1/2.$$

According to (9), the entropy of the probability distribution is $k \ln 2$. After erasure, we know the device is in state L. The probabilities are now

$$P(L) = 1, \quad P(R) = 0.$$

According to (9), the entropy of this probability distribution is 0. Erasure has reduced the entropy of the memory device by $k \ln 2$. Is this not just what Landauer's principle asserts?

No, it is not. Landauer's principle asserts that the *thermodynamic* entropy is reduced by $k \ln 2$. As we saw at the end of Section 2.2, expression (9) does not return the thermodynamic entropy of a system in thermal equilibrium unless the probability distributions inserted into it are canonical distributions. The initial probabilities $P(L) = P(R) = 1/2$ are not canonical distributions. They reflect our own uncertainty over the state of the device. While we may not know which, the device is assuredly in one of the states L_T or R_T. Each of the two states has its own canonical distribution, which represents the device's disposition in the region of phase space accessible to it. As a result the above argument fails to establish Landauer's principle for thermodynamic entropy.

What is dangerously misleading about the argument is that the distribution $P(L) = P(R) = 1/2$ will coincide with the canonical distribution of the device half way through the process of erasure, after the removal of the partition, when the device is in state $(L + R)_T$. The argument then returns the correct thermodynamic entropy reduction in the device during the second step, the compression of the phase space. But it remains silent on the first step, the "removal of the partition", the essential thermodynamic entropy generating step of the erasure process, and tempts us to ignore it.

2.4. A compendium

It will be useful for later discussion to collect the principal results in thermodynamics and statistical mechanics of this section.

Thermal equilibrium: A system in thermal equilibrium is represented by a canonical distribution (2). Its thermodynamic entropy is given uniquely by expression (8), which is a special case of (9) that arises when the probability distribution $p(x)$ is the canonical distribution.

Accessible regions of the phase space: These are the regions of the phase space that a system in thermal equilibrium can access over time as a part of its thermal motion. They are demarcated by the energy function $E(x)$ of the canonical distribution as those parts of the phase space to which finite energy is assigned. Since the problems of ergodicity raise issues that are apparently unrelated to Landauer's principle, I will assume here that the thermal systems under examination have the sorts of properties that the early literature on ergodic systems hoped to secure. Most notably, I assume that over time a system densely visits all portions of the accessible phase space and that the probability a canonical distribution assigns to each region in the accessible

phase coincides with the portion of time the system spends there. That the system visit these regions more rapidly than the rate of a reversible change is essential to the derivation above that identifies (8) as the thermodynamic entropy of the system.

Compression of the phase space: This compression arises when the accessible region of the phase space is reduced by external manipulation of the energy function $E(x)$. The compression is associated with a reduction in thermodynamic entropy of the system, insofar as the compression reduces the integral $k \ln \int_\Gamma \exp(-E(x)/kT)\,\mathrm{d}x$ of expression (8) for thermodynamic entropy. As long as suitable symmetry requirements are met, a halving of the accessible phase space will reduce the thermodynamic entropy by $k \ln 2$.

Creation of thermodynamic entropy in erasure: In the erasure process described, $k \ln 2$ of thermodynamic entropy is created in the first, irreversible step, the "removal of the partition". Without the thermodynamically irreversibility of this step, there would be no thermodynamic entropy cost associated with erasure.

3. Illicit ensembles and the failure of the many-to-one mapping argument

3.1. The use of ensembles in statistical mechanics

There is a standard procedure used often in statistical mechanics through which we can develop the probability distribution of a single component in its phase space by assembling it from the behavior of many like components.[7] One familiar way of doing this is to take a single component and sample its state frequently through its time development. The probability distribution of the component at one moment is then recovered from the occupation times, the fractional times the system has spent in different parts of its phase space during the history sampled. For example, we might judge that a molecule, moving freely in some chamber, spends equal time in all equal sized parts of the chamber. So we infer that its probability distribution at one time is uniformly distributed over the chamber. Another way of doing it is to take a very large collection of identical components with the same phase space—an "ensemble"—and generate a probability distribution in one phase space from the relative frequency of the positions of the components in their own phase spaces at one moment in time. For example, we may consider very many identical pollen grains suspended in water at temperature T and judge that the number with thermal energy E is proportional to $\exp(-E/kT)$. We immediately conclude that the probability that some particular pollen grain has thermal energy E is proportional to this same factor. At its very simplest, the procedure might just collapse the probability distributions of many phase spaces down to one phase space. We might take the probability distributions of one component at different times; or we might take the probability distributions of many components from their phase spaces.

[7] By a component, I mean a few degrees of freedom of a system with very many degrees of freedom: such as an individual atom in a crystal lattice; or several atoms forming a molecule in a gas; or a single spin in a magnetic spin system; or a pollen grain in water.

Carried out correctly, this form of the procedure is rather trivial, since all the distributions are the same. In all cases, the result is a probability distribution in one phase space at one moment that represents the thermodynamic properties of one component. This technique is so common that we freely move from individual components to ensembles and back and sometimes even speak of ensembles when we intend to speak of just one component.

Let us now consider how this process would proceed for forming a canonical distribution (2), $p(x) = \exp(-E(x)/kT)/Z$. First recall how this distribution can be derived. When we have many components in thermal equilibrium, the canonical distribution is generated *uniquely* from the demand that thermal equilibrium correspond to the most probable distribution of energy; and it is essential to that derivation that that energy function $E(x)$ represents the energy the component would have, were it at phase space position x, with x a position accessible to the component. Thus, in generating the canonical distribution for one component from an ensemble, one constraint is essential: *the phase spaces sampled, either through time or by visiting different components, must have the same energy function* $E(x)$. It is an obvious but absolutely fundamental point that one cannot assemble a canonical distribution properly representative of an individual component by sampling from a single component at times when the energy function $E(x)$ is different; and that one cannot form such a canonical distribution by collapsing the phase space position frequencies or probability distributions from components with different energy function $E(x)$ in their phase spaces. For then the energy function $E(x)$ would not represent the energies at phase space points accessible to the component; or it would not represent the correct energy for the component at accessible points in phase space. Whatever might result from such an illicit procedure would not correctly represent the thermodynamic properties of just one ensemble member. It would not be licit, for example, to apply the thermodynamic entropy formula (8) to it to recover the thermodynamic entropy of a component.

Consider sampling from the successive states of the compression process illustrated in Fig. 1. The sampling must take place during a sufficiently short time period so that the energy function is, for all intents and purposes, unchanged. Or consider what happens if we try to combine the initial and final states of the relocation process also illustrated in Fig. 1. Since they have disjoint phase spaces, neither state will be properly represented by the resulting distribution that spans both regions of the phase space. Fig. 3 illustrates the mathematical process of collapsing a canonical ensemble into a single canonical distribution that can properly represent each individual member of the ensemble; and the process of collapsing an illicit canonical ensemble to produce a distribution that properly represents no individual member.

Finally, even if one has an ensemble of canonically distributed systems, they cannot be treated as multiple clones of a *single* canonically distributed system unless the energy functions $E(x)$ is the same in each member of the ensemble. To do otherwise would be an error. Unfortunately, this error seems to be quite pervasive in the Landauer's principle literature.

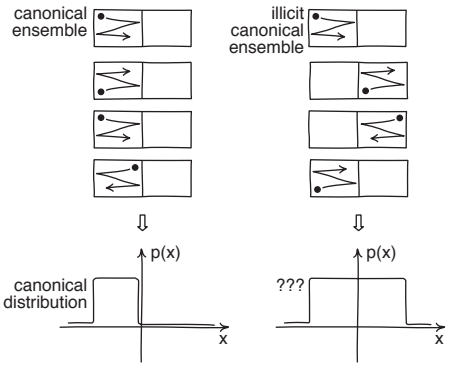

Fig. 3. Licit and illicit canonical ensembles and the distributions they produce.

3.2. Illicit ensembles

We must be grateful to Leff and Rex (2003) for giving us an uncommonly clear survey of this literature in the introductory chapter of their collection. While I will quote their text as clearly expressing the error, I want to emphasize that they are merely reporting more clearly than elsewhere what appears to be standard currency in the literature. In discussing the erasure procedure of Landauer's principle for the case in which the memory device is a partitioned box containing a single molecule, they note:

> The diffusion process in the erasure step (i) [removal of the partition] eradicates the initial memory state. Despite the fact that this process is logically irreversible, it is thermodynamically reversible for the special case where the ensemble has half its members in state L and half in state R. This is evident from the fact that partition replacement leads to the initial thermodynamical state (assuming fluctuations are negligibly small).[8]
>
> How has the ensemble entropy of the memory changed during the erasure process? Under our assumptions, the initial ensemble entropy per memory associated with the equally likely left and right states is $S_{LR}(\text{initial}) = k \ln 2$. After erasure and resetting, each ensemble member is in state L, and $S_{LR}(\text{final}) = 0$. Therefore $\Delta S_{LR} = -k \ln 2 = -\Delta S_{\text{res}}$. In this sense, the process is *thermodynamically reversible*; i.e. the entropy change of the universe is zero. This counterintuitive result is a direct consequence of the assumed uniform initial distribution of ensemble members among L and R states (Leff & Rex, 2003, p. 21).

[8](JDN) The apparent presumption is that the insertion of the partition does not alter the thermodynamic entropy of the memory devices. This directly contradicts the central assumption of the Szilard one-molecule gas engine (described in Section 4), in which the replacement of the partition reduces the thermodynamic entropy of the one-molecule gas by $k \ln 2$. This reduction in thermodynamic entropy is what Maxwell's demon seeks to exploit in the standard examples.

In the following paragraph, Leff and Rex consider the reverse process. They consider the memory cells without their partitions so the molecule has access to both L and R regions. They continue:

> Subsequent placement of the partition has zero entropic effect, because (approximately) half the ensemble members are likely to end up in each of the two states. (Leff & Rex, 2003, p. 21)

Fig. 4 helps us visualize the point.

The claim is that the thermodynamic entropy per cell in the set of cells with random data—as many L as R—is the same as the thermodynamic entropy of the cells in which the partition has been removed; and it is $k \ln 2$ greater than the thermodynamic entropy of an identical cell in the set of reset cells.

This is incorrect. The correct thermodynamic entropies are recovered by applying expression (8) to the canonical distributions of molecules in each cell, exactly as shown in Section 2.3. I went to some pains in Section 2.2 to show that this expression (or ones equivalent to it) is the only admissible expression for the thermodynamic entropy of a canonically distributed system. The calculation is straightforward and the results, given as (10), are unequivocal. When each of the cells with the random data has its partition removed, its accessible phase space doubled. That unavoidably increases its entropy by $k \ln 2$. A cell showing L has the same entropy whether it is a member of the cells carrying random data or a member of the cells that have all been reset to L. Thermodynamic entropy is a property of the cell and its physical state; it is not affected by how we might imagine the cell to be grouped with other cells.

How could we come to think otherwise? Whatever may have been their intention, the appearance is simply that the collection of cells carrying random data is being treated illicitly as a *canonical* ensemble, as suggested by the naming of the collection an "ensemble". Thus, all the results of Section 2 could be taken to apply. Each of the cells from the collection carrying random data occupy twice the volume of phase space; the cells are reset to a state that occupies half the volume of phase space; therefore, their entropy is reduced by $k \ln 2$ per cell. Yet the collection of cells carrying random data is clearly not a canonical ensemble. We cannot take the probability distributions for each individual cell and collapse them down to the one

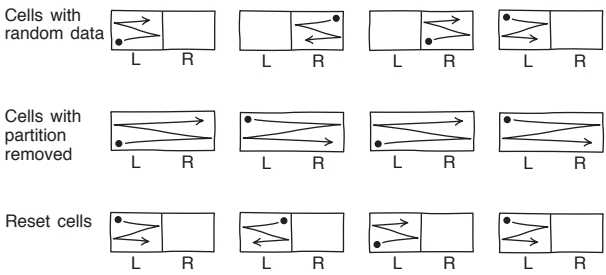

Fig. 4. Collections of memory cells.

phase space to produce a distribution that captures the properties of all. The collapse is illicit insofar as cells showing L and cells showing R have different energy functions $E(x)$. The energy function of the cells showing L is finite (and presumably small) in the L region of the phase space; it is infinite outside that region. Conversely, the energy functions of the cells showing R are finite in the R region; and infinite outside.[9] The resulting illicitly formed distribution extends over both L and R regions of the phase space. So we might take it to be equivalent to the canonical distribution of a cell with the partition removed. To do so would be to conclude incorrectly that each of the random data cells and each of the cells with the partitions removed have the same entropy.

If the collection is not being treated as a canonical ensemble, it is hard to understand how the results pertaining to thermodynamic entropy and heat generation could be recovered. It has been suggested to me that Leff and Rex's argument depends on introducing a second probability distribution that is intended to simulate the entropic properties of a canonical distribution. If a cell carries random data, its uncertainty would be represented by a probability distribution $F(u)$, where $u = L$ or R. Entropy formula (9) assigns what we will call an information theoretic entropy to the distribution

$$S_{\text{info}} = -k \sum_{u=L,R} F(u) \ln F(u). \tag{11}$$

The overall effect of erasure is to reduce the uncertainty of the data. That is, erasure might take equidistributed L and R data with $F(L) = F(R) = 1/2$ to reset data with $F(L) = 1$ and $F(R) = 0$. The associate change in S_{info} according to (11) would be a reduction by $k \ln 2$, just the number that Landauer's principle requires. Since the thermodynamic entropy of cells carrying random data and reset data is the same, may we not locate the entropy change of Landauer's principle in this information theoretic entropy change?

Promising as this possibility may seem, there is an immediate and, I believe, fatal difficulty (as already indicated in Section 2.3). *It is the wrong sense of entropy*. I showed in Section 2.1 that the sense of entropy at issue in exorcisms is thermodynamic entropy. S_{info} of Eq. (11) is not thermodynamic entropy; insertion of a probability distribution into formula (9) does not yield thermodynamic entropy unless the probability distribution is a canonical distribution. Thus, while it may appear that the erasure process

$$F(L) = F(R) = 1/2 \rightarrow F(L) = 1, \;\; F(R) = 0$$

is a kind of compression of the phase space, it is not the type of compression reviewed in Section 2.4 that would be associated with a reduction of thermodynamic

[9] Of course in practice, the molecule in the cells showing L is confined to the L portion of phase space by an impenetrable barrier, so the actual energy of the molecule, were it to get through might be quite small. What concerns us here is the form of the energy function needed to maintain the expression for the canonical distribution $p(x) = \exp(-E(x)/kT)/Z$. Since $p(x)$ must be zero for the R portions of phase space, $E(x)$ must be infinite there. For further discussion, see Appendix A.

entropy, for it is not the reduction in the accessible volume of phase space of a canonically distributed system. Finally, we cannot associate a quantity of heat with the change of information theoretic entropy. For we have no rule to associate its change with the exchange of heat with the surroundings. The rule (1) that associates heat transfer with entropy holds only for thermodynamic entropy and, indeed, defines it. No other entropy can satisfy it without at once also being thermodynamic entropy.[10]

The entropy Leff and Rex track through the erasure process is apparently the sum of the thermodynamic entropy (henceforth "S_{thermo}") and this information theoretic entropy S_{info}. Using the resulting augmented entropy,[11] $S_{\text{aug}} = S_{\text{thermo}} + S_{\text{info}}$, the thermodynamically irreversible process of the "removal of the partition" turns out to be a constant augmented entropy process. In it, for cells carrying random data, the increase of $k \ln 2$ of thermodynamic entropy S_{thermo} in each cell is exactly compensated by a decrease of $k \ln 2$ of information theoretic entropy S_{info}. In traditional thermodynamics, a thermodynamically reversible process has constant thermodynamic entropy.

Presumably this aids in motivating the labeling of the process of "removal of the partition" as "thermodynamically reversible" by Leff and Rex above. Further, there is a sense in which the process of "removal of the partition" is reversed by replacement of the partition. The replacement reconfines the molecule to one half of the box. But it is only a partial sense of reversibility. If we associate a state with a canonical distribution, then the process will only succeed in restoring the original

[10]We may seek a rule somehow analogous to (1) that would connect information theoretic information with transfers of heat in some sort of extended theory of thermodynamics. The difficulty is that the new notion is incompatible with virtually every standard property of thermodynamic entropy, so that the entire theory would have to be rebuilt from scratch. To begin, the augmented entropy S_{aug} is no longer a function of the state of a thermal system, as is thermodynamic entropy. One memory device in one fixed physical state, displaying an L, say, can have different entropies according to how we conceive the data. Is it carrying reset data? Or is it carrying random data?—in which case the S_{info} term increases its augmented entropy by $k \ln 2$. Also constant augmented entropy processes will no longer be the least dissipative and will no longer be thermodynamically reversible in the sense of being sequences of equilibrium states as indicated in the text. Then we would need a surrogate for the second law of thermodynamics. We cannot simply assume that summed augmented entropy will be non-decreasing for isolated systems, as is thermodynamic entropy. We must find the law that applies and we must find some appropriately secure basis for it, so it is a law and not a speculation. That basis should not be Landauer's principle itself, lest our justification of Landauer's principle becomes circular.

[11]That these two entropies can be added to yield augmented entropy follows if augmented entropy conforms to expression (9). Consider just one memory device. Its state is represented by a dual probability distribution that combines the new distribution $F(u)$ and the canonical distributions $p_L(x)$ and $p_R(x)$ of devices in states L_T and R_T, respectively. The combined distribution is $p(u, x) = F(u)p_u(x)$. Substituting this distribution into expression (9), we find

$$S_{\text{aug}} = \sum_{u=L,R} F(u)\left(-k \int_u p_u(x) \ln p_u(x)\, \mathrm{d}x\right) - k \sum_{u=L,R} F(u) \ln F(u) = S_{\text{thermo}} + S_{\text{info}},$$

where the separation into two added terms depends essentially on the probabilistic independence of $F(u)$ from $p_L(x)$ and $p_R(x)$. The first term, S_{thermo}, is a thermodynamic entropy term insofar as it is the weighted average of the thermodynamic entropies of the device in states L_T and R_T.

state half the time, for, in only half the trials, will the molecule be reconfined to the side from which it started.

If we adopt this partial sense of reversibility as the notion of reversibility to be employed in thermodynamic analysis, we adopt a ruinous redefinition of the notion that compromises the cogency of thermodynamics. To see this, recall that, traditionally, a thermodynamically reversible process is one that can proceed in either direction and a reversed process fully restores the original state. A basic result of thermodynamics, due originally to Carnot, is that these processes are the least dissipative. They are closely related to quasi-static processes; that is, sequences of equilibrium states or ones removed from equilibrium to arbitrarily small degree. The two are standardly taken to be co-extensive, in brief, for the following reasons. Any quasi-static process is reversible since the slightest disturbance to the equilibrium in either direction will enable the process to proceed in that direction. And only such processes are reversible, since disequilibrium processes, driven by large departures from equilibrium, are more dissipative than quasi-static processes.[12]

A paradigm of a process that is neither thermodynamically reversible nor quasi-static is the uncontrolled expansion that follows the "removal of the partition", in which a gas expands without performing work. It is a disequilibrium process. If we adopt a redefinition of thermodynamic reversibility so that this process is now to be counted as a thermodynamically reversible process, then such processes are no longer co-extensive with quasi-static processes.

One might be tempted to allow the two be decoupled. That would be a grave error. It is not a benign change that can be absorbed with minor adjustments into standard thermodynamics. It requires the complete reconstruction of thermodynamics. It entails that thermodynamic entropy, as defined by Clausius formula (1), is not a state function. To see this, take a memory cell holding random data; that is, one that is either in state L_T or in state R_T. Once the side of the cell holding the molecule is determined, that side undergoes a quasi-static, iso-thermal expansion in which it is expanded to fill the chamber, resulting in state $(L + R)_T$. During the expansion the cell absorbs $kT \ln 2$ of heat. The expansion is reversed by reinserting the partition, so the cell reverts to state L_T or R_T. If we count this reinsertion as a thermodynamically reversible step, we have for the complete cyclic process (illustrated in Fig. 5) that

$$\oint \frac{\mathrm{d}Q_{\mathrm{rev}}}{T} = k \ln 2$$

in contradiction with (1′). It immediately follows that entropy, as defined by the Clausius formula (1), is not a state function; thermal systems can no longer possess this entropy as a property. We must now try to build thermodynamics anew.

This new and risky notion of reversibility obscures the real reason that the standard erasure process LP5 passes a net amount of heat to the environment. While the first step ("removal of the partition") may be a constant augmented entropy process, it is still a thermodynamically irreversible process in the sense that it is a

[12]See Planck (1926, Sections 126–127); and for much detail on the complications of accommodating reversible processes with quasi-static processes, see Uffink (2001, Sections 7.1–7.2).

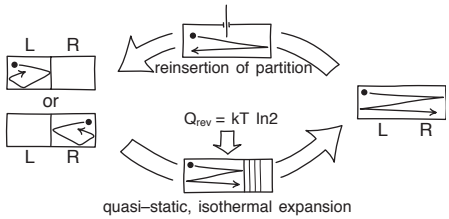

Fig. 5. Thermodynamic entropy no longer a state function with redefined notion of thermodynamic reversibility.

disequilibrium process that creates thermodynamic entropy. The thermodynamic entropy created appears as the net heat passed to the environment in the erasure process.

That Bennett is also in some way working with an illicit simulation of a canonical ensemble is the only way I can make sense of claims in his Bennett (1982, pp. 311–312). He considers the process of erasure of a bistable ferromagnet, in which the ferromagnet starts in one of the two states and is reset by a changing external field into one of them. If the initial state is "truly unknown, and properly describable by a probability equidistributed between the two minima [that would correspond to L and R above]", then he regards the process of erasure as thermodynamically reversible.[13] If on the other hand the initial state is "known (e.g. by virtue of its having been set during some intermediate stage of computation with known initial data)" then the erasure is logically and thermodynamically irreversible. Thus, he clearly maintains that two ferromagnets in identical physical states have entropies that differ by $k \ln 2$ according to whether we know what the state is or not.

Indeed, the "probability [distribution] equidistributed between the two minima" mentioned would seem to be the single probability distribution in the phase space of one memory device, intended to represent the thermodynamic properties of each of the memory devices carrying random data. While such a distribution can be defined in one phase space, it is not the canonical distribution that represents the thermodynamic properties of the collection of memory devices. It will employ an energy function $E(x)$ that allows all parts of the phase space to be accessible; each memory device (when recording data) will only have access to a portion of its phase space, so that it is represented by a different canonical probability distribution. This "equidistributed" distribution cannot be used to ascertain the thermodynamic entropy of each memory device by means of the results of Section 2.

We see in analogous remarks that Bennett has also adopted the redefined notion of thermodynamic reversibility:

> If a logically irreversible operation like erasure is applied to random data, the operation still may be thermodynamically reversible because it represents a

[13] Bennett also regards this first process as logically reversible. An anonymous referee has stressed to me that this is incompatible with Bennett's analyses elsewhere and seems to be a temporary aberration in his writings.

> reversible transfer of entropy from the data to the environment, rather like the reversible transfer of entropy to the environment when a gas is compressed isothermally. But if, as is more usual in computing, the logically irreversible operation is applied to known data, the operation is thermodynamically irreversible, because the environmental entropy increase is not compensated by any decrease of entropy of the data. (Bennett, 2003, p. 502)

What makes the difference in deciding whether an erasure process is thermodynamically reversible is whether the data are random. Apparently this refers to an analysis such as given above. The process of "removal of the partition" does increase thermodynamic entropy; but, only in case the data are random, is there a compensating reduction of information theoretic entropy, so that the augmented entropy of the process is constant, which is the new definition of a thermodynamically reversible process.

These approaches seem to be driven by the idea that randomness and thermalization can be equated. So a collection of devices carrying random data is supposed to be like a canonical ensemble. What is at stake, if we treat such collections of memory devices as canonical ensembles, is the additivity of entropy. To see this, we will draw on the states defined in Section 2.3. Consider a collection of $N/2$ memory devices L_T, each with thermodynamic entropy S_L, and $N/2$ memory device R_T, each with equal thermodynamic entropy $S_R = S_L = S$. If we collapse the distributions of all the memory devices down to one phase space in the obvious way, we recover a single probability distribution that is spread over both L and R regions of phase space that is actually the canonical distribution associated with the state $(L + R)_T$. So, recalling (10), if we insert this distribution into the entropy formula (9), we end up concluding that the thermodynamic entropy of each component is $S + k \ln 2$. Thus, the thermodynamic entropy of the entire collection would be $NS + Nk \ln 2$. That contradicts the additivity of thermodynamic entropy that assures us that a collection of N systems each in thermal equilibrium at temperature T and with entropy S has total entropy NS.

Or is the thought that the additivity of thermodynamic entropy is to be given up? Is the thought that each memory device might have the thermodynamic entropy $S_R = S_L = S$ individually; but the totality of N of them with random data, taken together, has a thermodynamic entropy greater than the sum NS of the individual entropies? That thought contradicts the standard formalism. Imagine, for example, that the N memory devices record some random sequence L, R, L, … It does not matter which it is or whether we know which it is. It is just some definite sequence. That thermal state is represented by a canonical distribution over the joint phase space of the N components:

$$\begin{aligned} p(x, y, z, \ldots) &= \exp(-E(x, y, z, \ldots)/kT)/Z \\ &= \exp(-(E_L(x) + E_R(y) + E_L(z) + \cdots)/kT)/Z, \end{aligned} \tag{12}$$

where the canonical phase space coordinates for the components are $x, y, z, \ldots$; $E_L(x)$ assigns finite energies to the L portion of phase space (and so on for the remainder); and Z is computed from (3). This is the *unique* canonical distribution

that represents the thermal property of the N devices. Applying the entropy formula (8) to this distribution, we recover that the entropy of the total collection is NS. That we might not know which particular sequence of data is recorded is irrelevant to the outcome of the computation. In every case, we recover the same thermodynamic entropy, NS.

To imagine otherwise—to imagine that the randomness of the data somehow defeats additivity—has the following odd outcome. Imagine that you know which particular sequence of L, R, … is recorded in the N memory devices, so for you the data are not random; but I do not know which is recorded, so the data are random for me. Then the supposition must be that the thermodynamic entropy of the N memory devices is $Nk\ln 2$ less for you than for me. And when you tell me which particular data sequence is registered, the thermodynamic entropy of the N devices for me will drop by $Nk\ln 2$ to NS, no matter which sequence I learn from you is the one recorded. We have the same outcome if we imagine that the thermodynamic entropy of a cell carrying data L, say, is increased by $k\ln 2$ if we happen not to know whether the cell carries data L or R. In these cases, thermodynamic entropy has ceased to be a function of the system's state.

3.3. Failure of the many-to-one mapping argument

The version of Landauer's principle of Section 2.3 has limited scope. It does not license a generation of $k\ln 2$ of thermodynamic entropy in the environment whenever a bit of information is erased. It licenses that generation only when a specific erasure procedure is followed. The common view in the literature is that the principle has broader scope. The argument advanced in support is based on the idea that erasure must map many physical states to one and this mapping is the source of the generation of thermodynamic entropy. The argument appeared in Landauer's (1961) early work. Remarking on the possibility that an erasure process might not immediately return the memory device fully to its reset state, he noted:

> Hence the physical "many into one" mapping, which is the source of the entropy change, need not happen in full detail during the machine cycle which performed the logical function. But it must eventually take place, and this is all that is relevant for the heat generation argument. (Landauer, 1961, p. 153)

Here is a recent version of this many-to-one mapping argument:

> While a computer as a whole (including its power supply and other parts of its environment), may be viewed as a closed system obeying reversible laws of motion (Hamiltonian or, more properly, for a quantum system, unitary dynamics), Landauer noted that the logical state often evolves irreversibly, with two or more distinct logical states having a single logical successor. Therefore, because Hamiltonian/unitary dynamics conserves (fine-grained) entropy, the entropy decrease of the [information bearing degrees of freedom] during a logically irreversible operation must be compensated by an equal or greater entropy

> increase in the [non-information bearing degrees of freedom] and environment. This is Landauer's principle. (Bennett, 2003, p. 502)[14]

That is, a single logical state is represented by a single physical state. A single physical state is represented by a volume of phase space. Hamiltonian dynamics conserves total phase space volume, so the reduction in phase space volume in one part of phase space must be compensated by an expansion elsewhere. And since entropy may be (cautiously!) associated with volumes of phase space, these changes in phase space volume may be translated into entropy changes.

The central assumption of this argument is that memory cells prior to erasure (the "many" state) occupy more phase space than the memory cells after erasure (the "one") state. It should now be very clear that this assumption is incorrect. Or at least it is incorrect if by "compression of the phase space" we mean the reduction of the accessible region of the phase space of a canonical ensemble as described in Section 2.4. And we must mean this if we intend the compression to be associated with a change of thermodynamic entropy for a system in thermal equilibrium at temperature T. To see the problem, take the memory cells described in Section 2.3. Prior to erasure, the memory device is in state L_T or it is in state R_T (but not both!). After the erasure it is in state L_T. Since the states L_T and R_T have the same volume in phase space, there is no change in phase space volume as a result of the erasure procedure. *A process of erasure that resets a memory device in state L_T or in state R_T back to the default state L_T does not reduce phase space volume in the sense relevant to the generation of thermodynamic entropy!*

How could such a simple fact be overlooked? Clearly part of the problem is that an intermediate stage of the common erasure procedure LP5 is an expanded state $(L + R)_T$ that occupies more volume in phase space. But that state is arrived at by expanding the phase space in the first step by exactly as much as it will be compressed in the second. Reflection on that fact should have revealed that the erasure process overall does not compress phase space.

What obscured such reflection is the real reason for the persistence of the many-to-one mapping argument. It was obscured by the assembly of many memory devices with random data into an ensemble that is taken to be or to simulate a canonical ensemble. As a result, the erasure of random data is incorrectly associated with a reduction in volume of the accessible region of a phase space of a canonical ensemble and a reduction of thermodynamic entropy is incorrectly assigned to the process.

The many-to-one mapping argument fails.

[14]The text identifies the principal problem with the many-to-one mapping argument. There are others. For example, the argument employs *fine-grained* entropy. Since *fine-grained* entropy can neither increase nor decrease in isolated systems, it does not correspond to thermodynamic entropy, which can increase in isolated systems. If we are to define entropy in terms of volumes of phase space, the more natural choice for this application is *coarse-grained* entropy, since coarse-grained entropy may increase over time. Its use, however, will greatly complicate the many-to-one argument. While it is extremely unlikely, Hamiltonian dynamics does allow the reduction of the coarse-grained entropy of isolated systems. It is just this possibility that was the original inspiration for the proposal of Maxwell's demons. Could they somehow convert "extremely unlikely" into "likely"?

3.4. Must erasure always be thermodynamically irreversible?

When we renounce illicit canonical ensembles and abandon the failed many-to-one mapping argument, we are left without any general reason to believe an unconditional Landauer's principle. We are able, however, to reformulate the question of the range of applicability of Landauer's principle, since we can now recognize that memory devices with random data have the same thermodynamic entropy as memory devices with default data. The association of an unavoidable thermodynamic entropy cost with an erasure process is equivalent to the necessity of the erasure process being thermodynamically irreversible.[15] So we ask: must erasure always be thermodynamically irreversible?[16]

It is the case that the erasure processes of the familiar examples are thermodynamically irreversible. For example, erasure processes actually used in real memory devices, of the type referred to by Landauer and others, employ thermodynamically irreversible processes. We also use thermodynamically irreversible processes in our standard implementations of erasure in thought experiments, such as memory cells that employ one-molecule gases. The standard process of this type, a thermodynamically irreversible expansion followed by a thermodynamically reversible compression, is the one incorporated into the restricted version of Landauer's principle stated in Section 2.3.

What we lack is a principled demonstration that all erasure processes must be thermodynamically irreversible. Our experience with familiar examples does count for something—but it is not enough. That we find thermodynamic irreversibility in real examples and the small stock of fictitious examples used and re-used in our thought experiments falls short of the general assurance needed. We need to be assured that all erasure processes must be thermodynamically irreversible, no matter how wildly they may differ from the familiar examples. For the claim to be examined is that Landauer's principle is sufficiently powerful to preclude all Maxwell's demons, which must include extraordinary devices that employ extraordinary processes.

Perhaps there are some unrecognized principles that govern the real or commonly imagined examples and if we could find them then we could give a really universal basis to Landauer's principle. But we are far from identifying them and, therefore, even further from knowing if those principles reflect some deeper fact of nature or merely a limitation on our imagination.

In this context, we should consider a demand sometimes explicitly placed on an erasure process (e.g. Bub, 2001, p. 573): that it must be indifferent in its operation to whether the state erased is L or R. First, it remains to be shown that this demand would force all possible erasure processes to be thermodynamically irreversible.

[15] If a thermodynamically reversible process takes an unerased memory device to an erased memory device at the same entropy, then there will be no net change in the entropy of the environment, since a thermodynamically reversible processes conserves thermodynamic entropy. An irreversible process increases total thermodynamic entropy; and that increase must appear in the environment as a thermodynamic entropy cost since the entropy of the memory device is unchanged.

[16] And if so, must it always come with a cost of $k \ln 2$ of thermodynamic entropy for each bit erased?

Rather, all we know is that the thermodynamically irreversible step of the ordinary erasure processes somehow seems associated with this indifference: the one removal of the partition allows an L state or an R state to expand irreversibly to fill the full phase space. But how are we to bridge the gap between our limited repertoire of standard examples and all possible erasure procedures? How do we know that this indifference cannot be implemented in some extraordinary, thermodynamically reversible process? Second, we seem to have no good reason to demand that the erasure procedure must be indifferent to the state erased. It certainly makes the erasure process easy in ordinary examples: we remove the partition and then compress. But that is no reason to believe that no other way is admissible. Leff and Rex (2003, p. 21) state the reason that may well be tacitly behind other assertions of the demand: "[dual procedures] would necessitate first determining the state of each memory. After erasure, the knowledge from that determination would remain; i.e. erasure would not really have been accomplished". The reason is plausible as long as we continue to think of ordinary devices and the eraser as, for example, a little computer with its own memory. But what assures us that, in all cases, the eraser must be a device of this type? Might it not function without recording states in a memory device of the type governed by Landauer's principle? Or if it does record states, why can it not use the very state under erasure to keep track of the procedure being followed?

How might such an erasure procedure look? I am loath to pursue the question since any concrete proposal invites a debate over the cogency of a particular example that once again obscures the real issue—that the burden of proving Landauer's principle remains unmet. As it turns out, however, such a procedure is described, in effect, in Section 4.2. We need to reconceive of the Szilard one-molecule gas engine as itself a memory device that records an L or an R according to the side of the partition on which the molecule is trapped. The "no-erasure demon" described a little later in the section erases that record using different processes according to whether the record is an L or an R and does it in a thermodynamically reversible manner.

4. Why Landauer's principle fails to exorcise Maxwell's demon

4.1. The challenge of Maxwell's demon

Our present literature on Maxwell's demon derives from the early twentieth century when it was finally established that thermal processes were statistical processes. It became clear that the second law of thermodynamics could not be unconditionally true. There were admissible mechanical processes that violated it. Small-scale violations arose in observable fluctuation phenomena, such as the Brownian motion of a pollen grain visible under a microscope. The obvious question was whether these microscopic violations of the second law could somehow be accumulated to produce macroscopic violations. Numerous mechanisms for doing just this were proposed. As we describe in our short survey in Earman and Norton

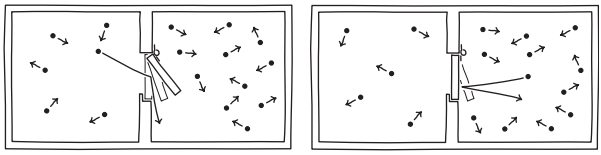

Fig. 6. Smoluchowski trapdoor.

(1998, Sections 4–6), a consensus rapidly developed. All the mechanisms imagined would fail since their intended operation would be disrupted by fluctuations within their own machinery. One of the best-known examples is the Smoluchowski trapdoor of Fig. 6.

A spring loaded trapdoor is installed in a wall separating gases initially at equal temperature and pressure. It is set up so that gas molecules striking the door from one side will pop the door open, allowing the molecule to pass. But it does not allow molecules to pass in the other direction since the impact of these molecules slams the trapdoor shut. The expected outcome is that the pressure of gas on one side spontaneously diminishes, while it increases on the other, a clear violation of the second law of thermodynamics. What was neglected in this expectation is that the spring restraining the trapdoor must be weak and the trapdoor very light weight if molecular collisions are to be able to open it. But under just these conditions, the trapdoors own thermal energy of $kT/2$ per degree of freedom will lead to wild flapping that will defeat its intended operation. The Smoluchowski trapdoor was just one of the many mechanisms proposed. They included mechanical devices with one-way ratchets and pawls and electrical systems in which charged colloids are spontaneously cooled by external absorption of the electromagnetic radiation they emit. All these devices were deemed to fail because of disruption by further overlooked fluctuation phenomena.

However, a nagging worry remained. What if these devices were operated by an intelligent agent who could somehow cleverly evade the disruptions of fluctuations? As we describe in Earman and Norton (1998, Sections 8–9) this was the problem tackled by Szilard in his 1929 paper that drew attention to the entropy costs of information processing. It is of course standard to naturalize the demon as a very complicated physical system, perhaps even as intricate as a human being, but still governed by ordinary physical laws. In this context, we can now pose:

Problem of Maxwell's demon.

Is there any device, possibly of extremely complicated construction and of devious operation, that is able to accumulate fluctuations into a macroscopic violation of the second law of thermodynamics?

Clearly what will not suffice as a solution is to notice that simple devices—spring loaded trapdoors, ratchets and pawls—fail. The concern is that something vastly more complicated might circumvent the weaknesses of the simple devices. Clearly what will not suffice is to notice that most *ordinary* systems adhere to the second law.

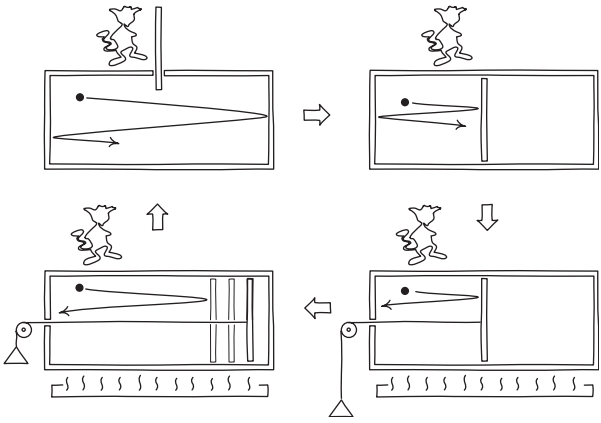

Fig. 7. Maxwell's demon operates Szilard's one-molecule gas engine.

The concern is that there might be some *extraordinary* device that does not. Purely mechanical considerations do not preclude it.

4.2. Landauer's principle fails to exorcise all demons

Let us review the standard way Landauer's principle is used to exorcise Maxwell's demon. It is in the context of the one-molecule gas engine introduced by Szilard. As shown in Fig. 7, a chamber contains a single molecule maintained at temperature T by contact with a heat sink.

The demon inserts a partition-piston to trap the molecule on one side. It then allows the trapped molecule to expand reversibly and isothermally against the partition-piston. In this process, $kT\ln 2$ of heat is drawn from the heat sink and converted to work. This conversion is the net effect of the completed cycle and it amounts to a reduction in the entropy of the heat sink by $k\ln 2$, in violation of the second law. If we think of the motion of the molecule from left to right and back as a rather extreme form of pressure fluctuation that violates the second law, the repeated operation of this cycle would appear to accumulate the violations without limit.

To save the second law, we must find a hidden source of entropy. It is supplied by noticing that a successful demon must learn where the molecule is after the insertion of the partition—on the left "L" or on the right "R".[17] In order to complete the cycle, the demon's memory must be reset to its initial state. That is, the demon must erase one bit of information. By Landauer's principle that erasure will increase the entropy of the thermal environment by $k\ln 2$, thereby restoring the second law.

[17]Bennett has urged that the older tradition erred in attributing an entropy cost to the acquiring of this information and that it can be acquired by thermodynamically reversible processes. However, as we noted in Earman and Norton (1999, pp. 13–14), the successful operation of the thermodynamically reversible processes proposed would be disrupted by fluctuations in exactly the same way as fluctuations disrupt the operation of simple, mechanical Maxwell's demons.

No one can doubt the elegance of this analysis. However, we should not allow that elegance to hide its shortcomings. They all come down to one problem. We face one particular attempt to reverse the second law that is dependent upon a particular, if natural, set of presumptions about the engineering design of the demon. We are to believe that the designs of all demons admissible under physical law will be sufficiently similar for them to be prone to the same exorcism. This might well be reasonable as long as we deal with ordinary devices that are not too different in their essential designs from the one just described. But that is not the problem to be solved. The challenge of exorcising Maxwell's demon is to demonstrate the failure of all demons, including extraordinary devices of potentially great complication devised by the most ingenious designers.

If this challenge is to be successfully overcome, we must assume that the engineering of these demons manifests all the essential elements of the standard examples: each demon must process information computationally; it must store information in a memory device; it must erase as part of its operation; and it must do it according to the two-step erasure procedure (''removal of the partition'' and ''compression of the phase space'') described in Section 2.3. This presumption is unrealistic for the analogy can fail on any or all of the above points. Yet if Landauer's principle is to be *the* agent that exorcises all Maxwell's demons, then the analogy must always succeed. Let us review some of the ways it can fail:

Demons that do not process information: We can readily conceive demonic devices in which information is neither acquired nor acted upon. In 1907, Svedberg (see Earman & Norton, 1998, pp. 443–444) proposed a simple device that was intended to reverse the second law. In brief, charged particles in a colloid emit their thermal energy as electromagnetic radiation that is excited by their thermal motions. A precisely shaped and located lead casing surrounds the colloid and absorbs the radiation. The colloid cools and the casing heats in direct violation of the second law of thermodynamics. The device acquires and processes no information in any obvious sense. It just sits there, supposedly warming and cooling. While you may find this device too simple to merit the name ''Maxwell's demon'', on what basis are we to preclude more complicated demonic devices that do not process information? If there is no information processed, there is no information erasure and Landauer's principle is irrelevant to the exorcism. There is no evident sense in which Landauer's principle explains the failure of Svedberg's device.[18]

Non-computational demons: There are Maxwell's demons that manipulate the system they act upon without being representable as computing devices. A simple example is the Smoluchowski trapdoor. Perhaps we might contrive to imagine the trapdoor as a sort of computer with its momentary position and motion a sort of memory of past processes. However, Landauer's principle is still not relevant to the exorcism since its two-step erasure process is clearly not present.

[18]Svedberg's device fails for familiar reasons. Thermal fluctuations in the lead casing would generate heat radiation that would pass back from the casing to the colloid and a thermal equilibrium would be established.

Non-standard computational demons: While some Maxwell's demons may be well represented as computational devices, our natural presumption is that they have standard architecture: a central processing unit that does the thinking and a memory device that must be erased. How can we be sure that no other architecture is possible in which there is no distinct memory device requiring erasure as a distinct step in the device's operation? Recall that the device need not have the power of a universal Turing machine that is able to run any program. The demon is a special purpose device that performs one function only. A suitable model would be a thermostat, which is able to respond differently to high and low temperatures without needing distinct memory devices. Or if we contrive to imagine some portion of its control circuitry as a memory device, is it a memory device of the right type that must be erased by the two-step procedure of Landauer's principle?

Different erasure protocols: In the case in which the demon does harbor a memory device that undergoes erasure, what assurance do we have that the erasure process must follow the two steps "removal of the partition" and "compression of the phase space" of Landauer's principle? I raised the possibility in Section 3 of alternative procedures that replaced the thermodynamically irreversible "removal of the partition" by a thermodynamically reversible step. What assurance do we have of the absolute impossibility of this or some other erasure procedure that is incompatible with Landauer's principle?

Must entropy costs match entropy gains? What is striking about the Szilard one-molecule gas engine (and its generalizations to n-fold compression) is that entropy reduction arising in the operation of engine is exactly balanced by the entropy cost of erasing the demon's memory. But this is just one example. What assurance do we have that the two will balance so perfectly no matter what the system is that the demon operates on and now matter how ingeniously the demon is designed?[19]

No erasure computational demons: Finally, even if one presumes that Maxwell's demon is a computational device with the standard architecture and that it erases using the procedures of Landauer's principle, it remains to be shown that such a demon must actually perform erasures. Consider for example a no-erasure Maxwell demon that operates on a Szilard one-molecule gas engine as described in greater detail in Earman and Norton (1999, pp. 16–17). The demon functions by combining operations that are accepted as admissible individually in the Landauer's principle–Maxwell's demon literature. It uses two subprograms, program-L and program-R, which are invoked according to whether the demon finds the molecule on the left- or on the right-hand side of the chamber. (Recall that the present orthodoxy holds that this detection can be carried out in a thermodynamically reversible process, contrary to the orthodoxy of the 1950s.) Which subprogram is to be run is recorded in a single memory device with two states L and R that also records the location of the molecule. The initial and

[19] Earman and Norton (1999, p. 19) describes a demon that is programmed sufficiently economically for the entropy cost of erasure to be less than the entropy reduction achieved in the operation of the engine.

default state of the device is L and, upon measurement of the position of the molecule, the first steps of the program leave the device in the L state if the molecule is on the left; or they switch it to the R state if it is on the right. The demon then runs program-L or program-R according to the content of the memory device. Program-L leaves the memory device unaltered. Program-R concludes with its last step by switching the memory device from R to L. This last step is a switching and not an erasure; Program-R takes a memory device known to it to be in the R state to the L state. At no point in this cycle is there an erasure operation, so Landauer's principle is never invoked.

The principal assumption of the no-erasure demon is that a memory device can be switched from L to R or conversely without thermodynamic entropy cost.[20] The process that effects this switching is illustrated in Fig. 1. The process is thermodynamically reversible and requires no net expenditure of work; the work needed to advance one wall is recovered from the recession of the other.[21]

What these questions and examples suggest is that Landauer's principle is far from the vehicle that exorcises all of Maxwell's demons. Rather the range of demons to which it applies is small and with ill-defined borders. The only demons that are assuredly covered are those that can be represented as computers that have distinct memory devices; and that have been programmed unimaginatively so that erasure is needed; and that use the specific memory erasure procedure LP5 of

[20]If this assumption is denied, then all computation will become thermodynamically enormously costly, in contradiction with the Landauer–Bennett tradition that ascribes an unavoidable entropy cost only to erasure. Or one might be tempted to allow that each switching or other program step is admissible if considered in isolation; but to insist that there is some sort of global constraint that precludes us combining them freely into any program we choose. Such a global constraint would incapacitate the present literature in the Landauer–Bennett tradition, in which individual program steps are routinely combined without recourse to global considerations.

[21]A referee has drawn my attention to an objection to the no-erasure demon by Maroney (2002, Section 8.3.3) that in turn draws on Shenker (1999). The objection is that branching of control to program-L or program-R requires that there be two global time evolutions for the system and that the system picks between them on the basis on its own state, a kind of "free will"; and that this in some way leads to a many-to-one mapping of states in time evolution that is incompatible with a Hamiltonian flow or, in the quantum mechanical case, with unitary time evolution. The scope of the objection is far greater than just the no-erasure demon. It also applies to the Szilard one-molecule gas engine, for that device also branches control depending on the location discovered for the molecule. It applies to any computational device in which the flow of control branches, thereby precluding most interesting computation tasks from being implemented on a device governed by Hamiltonian dynamics. I have been unable to see the cogency of the objection. The branching of the flow of control can be given a quite benign description with no need for "free will" or an obvious incompatibility with Hamiltonian dynamics: the system is governed by a single Hamiltonian and the sequence of states it generates is different according to whether the initial state is in one part of the phase space or in another. The concern with many-to-one mappings seems hasty. While a Hamiltonian flow maps total states one to one under time evolution, there need be no such one-to-one mapping for only a few degrees of freedom of the total system. That a computer's memory device is in state L or R specifies only a few degrees of freedom of the total system, for these memory devices are presumed to be in thermal equilibrium with a much larger environment. A great deal more would be needed to establish an incompatibility between the Hamiltonian dynamics of the total system and a many-to-one mapping of these degrees of freedom.

Section 2.3 (and not even all of these are exorcised—recall the economically programmed demon mentioned above).

5. Bennett's extension of Landauer's principle

The no-erasure demon described immediately above is immune to exorcism by Landauer's principle simply because it performs no erasures. Bennett (2003) has proposed in response, however, that it does succumb to a more general version of Landauer's principle. That extended version assigns an entropy cost not just to erasure but to the sort of merging of computational paths as arises at the end of the no-erasure demon's operation, when program-R switches the memory device back from R to L. The new principle reads (p. 501):

> Landauer's principle holds…that any logically irreversible manipulation of information, such as the erasure of a bit or the merging of two computational paths, must be accompanied by a corresponding entropy increase in non-information-bearing degrees of freedom of the information-processing apparatus or its environment.

The principle is supported by the many-to-one mapping argument as quoted in Section 3.3. The argument is taken to apply equally to the erasure of memory devices as to the merging of computational paths.

To make a cogent assessment of the extension, we should ask the same questions of it as asked in Section 2.3 of the original principle. What *precisely* does it assert? And why *precisely* should we believe it? Most urgently, just what constitutes a "computational path"? In the absence of precise answers, we might note that the many-to-one mapping argument failed to force a thermodynamic cost for information erasure in memory devices. Why should we expect it to fare better when it comes to the merging of computational paths? Indeed, we should expect it to fare worse. In ordinary erasure processes, such as described in Section 2.3, the memory devices pass through an intermediate state in which their phase spaces are expanded, so that a compression of the phase space ensues. When computational paths merge in computers, however, there seems to be no corresponding intermediate state. This merging is quite distant from a compression of the phase space as described in Section 2.4.

Let us take the specific example of the no-erasure demon that is the extended principle's target. We shall see quite quickly that no consideration advanced so far gives any expectation that an entropy cost must be associated with the merging of its computational paths. In particular, the merging of its computational paths is not associated with a compression of the phase space such as arises in LP5b.

Our no-erasure demon executes two programs tracked by a single memory device. If, for convenience, we imagine the memory device to be a molecule trapped in disjoint regions L and R of a chamber by two pistons, we can track at least this portion of the computational paths. When program-L is run, the molecule stays in the L region throughout. When program-R is run, the memory device is switched to

read R. To do this, the molecule is slowly moved over to the R region in a thermodynamically reversible, constant volume process (as shown in Fig. 1). The same thermodynamically reversible, constant volume process is used to switch the setting back to L at the end of the cycle by program-R. In all these processes, the entropy of the memory cell and the volume of phase space accessible to it, remain constant. Successive stages of this switching are shown in Fig. 8.

There is no thermodynamic entropy cost created by these processes. Indeed, there is not even a thermodynamically reversible transfer of entropy between the device and its environment. These processes are quite distinct from another process, labeled "expansion" and "compression" in Fig. 8, in which the volume of phase space accessible to the memory device expands and contracts.

Bennett presumably intends the operation of our no-erasure demon to be of this latter type of process, for it is through a two-fold reversible compression of the phase space that the $k \ln 2$ of thermodynamic entropy comes to be passed to the environment. How might we come to conceive the operation of the no-erasure demon in this way? We would need to sample its processes at different temporal stages and then illicitly collapse them into one process of expansion and compression. That is, the distributions associated with the two states at time 1 and time 2 in Fig. 8 would be combined to form a distribution that covers double the phase space volume. Just as before, the combination is illicit if it is intended to form a canonical ensemble. The energy functions $E(x)$ of the phase spaces at time 1 and time 2 are different. The first is finite only in the L region of the phase space; the second is finite only in the R region. Whatever is produced by the assembly is not a canonical ensemble that properly represents the thermodynamic properties of the two states. Its compression is not the compression of the volume of the accessible region of phase space of canonical ensemble as described in Section 2.4. So we cannot apply formulae like (8) for the entropy of a canonical distribution and infer that its thermodynamic entropy is $k \ln 2$ greater than the default L state.

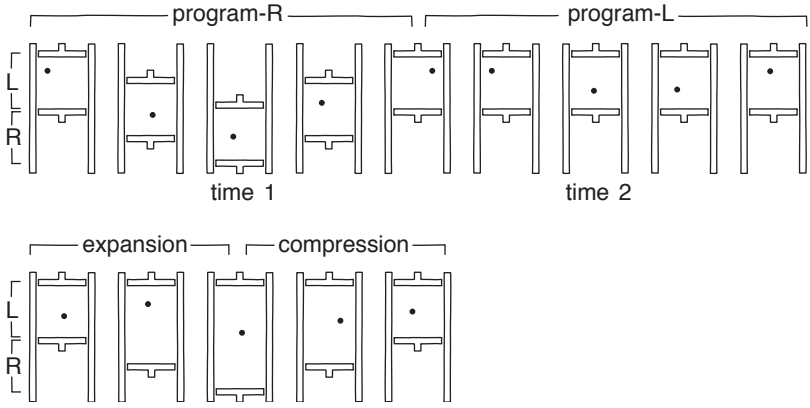

Fig. 8. Time development of the no-erasure demon's memory device and a different device that expands and contracts the phase space.

Finally, it should also be noted that, even if the process were equivalent to such an expansion and compression of the phase space, we would still have not have established that the merging of computational flows has generated thermodynamic entropy of $k \ln 2$ in the environment. If both expansion and compression were effected as thermodynamically reversible processes, any thermodynamic entropy passed to the environment by the compression would be balanced exactly by entropy drawn from the environment in the expansion phase.

This analysis yields no cogent grounds for associating an assured thermodynamic entropy cost with the merging of computational paths of our no-erasure demon. Thus, we have found no cogent grounds for believing the extended version of Landauer's principle when applied to this case and thus no grounds for believing it as a general principle.

6. Conclusion

It is hard to be optimistic about the literature on Landauer's principle and its use in exorcising Maxwell's demon. The fundamental weaknesses of the literature are at the methodological level. While there is talk of a general principle—the entropy cost necessarily incurred by erasure and the merging of computational paths—no general principle is stated precisely. Instead we are given illustrations and are somehow to intuit the generality from them. There is talk of reasons and justification. But we are only given imprecise hints as to what they might be. Erasure is a many-to-one mapping of physical states that somehow corresponds to a compression of the phase space that in turn somehow incurs an entropy cost. Yet the attempt to make precise the association of the many-to-one mapping with a compression of a phase space founders at least upon the illicit assembly of a canonical ensemble. This literature is too fragile and too tied to a few specific examples to sustain claims of the power and generality of the failure of all Maxwell demons.

Must the erasure of a bit of information be accompanied inevitably by a generation of $k \ln 2$ of thermodynamic entropy in the environment? As far as I know, the question remains open. While there may be cogent thermodynamic or statistical mechanical grounds for an inevitable entropy cost, the literature on Landauer's principle has yet to present them. Must all Maxwell's demons fail in their efforts to reverse the second law of thermodynamics? As far as I know, the question remains open. It does appear to me, however, that the success of any particular demon would probably be precluded if we were to accumulate enough realistic constraints on its particular operation—although the experimental successes of "Brownian ratchets" must cast some doubt into the minds of even the most ardent opponents of the demon. What is less sure is whether the particular realistic constraints that end up defeating particular demons could be assembled into a compactly expressible, general argument that would assure us in advance of the failure of all demons. This much is sure. If ever a successful exorcism of Maxwell's demon emerges, it will require an analysis more sophisticated and precise than those presently at hand.

Acknowledgements

I am grateful to Alexander Afriat, Jeffrey Bub, Ari Duwell, John Earman, Robert Rynasiewicz, and two anonymous referees for helpful discussion; and to Harvey Leff for stimulating and informative discussion of Bennett (2003) and the present paper; and to Jos Uffink for his insightful and diligent guidance both as reader of this paper and editor of this journal.

Appendix A. The telescoped distribution

Assume that we have before us one of the memory devices described in Section 2.3. It is either in state L_T or in state R_T. But since the device carries random data, we do not know which. We decide that there is probability $\frac{1}{2}$ of each. The states L_T and R_T may be described individually by the two canonical distributions

$$p_L(x) = \exp(-E_L(x)/kT)/Z_L \quad \text{and} \quad p_R(x) = \exp(-E_R(x)/kT)/Z_R,$$

where the subscripts L and R on the partition function Z (3) indicate that they are evaluated using energy functions $E_L(x)$ and $E_R(x)$. What of the canonical distribution associated with the state $(L+R)_T$ that arises when we remove the partition separating L and R? It is

$$p_{L+R}(x) = \exp(-E_{L+R}(x)/kT)/Z_{L+R}.$$

Might this not *also* represent the memory device carrying random data as described above? It is produced by simply telescoping down the two distributions $p_L(x)$ and $p_R(x)$ into one phase space in a process that is characterized in Section 3.1 as the illicit formation of a canonical ensemble. Nonetheless, it will correctly represent the probability that the device would be found to be in the state x were we to be able to check its state. Is that not good enough?

The difficulty is that the formula for $p_{L+R}(x)$ is not a complete presentation of the probabilistic properties of the device relevant to its thermodynamic properties. What is missing is correlation information. If the system is $(L+R)_T$ and we could somehow know that it happened to be some state x in L at a particular time, then (allowing some time to elapse) the distribution $p_{L+R}(x)$ would still tell us the probability of the system being in either region L or R of the phase space. If, on the other hand, the device is storing random data, this would no longer be so. If, at one time, we know the device happened to be in some state x in L, then we know it will remain in L for all time (assuming no disturbance). This correlation information is not represented in the expression for $p_{L+R}(x)$.

The effect of this difference is that expression (8) for the thermodynamic entropy of a canonically distributed system can be properly applied to the device in state $(L+R)_T$, but it cannot be properly applied to the device if it is carrying random data. To see why it fails in the latter case, let us reconstruct the derivation of (8) as given in Section 2.2, now applied specifically to the device with random data. We shall see that properly accommodating the probabilistic dependencies of the device

with random data leads to the conclusion that it carries the smaller quantity of thermodynamic entropy, $S_L = S_R$.

To begin, note that the mean energy of the memory device with random data, in the context of the thermodynamically reversible process of Section 2.2, can be expressed as

$$\bar{E}_{L+R} = \int_{L+R} E_{L+R}(x,\lambda)\, p_{L+R}(x,t)\, \mathrm{d}x. \tag{A.1}$$

We might proceed as in Section 2.2 to write down the rate of change of the mean energy as

$$\frac{\mathrm{d}\bar{E}}{\mathrm{d}t} = \int_{L+R} E_{L+R}(x,\lambda)\, \frac{\mathrm{d}p_{L+R}(x,t)}{\mathrm{d}t}\, \mathrm{d}x + \int_{L+R} \frac{\mathrm{d}E_{L+R}(x,\lambda)}{\mathrm{d}t}\, p_{L+R}(x,t)\, \mathrm{d}x. \tag{5′}$$

However, because of the probabilistic dependencies, the second term no longer represents the rate at which work is done on the system:

$$\frac{\mathrm{d}W}{\mathrm{d}t} \neq \int_{L+R} \frac{\mathrm{d}E_{L+R}(x,\lambda)}{\mathrm{d}t}\, p_{L+R}(x,t)\, \mathrm{d}x. \tag{A.2}$$

At this point the standard derivation of (8) is blocked.

The difficulty in applying (A.2) is that if the memory device actually happens to be in state L_T, then the portion of the integration over region R is not relevant; and conversely for state R_T. For example, if the device is in state L_T and the process parameterized by λ alters the energy function in region R only, then the process will supply no work to the system, whereas the integral in (A.2) indicates that the process will supply work. Conversely, if the device were in state R_T, the integral of (A.2) would underestimate the rate at which the work is supplied by a factor of 2 (since $p_{L+R}(x) = p_R(x)/2$ in region R).

If the classical Clausius formula (1) is to be used to assign a thermodynamic entropy to the memory device, we must find some way to represent a thermodynamically reversible process that the system may undergo. Such a process was represented in Section 2.2 by the operator $\mathrm{d}/\mathrm{d}t$, where this operator gave the rate of change of quantities when the temperature T and parameter λ were changed sufficiently slowly for the system to remain canonically distributed. The above discussion shows that two distinct operators must be used according to whether the device state is L_T or R_T. In the first case, the operator $\mathrm{d}/\mathrm{d}t_L$ would represent processes in which work is performed on the state L_T by manipulation of the energy function $E_L(x)$ and heat is passed to a device in state L_T. The operator $\mathrm{d}/\mathrm{d}t_R$ would represent the corresponding process if the state were R_T. In each case, the derivation would proceed as in Section 2.3 to conclude that the thermodynamic entropy of the device is S_L if it is in state L_T; and S_R if it is in state R_T, where

$$S = S_L = \frac{\bar{E}_L}{T} + k \ln \int_L \exp(-E_L/kT)\, \mathrm{d}x = \frac{\bar{E}_R}{T} + k \ln \int_R \exp(-E_R/kT)\, \mathrm{d}x$$
$$= S_R.$$

The device is assuredly either in state T_L or in state T_R; in either case the entropy is $S = S_L = S_R$.

Insofar as the memory device carrying random data can be said to be in a single thermodynamic state (as opposed to being in one of the two, we know not which), then its thermodynamic entropy is just $S = S_L = S_R$. While a common wisdom may be that we have to add $k \ln 2$ to the entropy to accommodate our uncertainty over the states, the above analysis reveals no grounds for characterizing such a term as thermodynamic entropy conforming to the Clausius formula (1). The additional term does arise if we substitute the probability distribution $p_{L+R}(x)$ into expression (8) for thermodynamic entropy. Yet we have just shown above that the derivation of this expression fails if the probability distribution $p_{L+R}(x)$ pertains to a memory device with random data.

References

Bennett, C. H. (1982). The thermodynamics of computation—a review. *International Journal of Theoretical Physics*, *21*, 905–940 reprinted in Leff and Rex (2003, Chapter 7.1).

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

Bub, J. (2001). Maxwell's demon and the thermodynamics of computation. *Studies in History and Philosophy of Modern Physics*, *32*, 569–579.

Earman, J., & Norton, J. D. (1998). Exorcist XIV: the wrath of Maxwell's demon; Part I: from Maxwell to Szilard. *Studies in History and Philosophy of Modern Physics*, *29*, 435–471.

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

Landauer, R. (1961). Irreversibility and heat generation in the computing process. *IBM Journal of Research and Development*, *5*, 183–191 reprinted in Leff and Rex (2003, Chapter 4.1).

Leff, H. S., & Rex, A. (2003). *Maxwell's demon 2: entropy, classical and quantum information, computing*. Bristol and Philadelphia: Institute of Physics Publishing.

Maroney, O. J. E. (2002). *Information and entropy in quantum theory*. Ph.D. dissertation, Birkbeck College, University of London.

Piechocinska, B. (2000). Information erasure. *Physical Review A*, *61*, 62314, 1–9; reprinted in Leff and Rex (2003, Chapter 4.4).

Planck, M. (1926). *Treatise on thermodynamics*. New York: Longman; reprinted New York: Dover.

Shenker, O. (1999). Maxwell's demon and Baron Munchausen: free will as a perpetuum mobile. *Studies in History and Philosophy of Modern Physics*, *30*, 347–372.

Shenker, O. (2000). *Logic and entropy*. Preprint: philsci–archive.pitt.edu/archive/00000115.

Shizume, K. (1995). Heat generation required by information erasure. *Physical Review E*, *52*, 3495–3499 reprinted in Leff and Rex (2003, Chapter 4.3).

Thomson, C. J. (1972). *Mathematical statistical mechanics*. Princeton: Princeton University Press.

Thomson, W. (1853). On the dynamical theory of heat, with numerical results deduced from Mr. Joule's equivalent of a thermal unit, and M. Regnault's observations on steam. *Philosophical Magazine, Series 4*, *VI*, 8–21.

Uffink, J. B. M. (2001). Bluff your way into the second law of thermodynamics. *Studies in History and Philosophy of Modern Physics*, *32*, 305–394.