The Connection between Logical and Thermodynamical Irreversibility

Tony Short, James Ladyman, Berry Groisman, Stuart Presnell
Departments of Physics and Philosophy, University of Bristol

July 22, 2005

Abstract

There has recently been a good deal of controversy about Landauer's Principle, which is often stated as follows: The erasure of one bit of information in a computational device is necessarily accompanied by a generation of $kT \ln 2$ heat. This is often generalised to the claim that any logically irreversible operation cannot be implemented in a thermodynamically reversible way. John Norton (2005) and Owen Maroney (2005) both argue that Landauer's Principle has not been shown to hold in general, and Maroney offers a method that he claims instantiates the operation reset in a thermodynamically reversible way.

In this paper we defend the qualitative form of Landauer's Principle, and clarify its quantitative consequences (assuming the second law of thermodynamics). We analyse in detail what it means for a physical system to implement a logical transformation L, and we make this precise by defining the notion of an *L-machine.* Then we show that logical irreversibility of L implies thermodynamic irreversibility of every corresponding L-machine. We do this in two ways. First, by assuming the phenomenological validity of the Kelvin statement of the second law, and second, by using information-theoretic reasoning. We illustrate our results with the example of the logical transformation 'reset', and thereby recover the quantitative form of Landauer's Principle.

1 Introduction

Computation can be theoretically described as a formal process of data manipulation when the only concern is with its logical form. However, any computation that is actually carried out is realised by some kind of physical process, and as such is subject to physical laws and the laws of thermodynamics in particular. It is therefore important to consider what connections there might be between the logical properties of computations and the thermodynamical properties of systems that realise them. The last few decades have witnessed a good deal of controversy about this matter following the seminal work of Rolf Landauer (1961). His main conclusion, which has subsequently become known as Landauer's Principle, is that the erasure of information in some computational device is necessarily accompanied by an appropriate increase in the thermodynamic entropy of the device and/or its environment. This result is often generalised as follows: (a) any logically irreversible process must result in an entropy increase in the non-information bearing degrees of freedom of the information-processing system or its environment; (b) any logically reversible process can be implemented thermodynamically reversibly (see for example Charles Bennett 2003). These two statements are believed to provide the sought after link between thermodynamics and computation. Recently however, these conclusions have been contested and today, some time after Landauer's paper, the implications of Landauer's arguments for the general connection between logic and thermodynamics is the subject of much debate. In particular, John Norton (2005) and Owen Maroney (2005) both argue that Landauer's Principle has not been shown to hold in general, and Maroney offers a method that he claims instantiates the operation reset in a thermodynamically reversible way.

In this paper we defend Landauer's Principle, however, we do not regard it as more fundamental than the second law of thermodynamics, and so we do not follow those authors who try to show that Landauer's Principle implies the impossibility of a Maxwell Demon. Rather, we assume the second law and show that Landauer's Principle follows. Hence, we follow the 'sound' rather than the 'profound' horn of the dilemma that John Earman and Norton (1998 and 1999) identified.

We believe that part of the reason for the above controversy is the fact that many important notions in this domain are loosely defined and not properly understood. We begin by constructing a clearly defined set of concepts for the conduct of reasoning about these issues, and we go on to establish the connection between logical and thermodynamic irreversibility via a general argument for a generalisation of Landauer's Principle. In particular, we will offer precise definitions of logical irreversibility and thermodynamic irreversibility, before analysing in detail what it means for a physical system to implement a logical transformation. To this end we introduce the notion of an *L-machine.* This is a hybrid physical-logical concept that combines a physical device, a specification of which physical states of that device correspond to various logical states, and an evolution of that device which corresponds to the logical transformation L. Then we address the question of whether

Possibilities	Thermodynamically reversible	Thermodynamically irreversible
Logically reversible	✓	✓
Logically irreversible	?	✓

Table 1: A table representing the different possibilities for logical and thermodynamic reversibility. Our paper addresses the issue of whether any logically irreversible transformation can be implemented thermodynamically reversibly.

the logical irreversibility of L implies thermodynamic irreversibility of every corresponding L-machine, and show that the answer is positive. This is our restatement and generalisation of Landauer's Principle.

Summarising the current state of the debate, we take it that everyone agrees that there are both logically reversible and irreversible transformations, and that every logically reversible transformation is implementable in a thermodynamically reversible way, and that any such transformation can also be implemented in a thermodynamically irreversible way. Everyone also agrees that a logically irreversible transformation can be implemented in a thermodynamically irreversible way. So the issue is whether there are any logically irreversible transformations that can be implemented in a thermodynamically reversible way (as illustrated in table 1). The conclusion of our present paper is that this is impossible. Thus we establish a complete link between logical and thermodynamic irreversibility.

The structure of this paper is as follows. In sections 2 and 3 we explain what we understand by logical and thermodynamic irreversibility respectively, and emphasise the importance of clearly distinguishing between logical and physical concepts, and between individual processes and families of processes. In section 4, we introduce the notion of an L-machine. In section 5, in order to make connection with concrete observable phenomena, we demonstrate that violation of Landauer's Principle leads to violation of the phenomenological validity of the Kelvin statement of the second law. In section 6, we offer a general and precise argument for Landauer's Principle using information-theoretic reasoning. In section 7, we illustrate our results with the example of the logical transformation 'reset', and argue that all logically irreversible processes can be thought of as combinations of logically reversible transformations and one or more reset transformations. In section 8, we clear up some confusions associated with the notions of known and unknown data. Finally, in section 9, we consider an individual process which reveals the irreversibility of an L-machine. In Appendix A we prove a result about bounds stated in section 5, and in Appendix B we give a justification for formulating the second law in terms of information theoretic entropies.

2 Logical Irreversibility

First we note that a logical transformation is a *mathematical* operation, consisting of a mapping L from a finite set X of input states, into a finite set Y of output states, where each input state is mapped to a unique output state. For example, consider the case of binary-valued logic, in which the input and output states are bit-strings (with 0 and 1 usually representing 'false' and 'true' respectively); the mapping L can be represented by a truth table, or as a digital circuit constructed from some set of universal gates (e.g. NAND and COPY). We say that a logical transformation is *logically reversible* if and only if $L : X \rightarrow Y$ is a one-to-one (injective) mapping. Hence with a reversible logical transformation, we can uniquely reconstruct the input state from the output state. If L is not a one-one mapping, we say that it is *logically irreversible.*

3 Thermodynamic irreversibility

Thermodynamic irreversibility, on the other hand is a feature of *physical* processes, described by the second law of thermodynamics. There is much controversy about the correct formulation of the latter, and about how it can be justified on the basis of statistical mechanics. We will not address these issues, but rather we will assume that the second law is valid.

It is crucial for our argument that we make a distinction between a logical transformation, which is a map from a *set* of logical states to a *set* of logical states, and a physical process, which is a change in a physical system whereby it goes from a *particular* physical state to a *particular* physical state. It follows that it makes no sense to talk of the implementation of a logical transformation by a physical process, rather in so far as logical transformations are implemented using physical systems, they are implemented by a family of processes. For the physical system to implement the logical transformation reliably, the family of processes must take each of the physical states that represent the logical input states to the appropriate physical state, that is the one that represents the right logical output state (Our point here is clearly illustrated in the case of a truth table, where each member of the family of processes corresponds to a single row). We will say that a family of physical processes is thermodynamically irreversible if and only if at least one of its members is. This is the rationale behind our definition of an L-machine in the next section. We raise this issue here since we now discuss the thermodynamics of a physical process.

Consider a system in a heat reservoir at temperature T undergoing some thermodynamic process p. If $\Delta S_{sys}(p)$ is the change in the entropy of the system during the process p, and $\Delta Q(p)$ is the heat flow from the system into the reservoir during the same process, then the second law requires that

$$\forall\, p, \quad \Delta S_{sys}(p) + \frac{\Delta Q(p)}{T} \geq 0 \tag{1}$$

Identifying $\Delta S_{res}(p) = \Delta Q(p)/T$ as the entropy change of the heat reservoir, we define

$$\Delta S_{tot}(p) = \Delta S_{sys}(p) + \Delta S_{res}(p) \tag{2}$$

as the total entropy change of the system and reservoir together. The second law can then be restated in the familiar form

$$\forall p, \quad \Delta S_{tot}(p) \geq 0 \tag{3}$$

i.e. total entropy is non-decreasing over time.

A process p is *thermodynamically reversible* if and only if $\Delta S_{tot}(p) = 0$.

If $\Delta S_{tot}(p) > 0$, the physical process p cannot be run in reverse, as the reverse process p' would have $\Delta S_{tot}(p') < 0$, and hence violate the second law. We therefore refer to any process p for which $\Delta S_{tot}(p) > 0$ as *thermodynamically irreversible*. As is well known, there are a number of formulations of the second law that are provably equivalent to this, modulo certain assumptions.

4 Implementing a logical transformation with a physical device

In order to analyze the connection between *logical* transformations, and *physical* thermodynamic processes, we must consider what it means for a physical system to implement a logical transformation. As we said above, a physical system can only implement a logical transformation through a family of processes. To physically implement a logical transformation, we require: A physical device, a specification of which physical states of that device correspond to the possible logical states (we call the former *representative states*), and a time evolution operator of that device. The time evolution operator must generate the relevant family of processes, and the reliability of the implementation consists in the time evolution operator being such as to ensure that *whichever* of the representative physical states the device is prepared in, it ends up in the appropriate representative state. This insistence on generality is an important difference between our approach and that of Maroney (2005) who considers only individual processes.

We refer to this combined system as an *L-machine*. Note that L names a particular logical transformation, so we have L_{AND}-machines, and so on.

Formally, we define an L-machine as an ordered set

$$\left\{D, \left\{D_{in}(x) | x \in X\right\}, \left\{D_{out}(y) | y \in Y\right\}, \Lambda_L\right\} \tag{4}$$

consisting of

- A physical *device* D, situated in a heat bath at temperature T.
- A set $\{D_{in}(x)|x \in X\}$ of macroscopic input states of the device, which are distinct thermodynamic states of the system (i.e. no microstate is a component of more than one thermodynamic state). $D_{in}(x)$ is the representative physical state of the logical input state x.
- A set $\{D_{out}(y)|y \in Y\}$ of distinct thermodynamic output states of the device. $D_{out}(y)$ is the representative physical state of the logical output state y. Note that the set of input states and output states may overlap.
- A time-evolution operator Λ_L for the device, such that $\forall\, x \in X, \Lambda_L(D_{in}(x)) = D_{out}(L(x))$.

An L-machine $\{D, \{D_{in}(x)|x \in X\}, \{D_{out}(y)|y \in Y\}, \Lambda_L\}$ physically implements L in the following sense. If D is prepared in the input state $D_{in}(x)$ corresponding to the logical input state $x \in X$, and is then evolved using Λ_L, it will be left in the output state $D_{out}(y)$ corresponding to the logical output state $y = L(x) \in Y$. We will denote this physical process by p_x.

Consider the thermodynamics of the process p_x. If the entropy of the system in the state $D_{in}(x)$ is $S_{in}(x)$, the entropy of the system in state $D_{out}(y)$ is $S_{out}(y)$, and the heat flow from the system into the reservoir during the process is $\Delta Q(p_x) = T\Delta S_{res}(p_x)$, the total entropy change $\Delta S_{tot}(p_x)$ for the process will be given by

$$\Delta S_{tot}(p_x) = S_{out}(L(x)) - S_{in}(x) + \frac{\Delta Q(p_x)}{T} \geq 0, \tag{5}$$

This particular process will be thermodynamically reversible if $\Delta S_{tot}(p_x) = 0$. Note that in the commonly considered case in which the initial and final entropies of the system are the same, ΔS_{tot} is proportional to the heat flow from the system into the reservoir. Furthermore if the initial and final energies of the system are the same as well, then from the first law of thermodynamics, this heat flow is equal to the work done on the system.

We say that the L-machine is *thermodynamically reversible* if and only if $\forall\, x \in X, \Delta S_{tot}(p_x) = 0$ (i.e. if all of the processes p_x are thermodynamically reversible). An L-machine is therefore *thermodynamically irreversible* if there exists an $x \in X$ for which $\Delta S_{tot}(p_x) > 0$.

5 Implementing a logically irreversible transformation

Application of the statement of the second law in terms of entropy is controversial for many reasons. In particular, because of concerns about the status of the thermodynamic entropy as an objective property of physical systems given its connection to concepts such as uncertainty, probability and so on, and furthermore, because such applications often proceed via information theoretic definitions of entropy (for systems not in thermal equilibrium). In this section we offer a proof of our main result which appeals only to the more concrete statement of the second law of thermodynamics which is usually referred to as the Kelvin formulation:

> "It is impossible to perform a cyclic process with no other result than that heat is absorbed from a reservoir, and work is performed." (Uffink 2001, p. 328)

Consider an L-machine which implements an irreversible logical transformation L. As L is logically irreversible, it is not a 1-1 mapping. It is therefore possible to select two logical input states $(x_1, x_2) \in X$ which map onto the same logical output state $y \in Y$ (such that $L(x_1) = L(x_2) = y$).

Now consider a composite system S_{DB}, consisting of the device D (prepared in the output state $D_{out}(y)$) and a box B containing a single gas molecule trapped between two pistons, in a heat reservoir at temperature T.

We investigate the average heat flow from the composite system S_{DB} into the heat reservoir during the following 4-stage cyclic process P^{DB} (note that whenever heat flows from the reservoir ΔQ is negative):

1. Insert a partition into the box B (parallel to the two pistons), such that there is a probability q_1 $(0 < q_1 < 1)$ of the particle being confined to the left of the partition, and a probability $q_2 = (1 - q_1)$ of it being confined to the right of the partition.

2. Perform a controlled operation on the device D, which depends on the position of the particle in the box B, such that

   (a) If the particle is to the *left* of the partition, the device is transformed thermodynamically reversibly from the state $D_{out}(y)$ to the state $D_{in}(x_1)$, causing a heat flow $\Delta Q_{D_1} = T(S_{out}(y) - S_{in}(x_1))$ into the heat reservoir.

   (b) If the particle is to the *right* of the partition, the device is transformed thermodynamically reversibly from the state $D_{out}(y)$ to the state $D_{in}(x_2)$, causing a heat flow $\Delta Q_{D_2} = T(S_{out}(y) - S_{in}(x_2))$ into the heat reservoir.

3. Perform a controlled operation on the box B, which depends on the state of the device D, such that

(a) If the device is in state $D_{in}(x_1)$, the *right*-hand piston is moved up to the partition, then the partition is removed, and the gas particle is allowed to expand isothermally against the piston (performing work $-kT \ln q_1$, and causing heat to flow from the reservoir so that $\Delta Q_{B_1} = kT \ln q_1$) until the piston is returned to its initial position.

(b) If the device is in state $D_{in}(x_2)$, the *left*-hand piston is moved up to the partition, then the partition is removed, and the gas particle is allowed to expand isothermally against the piston (performing work $-kT \ln q_2$, and causing heat to flow from the reservoir so that $\Delta Q_{B_2} = kT \ln q_2$) until the piston is returned to its initial position.

4. Perform the evolution Λ_L on the device D, such that

(a) If the device is in the state $D_{in}(x_1)$, the process p_{x_1} occurs and transforms the state of the device into $D_{out}(y)$, causing a heat flow $\Delta Q(p_{x_1})$ into the reservoir.

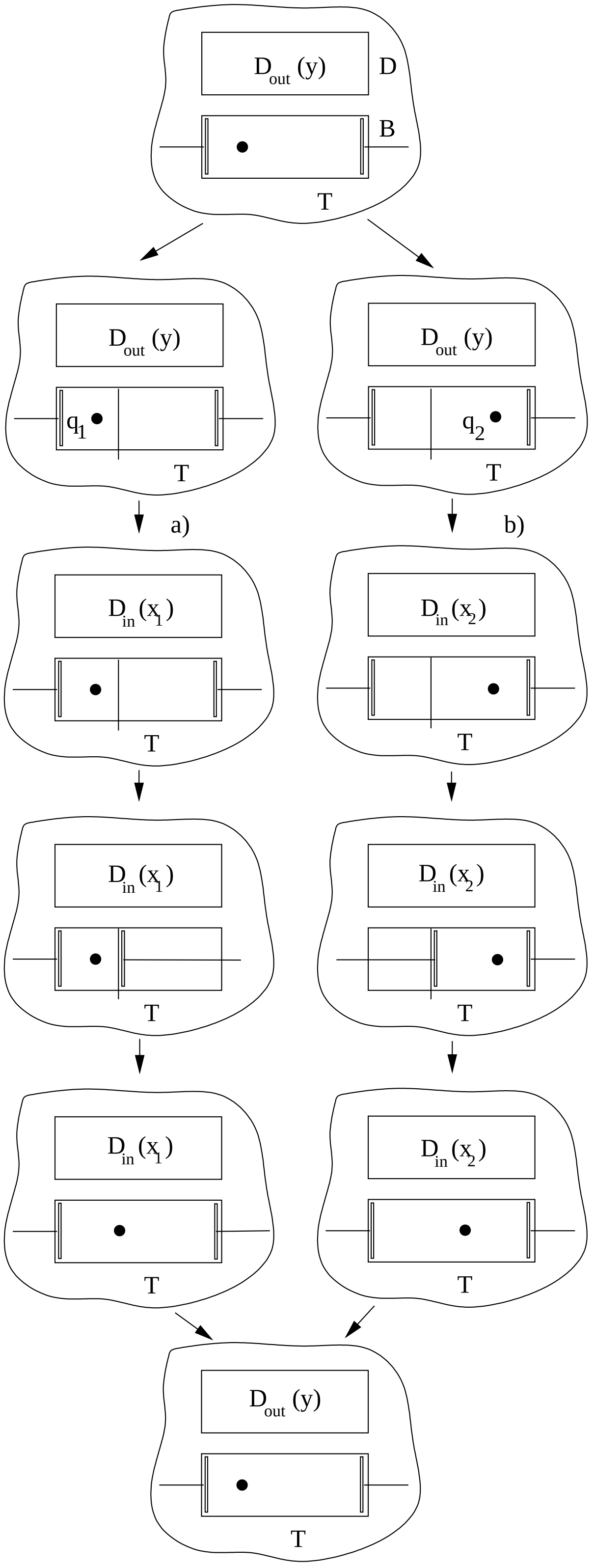
(b) If the device is in the state $D_{in}(x_2)$, the process p_{x_2} occurs and transforms the state of the device into $D_{out}(y)$, causing a heat flow $\Delta Q(p_{x_2})$ into the reservoir.

Note that the device D is analogous to the memory of the familiar kind of Maxwell's demon inspired by the engine described in Szilard (1929) and analysed in Bennett (1987): step 2 corresponds to the demon measuring the position of the particle; step 3 is the demon extracting work from the gas; and step 4 is the demon resetting its memory.

After this process, the combined system S_{DB} will have returned to its initial state. The average heat flow into the heat reservoir during the process p^{DB} is given by

$$\begin{aligned} \langle \Delta Q(p^{DB}) \rangle &= \sum_{n=1,2} q_n(\Delta Q_{D_n} + \Delta Q_{B_n} + \Delta Q(p_{x_n})) && (6) \\ &= \sum_{n=1,2} q_n(T(S_{out}(y) - S_{in}(x_n)) + kT \ln q_n + \Delta Q(p_{x_n})) && (7) \\ &= \sum_{n=1,2} q_n(T\Delta S_{tot}(p_{x_n}) + kT \ln q_n) && (8) \end{aligned}$$

As the process is cyclic, the internal energy of the initial and final states of the entire system is the same. Hence, from the first law of thermodynamics, if the heat flow is negative (i.e. heat flows from the reservoir into the system), then that heat must have been converted into work by the system. Clearly, by performing this cycle many times we could obtain a well-defined heat flow



Step 1: Inserting a partition.

Step 2: Performing a controlled operation on. the device D.

Heat flow takes place

Step 3a: Performing a controlled operation on the box B − moving the piston up to the partition.

Step 3b: Performing a controlled operation on the box B − removing the parition and isothermal expantion against the piston.

Heat flow takes place

Step 4: Performing the evolution Λ_L on the device D.

Heat flow takes place

Figure 1: The evolution of the system S_{DB}

proportional to the average heat flow in a single cycle. In order to satisfy the Kelvin statement of the second law above we must therefore have:

$$\langle \Delta Q(p^{DB}) \rangle \geq 0. \tag{9}$$

and hence

$$\sum_{n=1,2} q_n \Delta S_{tot}(p_{x_n}) \geq -k \sum_{n=1,2} (q_n \ln q_n) > 0. \tag{10}$$

As $\sum_{n=1,2} q_n \Delta S_{tot}(p_{x_n}) > 0$, one or both of $\Delta S_{tot}(p_{x_1})$ or $\Delta S_{tot}(p_{x_2})$ must be greater than 0. Following the definition in the previous section, this proves:

> **Theorem**: If L is logically irreversible, then every L-machine is thermodynamically irreversible.

In fact, by considering specific initial states, it is possible to prove that both $\Delta S_{tot}(p_{x_1}) > 0$ and $\Delta S_{tot}(p_{x_2}) > 0$.

For any non-zero value of $\Delta S_{tot}(p_{x_1})$, consider the case in which $q_1 = \exp(-\Delta S_{tot}(p_{x_1})/k)$. Inserting this into the above equation we find that

$$\Delta S_{tot}(p_{x_2}) \geq -k \ln(1 - q_1) > 0. \tag{11}$$

Hence if $\Delta S_{tot}(p_{x_1}) > 0$, it must also be true that $\Delta S_{tot}(p_{x_2}) > 0$. Similarly, we can show that if $\Delta S_{tot}(p_{x_2}) > 0$ then $\Delta S_{tot}(p_{x_1}) > 0$. As one of the entropy changes *must* be greater than zero in order to satisfy (10), the other entropy change must also be greater than zero.

The process p_x of operating the L-machine with input state $D_{in}(x)$ will therefore be thermodynamically irreversible for each logical input state x which cannot be deduced unambiguously from its logical output state $y = L(x)$.

Furthermore, if the total entropy change for each process is bounded above by a large but finite amount ΔS_{tot}^{max} (as it must be in any realistic scenario), then the minimum entropy change of each irreversible process is bounded below by a corresponding small but finite amount ΔS_{tot}^{min}, where,

$$\Delta S_{tot}^{min} = -k \ln \left(1 - \exp\left(\frac{-\Delta S_{tot}^{max}}{k} \right) \right) \tag{12}$$

We show in appendix A that this is the tightest lower bound. Note that as ΔS_{tot}^{max} decreases, ΔS_{tot}^{min} increases. The lowest possible value for ΔS_{tot}^{max} is therefore achieved when $\Delta S_{tot}^{max} = \Delta S_{tot}^{min}$ (because $\Delta S_{tot}^{max} < \Delta S_{tot}^{min}$ cannot be satisfied). From equation (12) this occurs when $\Delta S_{tot}^{max} = \Delta S_{tot}^{min} = k \ln 2$. Note also that when $q_1 = q_2 = 1/2$, equation (10) implies that $(\Delta S_{tot}(p_{x_1}) + \Delta S_{tot}(p_{x_2}))/2 \geq k \ln 2$. In the standard case in which the entropies of the input and output representative states are equal, $\Delta S_{tot}(p_{x_n}) = \Delta Q(p_{x_n})/T$. We therefore make contact with the standard quantitative formulation of Landauer's Principle according to which resetting one bit of information requires an average heat flow of at least $kT \ln 2$. We must emphasise that for an individual process (for example, the process whereby a '1' is reset to a '0'), the heat flow may be less than $kT \ln 2$, and thus it is only the average heat flow that satisfies the quantitative form of Landauer's Principle (and then only when $q_1 = q_2 = 1/2$). (Note however, that equation (10) is completely general.)

The above theorem depends on the application of thermodynamics to a family of processes rather than just to one or another of them. It is trivially true that any *individual* process in the family of processes that implement a logical transformation can be carried out in a thermodynamically reversible way. To see this, consider the transformation of an arbitrary logical input state, x, into an arbitary logical output state, y. Then choose an arbitrary thermodynamically reversible process and stipulate that its initial state represents x and its final state represents y. This may seem like cheating, but without the requirement that we consider a device which is guaranteed to implement a logical transformation for *multiple* input states, there is no thermodynamic constraint.

6 Irreversibility directly from entropies

In this section, as is becoming common, we make use of information theoretic entropies to derive the main result. Previous arguments in the literature are usually restricted to specific examples of L-machines, from which the authors generalise without rigorous proof. For those readers, such as John Norton, who are dubious about this latter aspect of defences of Landauer's Principle, we offer here a general argument, stating explicitly the reasoning which may have been implicit in previous work. Norton also raises concerns about the use of information theoretic entropies in conjunction with thermodynamic results in this context. At the end of this section we respond to some of Norton's concerns. We also refer the reader to Appendix B.

If the different microstates of a system are discrete (as for example with energy eigenstates in quantum theories) and occupied with probability λ_n, we define the (information theoretic) entropy of that system by

$$S = -k \sum_n \lambda_n \ln \lambda_n. \tag{13}$$

Note that this coincides with the thermodynamic entropy when the system has a canonical probability distribution, and that it differs from the standard information-theoretic entropy by a normalisation factor of $k \ln 2$.

As before, consider an L-machine which implements an irreversible logical transformation L, and two logical input states $(x_1, x_2) \in X$ which map onto the same logical output state $y \in Y$ (such that $L(x_1) = L(x_2) = y$).

Now consider the following process $p^{m(q_1)}$: Prepare the device D in the state $D_{in}^{m(q_1)}$ which is a mixture of the states $D_{in}(x_1)$ and $D_{in}(x_2)$ with probabilities q_1 and $q_2 = (1 - q_1)$ respectively, and evolve it using Λ_L. This process yields the final state $D_{out}(y)$ with certainty.

Note that $p^{m(q_1)}$ is not one of the family of processes that implement L, because $p^{m(q_1)} \notin \{p_x | x \in X\}$, and because $D_{in}^{m(q_1)}$ is not a representative state. The fundamental difference between our approach and that of Maroney (2005) is that he isn't sensitive to these distinctions.

However, we can use the process $p^{m(q_1)}$ to prove a result for p_{x_1} and p_{x_2}. Using the definition of the information-theoretic entropy given by (13), the entropy of the device in the state $D_{in}^{m(q_1)}$ can be shown to be (e.g. Jones 1979)

$$S_{in}^{m(q_1)} = \sum_{n=1,2} q_n S_{in}(x_n) - k \sum_{n=1,2} q_n \ln q_n. \tag{14}$$

Hence the entropy change of the device when it is evolved using Λ_L is

$$\Delta S_{sys}(p^{m(q_1)}) = S_o(y) - \sum_{n=1,2} q_n (S_{in}(x_n) - k \ln q_n) \tag{15}$$

The entropy change of the heat reservoir during the process is given by

$$\Delta S_{res}(p^m) = \sum_{n=1,2} q_n \left(\frac{\Delta Q(p_{x_n})}{T} \right). \tag{16}$$

The total entropy change during the process $p^{m(q_1)}$ is therefore

$$\Delta S_{tot}(p^m) = \sum_{n=1,2} q_n \left(S_{out}(y) - S_{in}(x_n) + k \ln q_n + \frac{\Delta Q(p_{x_n})}{T} \right) \geq 0. \tag{17}$$

Using the definition of $\Delta S_{tot}(p_{x_n})$ given in equation (5), this implies that

$$\sum_{n=1,2} q_n \Delta S_{tot}(p_{x_n}) \geq -k \sum_{n=1,2} q_n \ln q_n > 0. \tag{18}$$

This is the same result that we obtained in equation (10) in the last section from which can be derived the results concerning bounds, and our Theorem.

In his criticism of the common approaches to this problem, Norton claims that thermodynamics cannot be applied to states in general probabilistic mixtures (such as the state $D_{in}^{m(q_1)}$), but only to those corresponding to a canonical mixture in which the microstate n, with energy E_n, is occupied with probability

$$\lambda_n = \frac{1}{Z} \exp\left(\frac{-E_n}{kT}\right) \tag{19}$$

where

$$Z = \sum_n \exp\left(\frac{-E_n}{kT}\right) \tag{20}$$

is the partition function for the state.

However, given any two thermodynamic states $D_{in}(x_1)$ and $D_{in}(x_2)$ which are canonically distributed, it is always possible to construct a canonical mixture by taking

$$q_1 = \frac{Z_1}{Z_1 + Z_2} \qquad q_2 = \frac{Z_2}{Z_1 + Z_2}, \tag{21}$$

where Z_1 and Z_2 are the partition functions for states $D_{in}(x_1)$ and $D_{in}(x_2)$ respectively. Considering the evolution of this canonical state is sufficient to prove our main result, that one or both of $\Delta S_{tot}(p_{x_1})$ or $\Delta S_{tot}(p_{x_2})$ must be greater than 0, and hence that the L-machine is thermodynamically irreversible. However, it is not sufficient to prove the stronger claim that both p_{x_1} and p_{x_2} must be thermodynamically irreversible.

7 A simple example: Reset

As an example of the above, consider the logical transformation 'Reset' represented by the truth table

Input	Output
0	0
1	0

The logical process 'Reset' represents the simplest logically irreversible transformation. Defining the input states x_1='0' and x_2='1', and the output states y_1='0' and y_2='1', the map $L_{Reset} : X \rightarrow Y$ corresponding to the reset operation is given by

$$L_{Reset}(x_1) = L_{Reset}(x_2) = y_1. \tag{22}$$

Often in the literature the reset operation is regarded as equivalent to the process of information *erasure* (in which the output bit is a maximally random bit independent of the input bit). Here we make a clear distinction between the two. The reset operation, L_{Reset}, is an irreversible *logical* operation, which maps logical states to logical states. The erasure operation, however, is a *non-logical* operation, since it maps each logical state to a probabilistic mixture of logical states (which are not logical states by definition). Thus, erasure is not in the scope of the present article. We note that, as reset is a logical operation, and a maximally random (i.e. erased) bit is useless for further information processing, reset and not erasure is used in practical computation.

We claim that every logically irreversible transformation is equivalent to a logically reversible transformation plus one or more reset operations. To see this consider an arbitrary logically irreversible transformation. It can be converted into a reversible transformation if a copy of the input state is appended to its output. This clearly allows the input state to be recovered from the output state. To obtain a transformation logically equivalent to the original irreversible transformation we simply reset all the copies.

Note also that in the proofs above, the only step which incorporates the irreversibility of the logical transformation L is that in which L is represented as having inputs x_1 and x_2 and an output y such that $L(x_1) = L(x_2) = y$, which is just how we characterised reset above. We therefore explicitly made use of the fact that every irreversible transformation incorporates reset.

A common choice of physical device D to implement L_{Reset}, considered by other authors and first introduced by Szilard, is the box B (of section 5) containing a single gas molecule and a moveable partition, in a heat reservoir at temperature T. We define the input and output states as follows:

$$D_{in}(x_1) = D_{out}(y_1) \quad = \quad \begin{array}{l}\text{Partition in the middle of the box,}\\ \text{and particle on the } \textit{left.}\end{array} \tag{23}$$

$$D_{in}(x_2) = D_{out}(y_2) \quad = \quad \begin{array}{l}\text{Partition in the middle of the box,}\\ \text{and particle on the } \textit{right.}\end{array} \tag{24}$$

The entropy of the system in all these states is the same.

One way to implement L_{Reset} using this box is to perform the following procedure $\Lambda_{L_{Reset}}$:

1. Remove the central partition

2. Isothermally compress the gas into the left hand half of the box using a piston. This requires work $kT\text{ln}2$ and causes heat flow $\Delta Q = kT \ln 2$ into the heat reservoir.

3. Replace the central partition

4. Withdraw the piston (to the right)

With this procedure we have

$$\Delta S_{tot}(p_{x_1}) = \Delta S_{tot}(p_{x_2}) = \frac{\Delta Q}{T} = k \ln 2 \tag{25}$$

and hence resetting either initial state is a thermodynamically irreversible process, and the L_{Reset}-machine $\{D, \{D_{in}(x)|x \in X\}, \{D_{out}(y)|y \in Y\}, \Lambda_{L_{Reset}}\}$ is *thermodynamically irreversible.*

However, consider the process $p^{m(1/2)}$ in which the device is prepared in the thermodynamic state $D_{in}^{m(1/2)}$ which is an equal mixture of $D_{in}(x_1)$ and $D_{in}(x_2)$ and then evolved (as in Maroney 2005). We have $\Delta S_{tot}(p^{m(1/2)}) = 0$, and hence this particular process $p^{m(1/2)}$ *is* thermodynamically reversible. Indeed, we can see that the reverse procedure $\Lambda'_{L_{Reset}}$ described thus:

1. Move the piston into the centre of the box (from the right)

2. Remove the central partition

3. Allow the gas to expand isothermally into the right-hand half of the box by pushing against the piston. During this process, we can extract work $kT \ln 2$ from the gas, causing heat to flow from the reservoir so that $\Delta Q = -kT \ln 2$.

4. Replace the central partition

will take the final state $D_{out}(y_1)$ back onto an initial state $D_{in}^{m(1/2)}$. However, as $D_{in}^{m(1/2)}$ does not correspond to a logical input state, and p^m does not correspond to a logical process, this cannot be considered an implementation of L. Furthermore, the evolution $\Lambda_{L_{Reset}}$ followed by $\Lambda'_{L_{Reset}}$ will leave *all* input states in the final state $D_{in}^{m(1/2)}$, so clearly the system cannot truly be considered as reversible.

Note that by modifying the procedure carried out by the device, it is possible to obtain arbitrarily small (non-zero) values of $\Delta S_{tot}(p_{x_1})$. However $\Delta S_{tot}(p_{x_2})$ must then become arbitrarily large to compensate (see the discussion of bounds at the end of section 5 and in Appendix A). One possible scheme is the following

1. Isothermally compress the gas in the right-hand half of the box into the fraction v of its initial volume nearest the partition, using a piston. If the particle is initially in the right-hand side of the box, this requires work $-kT \ln(v)$ and causes a heat flow $\Delta Q = -kT \ln(v)$ into the reservoir.

2. Remove the central partition

3. Isothermally compress the gas into the left hand half of the box by further moving the piston. This requires work $kT\ln(1+v)$ and dissipates heat $\Delta Q = kT\ln(1+v)$ into the reservoir.

4. Replace the central partition

5. Withdraw the piston (to the right)

This gives $\Delta S_{tot}(p_{x_1}) = k\ln(1+v) < k\ln 2$ and $\Delta S_{tot}(p_{x_2}) = k\ln(1+1/v) > k\ln 2$.

8 Known and unknown data

It is important to be clear about what is meant by the notions of known and unknown data. First note that in practice data is always about some target system, however, what is relevant to computation is only what is accessible to the computational device. Hence, when we refer to known and unknown data this must be understood in terms of what is 'known' by the device. Consider a computational device D that contains a register R, and a memory M which can be read. Both R and M have representative states, and we assume without loss of generality that they have the same number of them. We stipulate that everything accessible to the device throughout a computation is contained within the register and the memory; nothing else is allowed to influence its evolution. (Clearly then if an operator was allowed to input further data in the middle of a computation then we must regard them as part of the device.)

We define what it is for the register R to contain known or unknown data as follows:

1. *Known data*: The representative state of the register R and the representative state of the memory M both represent the same logical state.

2. *Unknown data*: The representative state of the register R and the representative state of the memory M need not both represent the same logical state.

Now suppose that we wish to implement the logical transformation L on the register R, whilst keeping the state of the memory M unchanged. As an example, let $L = L_{reset}$ as defined in section 7. Then we can define the global operation on the register and the memory respectively as $L' = L_{reset} \times I$, where I is the identity operation.

L' can be represented by the following truth-table:

RM	RM
00	00
01	01
10	00
11	01

As L' is logically irreversible, by our Theorem, any L'-machine is thermodynamically irreversible.

However, in the case of known data the reset operation on R can be implemented by the global operation L_{uncopy} which is given by the following truth-table:

RM	RM
00	00
01	11
10	10
11	01

This is because the middle two lines where L_{uncopy} differs from L' are guaranteed not to occur for known data. However, note that L_{uncopy} is logically reversible and therefore it is possible to implement it thermodynamically reversibly.

In the light of what we say in section 7 about the relationship between logically irreversible transformations and reset, this result generalises immediately to the result that there are no irreversible logical transformations on known data, provided the memory is not changed by the transformation (c.f. Bennett (2003)).

9 Single-process Irreversibility

Although the thermodynamic reversibility of a particular process $p^{m(q_1)}$ is not indicative of whether an L-machine is thermodynamically reversible or not (as discussed in section 6), it is interesting to consider whether there is any single process which represents the thermodynamic irreversibility of the L-machine.

In fact there is, but to express it we must consider not simply the device D, but the target system A (external to the device) which is the source of the data on which the L-machine is to operate. Consider the following process $p^{A(L)}$:

1. Prepare the target system A in an equal probabilistic mixture of its representative states $\{A(x)|x \in X\}$ (representing uncertainty in the initial data), and the device D in the standard state $D_{in}(x_1)$ (its ready state).

2. Perform a controlled operation from A to D such that when A is in the state $A(x)$, the state $D_{in}(x_1)$ is transformed thermodynamically reversibly to the state $D_{in}(x)$. This corresponds to copying the initial data from the target into the device.

3. Perform the evolution Λ_L of the L-machine. This corresponds to performing the required computation.

By considering the information-theoretic entropies of the combined probability distribution for D and A, it is easily seen that

$$\Delta S_{tot}(p^{A(L)}) = \frac{1}{|X|} \sum_x S_{tot}(p_x), \tag{26}$$

where $|X|$ is the number of possible logical inputs.

Hence, $\Delta S_{tot}(p^{A(L)}) > 0$ (i.e. the process $p^{A(L)}$ will be thermodynamically irreversible) if and only if at least one of the processes p_x is thermodynamically irreversible. From our definition of the thermodynamic irreversibility for an L-machine, it is clear that the process $p^{A(L)}$ will be thermodynamically irreversible if and only if the L-machine with which it is implemented is thermodynamically irreversible.

To our Theorem, we can therefore add the following corollary: If L is logically irreversible, then the process $p^{A(L)}$ will be thermodynamically irreversible.

10 Conclusions

We note the clarification that is achieved by carefully distinguishing between logical and physical concepts, and between individual processes and families of processes. The introduction of the conceptual tool of L-machines, and the definition of what it is for one to be thermodynamically irreversible enables us to achieve generality without loss of precision. An L-machine is thermodynamically irreversible if at least one of the processes p_x, corresponding to activation of the device with the input state corresponding to x, is thermodynamically irreversible. In particular, the process p_x will be thermodynamically irreversible whenever x cannot be unambiguously determined from $L(x)$ (and we give quantitative bounds on this irreversibility compatible with the quantitative form of Landauer's Prinicple).

Hence, we are able to prove the generalised qualitative form of Landauer's principle:

> If L is logically irreversible, then every L-machine is thermodynamically irreversible.

This completes the sought after connection between logical and thermodynamical irreversibility.

Acknowledgements

The authors would like to thank Sandu Popescu, Stephen Leeds, Craig Calender, and Adam Elga for helpful discussions. A.Short acknowledges support from the U.K. Engineering and Physical Sciences Research Council (IRC in "Quantum Information Processing"). B.Groisman acknowledges support from the EU under European Commission project RESQ.

Appendix A

In section 5, we showed that once $\Delta S_{tot}(p_{x_1})$ is fixed, $\Delta S_{tot}(p_{x_2})$ is bounded from below by

$$\Delta S_{tot}(p_{x_2}) \geq -k \ln \left(1 - \exp\left(\frac{-\Delta S_{tot}(p_{x_1})}{k}\right)\right). \tag{27}$$

However, this bound was obtained by considering a specific initial setup, in which $q_1 = \exp(-\Delta S_{tot}(p_{x_1})/k)$, and it is conceivable that a stronger lower bound could be obtained by considering other values of q_1 in the range $0 < q_1 < 1$. In this appendix, we prove that the above choice of q_1 is optimal, and that the bound given by (27) is the strongest lower bound on $\Delta S_{tot}(p_{x_2})$ that can be obtained.

From equation (10) the lower bound $L(q_1)$ on $\Delta S^{LB}_{tot}(p_{x_2})$ for a general q_1 is given by

$$\Delta S_{tot}(p_{x_2}) \geq L(q_1) = \frac{1}{q_2}\left(-q_1 \Delta S_{tot}(p_{x_1}) - k \sum_{n=1,2} q_n \ln q_n\right) \tag{28}$$

where $q_2 = 1 - q_1$. Hence

$$\begin{aligned} L(q_1) &= -\frac{q_1}{q_2}\left(\Delta S_{tot}(p_{x_1}) + k \ln q_1\right) - k \ln q_2, & (29)\\ \frac{\mathrm{d}L(q_1)}{\mathrm{d}q_1} &= -\frac{1}{q_2^2}\left(\Delta S_{tot}(p_{x_1}) + k \ln q_1\right), & (30)\\ \frac{\mathrm{d}^2L(q_1)}{\mathrm{d}q_1^2} &= \frac{2}{q_2}\frac{\mathrm{d}L(q_1)}{\mathrm{d}q_1} - \frac{k}{q_1 q_2^2}. & (31) \end{aligned}$$

As $q_1 > 0$ and $q_2 > 0$, the lower bound $L(q_1)$ has a unique extremal point, which is a maximum, at

$$q_1 = \exp\left(-\frac{\Delta S_{tot}(p_{x_1})}{k}\right). \tag{32}$$

But this is just the value of q_1 we originally selected, proving that the lower bound being given by equation (27) is the strongest lower bound that can be obtained.

Appendix B

Here we offer a general argument for applying the second law to a system in a heat reservoir even when it does not have a canonical probability distribution (such as in the process $p^{m(q_1)}$ of section 6). We assume that it makes sense to talk about the state of the universe and any subsystem of it in probabilistic terms, but we remain neutral about whether such probabilities are epistemic or ontic. We also assume for simplicity that all the relevant microstates are discrete and hence that we can use the definition of the information theoretic entropy in equation (13).

Suppose the universe consists of system, in which we are interested, and everything else, which we treat as a heat reservoir. We define $S_{universe}$ as the entropy of the probabilistic state of the universe. As for any composite object, the entropy of the universe will in general be less than the sum of the entropies of the system and the heat reservoir individually, due to correlations between them. We therefore write

$$S_{universe} = S_{sys} + S_{res} - I_{(sys,res)} \tag{33}$$

where $I_{(sys,res)}$ is a positive quantity (usually referred to as the mutual information) corresponding to the entropy associated with the correlations between system and reservoir.

As a simple example of this, consider an equal mixture of the states '$0_A 0_B$' and '$1_A 1_B$' of a bipartite binary system, which has $S_{AB} = S_A = S_B = I_{(A,B)} = k \ln 2$.

For a general process p on a system in a reservoir, we have

$$\Delta S_{universe}(p) = \Delta S_{sys}(p) + \Delta S_{res}(p) - \Delta I_{(sys,res)}(p) \tag{34}$$

We then make the following assumptions:

1. We assume that the underlying dynamics are reversible such that the total entropy of the universe does not decrease over time.

   I.e. $\forall p, \Delta S_{universe}(p) \geq 0$.

2. We assume that with respect to time scales that are sufficiently large, the correlations between the system and the heat reservoir increase or are constant, that is, that $\forall p, \Delta I_{(sys,res)}(p) \geq 0$. This encapsulates the fact that it is very hard to undo correlations made with a large thermal environment, as the necessary information disperses away rapidly into many degrees of freedom.

Using these assumptions we can recover a statement of the second law applicable to general information-theoretic entropies:

$$\forall p, \quad \Delta S_{tot} = \Delta S_{sys} + \Delta S_{res} \geq 0 \tag{35}$$

References

Bennett, C. H. (1973). The logical reversibility of computation. *IBM Journal of Research and Development*, 17, 525-532.

Bennett, C. H. (1982). The Thermodynamics of ComputationA Review. *International Journal of Theoretical Physics, 21*, 905940. Reprinted in Leff and Rex (1990), 213248.

Bennett, C. H. (1987). Demons, Engines and the Second Law. *Scientific American, 257*, 108116.

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

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

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

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

Feynman, R. P. (1996). *Feynman Lectures on Computation*, edited by J. G. Hey and W. Allen. Reading, MA: Addison-Wesley

Jones, D. S. (1979). *Elementary Information Theory*, Oxford: Clarendon Press .

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 (1990)).

Landauer, R. (1961). Dissipation and heat generation in the computing process. *IBM Journal of Research and Development*, 5, 183191

Leff, H. S., and Rex, A. F., (Eds.)(1990). *Maxwell's demon: Entropy, information, computing*, Bristol: Adam Hilger.

Leff, H. S., and Rex, A. F. (Eds.) (2003). *Maxwell's demon 2: Entropy, classical and quantum information, computing.* Bristol: Institute of Physics.

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

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

Szilard, L. (1929). On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings. *Zeitschrift für Physik*, 53, 840-856. Reprinted in Leff and Rex (1990), 124-133.

Uffink, J. (2001). Bluff Your Way in the Second Law of Thermodynamics. *Studies In History and Philosophy of Modern Physics*, 32, 305-394.