Physics and Computation: The Status of Landauer's Principle
(Extended Abstract)

James Ladyman

Department of Philosophy, University of Bristol

1 Introduction

Realism about computation is the view that whether or not a particular physical system is performing a particular computation is at least sometimes a mind-independent feature of reality. The caveat 'at least sometimes' is necessary here because a realist about computation need not believe that all instances of computation should be realistically construed. The computational theory of mind presupposes realism about computation. If whether or not the human nervous system implements particular computations is not a natural fact about the world that is independent of whether we represent it as doing so, then the computational theory of mind fails to naturalise the mind. Realism about computation is also presupposed by attempts to use computational principles such as Landauer's Principle to dispel Maxwell's Demon. Realism about computation has been challenged by Hilary Putnam and John Searle among others. Various arguments have been put forward purporting to show that any physical system of sufficient complexity trivially implements all computations. Ladyman et al. (2007) offer a precisification and general proof Landauer's Principle. In order to do this they present an analysis of what it is for a physical process to implement a logical transformation. In this paper, their analysis is explained and its implications for realism about computation and the use of Landauer's Principle in foundational debates is assessed.

When we are concerned with the logical form of a computation and its formal properties then it can be theoretically described in terms of functions and relations between abstract entities. However, actual computation is realised by some physical process, and the latter is of course subject to physical laws and the laws of thermodynamics in particular. It is therefore important to consider whether or not there are any systematic connections between the logical properties of computations consider abstractly and the thermodynamical properties of their realizations. Rolf Landauer (1961) proposed such a general connection, known as Landauer's Principle, namely that the erasure of information in any 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

S.B. Cooper, B. Löwe, and A. Sorbi (Eds.): CiE 2007, LNCS 4497, pp. 446–454, 2007.

reversible process can be implemented thermodynamically reversibly (see for example Charles Bennett 2003). Landauer's Principle 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.

In order to clarify the status of Landauer's Principleis it is necessary to precisely define a computation, and what it means to say that a computation is physically realized. In particular, this paper offers precise definitions of logical irreversibility and thermodynamic irreversibility, and a detailed analysis of what it means for a physical system to implement a logical transformation. The result of this analysis is the notion of an *L-machine*. This is a hybrid physical-logical entity 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. Landauer's Principle can be restated and generalized to the claim that the logical irreversibility of L implies thermodynamic irreversibility of every corresponding L-machine.

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).

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

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

It is important to make a clear distinction between the logical and physical domains, and to avoid talk of logical 'processes' and refer instead to logical transformations and their implementation by *families* of physical processes. The term 'process' always refers to a physical process in which a system starts in some particular state and is guaranteed to end in some particular state[1]. Landauer's

[1] In general, the particular end state may be a probabilistic mixture of thermodynamic states, but usually the final state is not such a mixture. Although, in the former case, the system may be supposed to actually be in some specific component of the mixture, it is not guaranteed to end up in that component, and so this component state cannot be considered as the final state of the process.

Principle is only considered in the general and precise form introduced above: If L is logically irreversible, then every L-machine is thermodynamically irreversible[2].

2 Logical Irreversibility

A logical transformation is a *mathematical* operation, consisting of a single-valued map L from a finite set X of input states, into a finite set Y of output states (i.e. each input state is mapped by L 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). A logical transformation is *logically reversible* if and only if $L : X \rightarrow Y$ is a one-to-one (injective) mapping[3]. Hence with a reversible logical transformation, it is possible to uniquely reconstruct the input state from the output state. If L is not a one-to-one mapping, then it is *logically irreversible*.

It is crucial that there is 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 (The point here is clear in the case of a truth table, where each member of the family of processes corresponds to a single row). The notion of implementation of a logical transformation by a physical device is discussed in section 4 below.

3 Thermodynamic Irreversibility

Thermodynamic irreversibility is a feature of *physical* processes, expressed by the second law of thermodynamics. There is much controversy about how the latter can be justified on the basis of statistical mechanics. Without assuming anything about the relationship between phenomenological thermodynamics and statistical mechanics, its is assumed that the second law stated in terms of *thermodynamic entropy* is valid.

In thermodynamics various operational assumptions are made that allow the definition of the thermodynamic entropy of individual macroscopic states (up to

[2] An L-machine is just the most general way of capturing the idea of physically implementing a logical transformation L.

[3] Note that whether or not L is surjective is irrelevant for the present paper. This is because if there are output states that do not get arrived at by the implementation of the transformation these are irrelevant to thermodynamic considerations.

a constant)[4]. This is almost universally accepted, however, there is controversy about the assignment of entropy to probabilistic mixtures of macrostates (for example, see Norton (2005)). For example, consider the mixture of macrostates M_i, with probabilities q_i. Assuming that the assignment of entropy to such a state is legitimate, it might be supposed that it is simply the average of the individual entropies $S(M_i)$; explicitly, $\sum_i q_i S(M_i)$. However, it is common to also include a term to represent the contribution to the entropy of the probability distribution itself; explicitly:

$$S_{mixture} = \sum_i q_i S(M_i) - k \sum_i q_i \ln q_i \tag{1}$$

The latter term is an information theoretic entropy and its inclusion in thermodynamic calculations currently lacks rigorous foundational justification[5]. Ladyman et al (2007) offers a proof of Landauer's Principle that depends on the use of the information theoretic entropy and a proof that is independent of it.

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

$$\forall\, p, \quad \Delta S_{sys}(p) + \frac{\Delta Q(p)}{T} \geq 0 \tag{2}$$

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

$$\Delta S_{tot}(p) = \Delta S_{sys}(p) + \Delta S_{res}(p) \tag{3}$$

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

$$\forall p, \quad \Delta S_{tot}(p) \geq 0 \tag{4}$$

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. Therefore any process p for which $\Delta S_{tot}(p) > 0$ is *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.

A family of physical processes is thermodynamically irreversible if and only if at least one of its members is. This is important for the definition of irreversibility for L-machines in the next section.

[4] See, for example, Fermi (1936), Chapter IV.

[5] However, such a justification is the subject of work in progress by Ladyman, Presnell and Short.

4 Implementing a Logical Transformation with a Physical Device

In order to analyze the connection between *logical* transformations, and *physical* thermodynamic processes, it is necessary to consider what it means for a physical system to implement a logical transformation. As stated above, a physical system can only implement a logical transformation through a family of processes. To physically implement a logical transformation, there must be: A physical device, a specification of which physical states of that device correspond to the possible logical states (call the former *representative states*), and a time evolution operator of that device. This combined system is an *L-machine*. Note that L names a particular logical transformation, so there are L_{AND}-machines, and so on.

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 the present approach and that of Maroney (2005) who considers only individual processes.

Furthermore, it is important to note that the time evolution operator must encode everything about the behaviour of the device, and so the possibility of an external agent intervening during its operation is ruled out. In particular this prohibits any such external agent affecting the time evolution of the system by making use of information about its state while it is running. In other words, intelligent agents (such as demons) may be introduced only if their knowledge and actions affecting the operation of the device are included in the specification of the L-machine and its time evolution. Heuristically, suppose that the interaction between the L-machine and the rest of the world is limited to the setting of the input state and the pressing of the 'go' button.

Formally, an L-machine is an ordered set

$$\{D, \{D_{in}(x)|x \in X\}, \{D_{out}(y)|y \in Y\}, \Lambda_L\} \tag{5}$$

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

$$\forall\, x \in X, \Lambda_L(D_{in}(x)) = D_{out}(L(x)). \tag{6}$$

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 $y = L(x) \in Y$. This physical process is denoted by p_x.

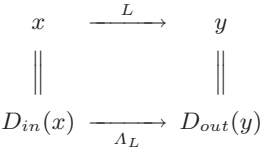

Fig. 1. An illustration of the relationship between the logical states x and y and their representative physical states $D_{in}(x)$ and $D_{out}(y)$, showing the logical transformation L and the physical time evolution operator Λ_L

Note that the labelling of the states is essential to the identity of a L-machine. For example, exactly the same device and time-evolution operator could be used as part of both an L_{AND}-machine, and an L_{OR}-machine by the appropriate relabelling of the physical input and output states.

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 $D_{out}(L(x))$ is $S_{out}(L(x))$, 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

$$\Delta S_{tot}(p_x) = S_{out}(L(x)) - S_{in}(x) + \frac{\Delta Q(p_x)}{T} \geq 0. \tag{7}$$

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.

An L-machine is *thermodynamically reversible* if and only if for all $x \in X$, $\Delta S_{tot}(p_x) = 0$ (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 $\Delta S_{tot}(p_x) > 0$.

Note implementing L by implementing some other 'stronger' L' from which the outputs of L can be deduced is ruled out; for example, the logical transformation L' corresponding to the combination of L and keeping a copy of the input. Formally, a logical transformation L' is *stronger* than a logical transformation L just in case, for every input x, $L(x)$ can be recovered from $L'(x)$, but for at

least one x, $L'(x)$ cannot be recovered from $L(x)$. It follows that if L' is stronger than L, then, for every x, $L(x) = L^*(L'(x))$, where L^* is a logically irreversible transformation[6]. In general an implementation of a logically stronger L' is not an implementation of L, and is *unfaithful* in the following sense: it allows that, for some x, more can be learnt about the value of x from the output $L'(x)$ than from $L(x)$ itself. Allowing that L can be implemented by the implementation of a logically stronger transformation L' must also be ruled out because it begs the question at issue here by implicitly assuming that Landauer's Principle is false: it would always be possible to implement a logically irreversible process by implementing a stronger logically reversible process, and all sides agree that this could be done in a thermodynamically reversible way.

Note also that in the above definition a unique representative state is assigned to each logical state as this makes for a clear and simple analysis. However, in general it could be allowed that more than one physical state represents the same logical state, in which case, for each x, $D_{in}(x)$ would be replaced by a set $\{D_{in}^{(1)}(x), D_{in}^{(2)}(x) \ldots\}$ of distinct physical states (and similarly for each y). Call such a generalisation a 'multi-L-machine'. The condition (6) on the time-evolution operator of the device would then generalise in an obvious way to

$$\forall\, x \in X, \forall\, D_{in}^{(i)}(x), \exists\, D_{out}^{(j)}(L(x)) \;:\; \Lambda_L(D_{in}^{(i)}(x)) = D_{out}^{(j)}(L(x)). \tag{8}$$

By definition, each representative state is a physically distinguishable macro-state, so assume that the device can be prepared in a specific $D_{in}^{(i)}(x)$, and it can be determined which of the $D_{out}^{(j)}(y)$ it ends up in. Hence, a *refinement* of any multi-L-machine, is the multi-L-machine obtained by choosing a particular representative state for each logical input state, and their corresponding output states, and keeping the device and time evolution operator the same.

For many multi-L-machines, every refinement is an L-machine and in such cases nothing is gained by considering the generalisation. However, in every other case there exists a refinement which under relabelling of its output states is an L'-machine, for some L' that is logically stronger than L. This is unfaithful in the sense defined above, and hence is ruled out. Furthermore, without ruling out these cases then, for any logically irreversible L, a machine that implements L merely in virtue of the fact that it is stipulated that for every logical input state x, the same physical state represents x and $L(x)$, where the time evolution operator is the identity operator could be considered. This clearly trivialises the notion of implementing a logical transformation. It is ruled out by the prescription above since it could be used to implement the logically stronger identity operation.

On the basis of the above definitions it is possible to prove Landauer's Principle from the Kelvin statement of the Second Law of Thermodynamics using a thermodynamic cycle.

[6] Note that $L'(x)$ can itself be logically irreversible, such as the logical transformation L' corresponding to the combination of L_{AND} and keeping a copy of the second input bit. L' is stronger than L_{AND} but is still logically irreversible.

Acknowledgements. This abstract is based on Ladyman, J., Presnell, S., Short, A. and Groisman, B. (2007), 'The connection between logical and thermodynamic irreversibility', in Studies in History and Philosophy of Modern Physics.

References

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

Bennett, C.H.: The Thermodynamics of Computation?A Review International Journal of Theoretical Physics 21, 905–940 (1982) (Reprinted in Leff and Rex (1990), 213–248)

Bennett, C.H.: Demons, Engines and the Second Law. Scientific American, vol. 257, pp. 108–116

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

Brillouin, L.: Maxwell's demon cannot operate: Information and entropy. I. Journal of Applied Physics 22, 338–343 (1951)

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

Earman, J., Norton, J.D.: 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 (1998)

Earman, J., Norton, J.D.: 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 (1999)

Fermi, E.: Thermodynamics. Dover, New York (1936)

Feynman, R.P.: Feynman Lectures on Computation. In: Hey, J.G., Allen, W. (eds.) Reading, MA, Addison-Wesley, London (1996)

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

Ladyman, J., Presnell, S., Short, A., Groisman, B.: The connection between logical and thermodynamic irreversibility. Studies in History and Philosophy of Modern Physics (2007)

Landauer, R.: Irreversibility and heat generation in the computing process. IBM Journal of Research and Development (Reprinted in Leff and Rex (1990)) 5, 183–191 (1961)

Landauer, R.: Dissipation and heat generation in the computing process. IBM Journal of Research and Development 5, 183–191 (1961)

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

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

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

Norton, J.D.: 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 (2005)

Piechocinska, B.: Information erasure. Physical Review A, 61, 062314, 1–9 (2000)

Shizume, K.: Heat generation required by information erasure. Physical Review E 52, 3495–3499 (1995)
Szilard, L.: On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings. Zeitschrift für Physik 53, 840–856 (1929) (Reprinted in Leff and Rex (1990), 124–133)
Uffink, J.: Bluff Your Way in the Second Law of Thermodynamics. Studies In History and Philosophy of Modern Physics 32, 305–394 (2001)