Landauer's Erasure Principle and Data Compression

Stefan Wolf*†
* Faculty of Informatics, Università della Svizzera italiana, 6900 Lugano, Switzerland
† Facoltà indipendente di Gandria, 6978 Gandria, Switzerland

***Abstract*—A consequence of Landauer's slogan "Information Is Physical," when combined with the second law of thermodynamics, is *Landauer's erasure principle*: Deleting a binary string of length N requires work proportional to N. We suggest to modify the principle in different respects. First, we consider the erasure process constructively, *i.e.*, as an algorithm carried out by a Turing machine. Second, we claim the erasure price to be lower for *redundant*, *i.e.*, compressible strings (given the demon's algorithmically constructive information about the string). Third, our bounds are functions only of the objects in question (the string and the extractor's knowledge) and do not depend on any context such as a probability distribution. We pursue the idea that the erasure cost measures intrinsic randomness (applicable, *e.g.*, to quantum correlations), and we finally turn back our attention to the second law of thermodynamics for which we propose a version relating it more closely to Turing- than steam machines.**

I. Landauer's Principle and Its Converse

According to Landauer [13], "information is physical:" Any information generation, storage, processing, and transmission is ultimately physical and must be understood as such. A consequence of this insight is *Landauer's principle* [12]: Erasing N bits of information costs an amount of at least $kTN \ln 2$ of free energy which is then dissipated as heat to the environment (of temperature T; k is Boltzmann's constant; "erasing N bits" stands for "forcing the corresponding N binary degrees of freedom into the state 0"). The principle has been derived by Landauer from the second law of thermodynamics: The reduction of entropy within the storage device must be compensated by an increase, of at least the same amount, in environmental entropy.

We suggest to modify Landauer's principle in the following respects: First, we claim the erasure cost to be proportional to the *length of a compression* of the string in question, not to its full length. Second, the erasure device's *knowledge* is taken into account, and it can reduce the erasure cost. In line with the *Church-Turing hypothesis*,[1] the erasure process, and, hence, also the nature of this knowledge, are understood algorithmically-constructively, *i.e.*, to be computed by a Turing machine. The resulting erasure price is context-free and depends only on the objects in question, *i.e.*, on the string and the Turing machine including its tape's initial state (the "knowledge"), but does not involve entropies, not even probability distributions.

In the context of his resolution of *Maxwell's demon's paradox*, Bennett [3] stated the converse of Landauer's principle: If the demon's memory initially is in the all-0-state, and it is randomized after the sorting procedure, then that initial 0-string can be regarded as the *resource* carrying the free energy required for the sorting. Explicitly, the converse of Landauer's principle states that (a physical representation of) the string 0^N of length N has *work value* $kTN \ln 2$. If, for example, the N bits are encoded in N gas molecules being on the left *vs.* the right half of a container, respectively, then the all-0-string corresponds to a *compressed gas* — with work value.

We directly connect, for any physical representation of a binary string, its *work value* and *erasure cost*: They add up to the full length of the string. So, any bound on the work value yields a bound on the erasure cost. We review the state of the art on *work extraction*.

II. Work Extraction: State of the Art

A. The Results by Bennett and by Zurek

Bennett [3] claimed the work value of a string S to be *its length minus the algorithmic entropy*, the latter being the length of the shortest program that lets a fixed universal Turing machine $\mathcal{U}$ output S. The algorithmic entropy of S has also been called *Kolmogorov complexity* $K(S)$ *of* S [11]:

$$\mathrm{WV}(S) = \mathrm{len}(S) - K(S)$$

(let for simplicity $kT \ln 2 = 1$). Bennett's argument is that S can be logically, hence, thermodynamically [10] *reversibly* mapped to the string $P||000\cdots0$, where P is the shortest program for $\mathcal{U}$ generating S and the length of the generated 0-string is $\mathrm{len}(S) - K(S)$ (see Figure 1).

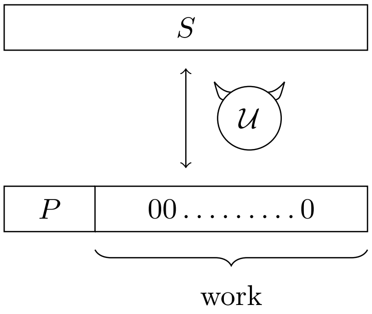


Figure 1. Bennett's argument

[1] The *Church-Turing hypothesis* states that all physically possible processes can be simulated by a Turing machine.

It was already pointed out by Zurek [18] that whilst it is true that the reverse direction exists and is computable by a universal Turing machine, its *forward direction*, *i.e.*, obtaining P from S, is *uncomputable*. This means that a demon that could carry out this work-extraction computation on S does not exist under the Church-Turing hypothesis. We will see, however, that Bennett's value is an *upper bound* on the work value of S (and may even be reasonably tight in many cases). Bennett also links the string's erasure cost to its probabilistic entropy [5].

B. The Results by Szilárd and by Dahlsten et al.

Dahlsten *et al.* [7] follow Szilárd [14] in putting the *knowledge* of the demon extracting the work to the center of their attention. More precisely, they claim

$$\mathrm{WV}(S) = \mathrm{len}(S) - D(S)\,,$$

where the "defect" $D(S)$ is bounded from above and below by a *smooth Rényi entropy* of the distribution of S from the demon's viewpoint, modeling her ignorance.

Building on the mentioned results and in the same probabilistic model, the cost of erasure [8] as well as of general computations [9] have been linked to entropic expressions of (conditional) probability distributions.

Comparison and Discussion. The work [7] does not consider the *algorithmic* aspects of the demon's actions extracting the free energy, and it is based on the demon's *probabilistic-entropic knowledge on S*. If we want, in the spirit of the Church-Turing hypothesis, to model the demon as an algorithmic apparatus, then we should specify the nature of that knowledge explicitly: Vanishing conditional entropy only says that S is uniquely determined from the demon's viewpoint; this can either mean that the demon has a *copy* of S (or at least the *ability* to compute one), or the knowledge is weaker, merely singling out S in a *non-constructive* way. We will see below in detail how this ambiguity sits at the origin of the gap between the two described groups of results; it is maximal when the demon fully "knows" S which, however, still has maximal Kolmogorov complexity given her internal state. In this case, the first result claims $\mathrm{WV}(S)$ to be 0, whereas $\mathrm{WV}(S) \approx \mathrm{len}(S)$ according to the second. The gap disappears if "knowing S" is understood in the *constructive* as opposed to entropic sense, and the demon can access an extra copy of S — besides the "original" S from which work is to be extracted. If that extra copy is included in Bennett's reasoning, then his result reads

$$\mathrm{WV}(S||S) \approx 2\,\mathrm{len}(S) - K(S) \approx \mathrm{len}(S)\ .$$

III. Work Extraction as Data Compression

We analyze the case of a demon with knowledge and understand work extraction to be a *computation* carried out by this demon.

A. The Model

We assume the demon to be a *universal Turing machine* $\mathcal{U}$ the memory tape of which is sufficiently long for the inputs and tasks in question, but *finite*. The tape initially contains S, the string the work value of which is to be determined, X, a finite string modeling the demon's *knowledge about* S, and 0's for the rest of the tape. After the extraction computation, the tape contains, at the bit positions initially holding S, a (shorter) string P plus $0^{\mathrm{len}(S)-\mathrm{len}(P)}$, whereas the rest of the tape is (again) the same as before work extraction. The demon's operations are *logically* reversible and can, hence, be carried out *thermodynamically* reversibly [10]. Logical reversibility is the ability of the same demon to carry out the backward computation step by step, *i.e.*, from $P||X$ to $S||X$. We denote by $\mathrm{WV}(S|X)$ the *maximal length of an all-*0*-string extractable logically reversibly from S, given the knowledge X*, *i.e.*,

$$\mathrm{WV}(S|X) := \mathrm{len}(S) - \mathrm{len}(P)$$

if P's length is minimal (see Figure 2).

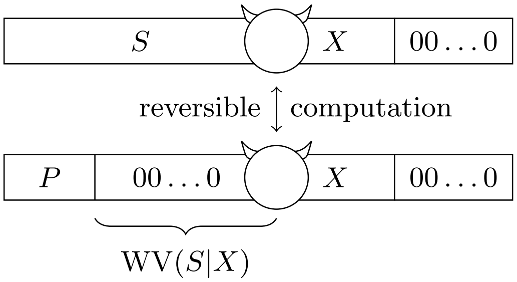


Figure 2. The model of work extraction with knowledge

B. Lower Bound on the Work Value

We show that every specific data-compression algorithm leads to a lower bound on extractable work. Let C be a computable function

$$C\,:\,\{0,1\}^* \times \{0,1\}^* \longrightarrow \{0,1\}^*$$

such that

$$(A,B) \mapsto (C(A,B),B)$$

is injective. We call C a *data-compression algorithm with helper*. Then we have

$$\mathrm{WV}(S|X) \geq \mathrm{len}(S) - \mathrm{len}(C(S,X))\,.$$

This can be seen as follows. First, note that the function

$$A||B\ \mapsto\ C(A,B)||0^{\mathrm{len}(A)-\mathrm{len}(C(A,B))}||B$$

is computable and bijective. From the two (possibly irreversible) circuits computing the compression and its inverse, one can obtain a *reversible* circuit realizing the function such that no further input or output bits are involved. This can be achieved by first implementing all logical operations with Toffoli gates and uncomputing the "junk" [4] in both circuits.

The resulting two circuits have now still the property that the input is part of the output. As a second step, we can simply combine the two such that the first circuit's first and second outputs become the second's second and first inputs, respectively. Roughly speaking, the first computes the compression and the second reversibly uncomputes the raw data. The combined circuit has only *the compressed data plus the 0's* as the output, sitting on the bit positions carrying the input before. (This circuit is roughly as efficient as the less efficient of the two irreversible circuits for data compression and decompression, respectively.) We assume the reversible circuit to be hard-wired in the demon. A typical example for an algorithm that can be used here is universal data compression *à la* Ziv-Lempel [17].

C. Upper Bound on the Work Value

We have the following upper bound on the extractable work:

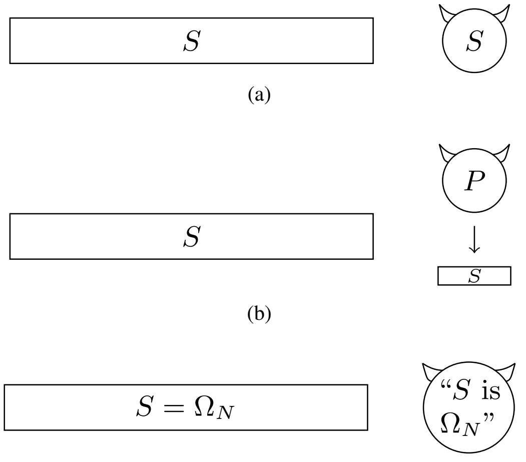
$$\mathrm{WV}(S|X) \leq \mathrm{len}(S) - K_{\mathcal{U}}(S|X)\,,$$

where $K_{\mathcal{U}}(S|X)$ is the conditional Kolmogorov complexity (with respect to the demon $\mathcal{U}$) of S given X, *i.e.*, the length of the shortest program P for $\mathcal{U}$ that outputs S, given X. The reason is that the demon is only able to carry out the computation in question (logically, hence, thermodynamically) reversibly if she is able to carry out the reverse computation as well. Therefore, the string P must be at least as long as the shortest program for $\mathcal{U}$ generating S if X is given.

Although the same is not true in general, this upper bound is *tight* if $K_{\mathcal{U}}(S|X) = 0$. The latter means that X itself is a program for generating an additional copy of S. The demon can then bit-wisely XOR this new copy of S to the original S (to be work-extracted) on the tape, hereby producing $0^{\mathrm{len}(S)}$ *reversibly* to replace the original S, at the same time saving the new one, as reversibility demands. When Bennett's "uncomputing trick" is used — allowing to make any computation by a Turing machine logically reversible [4] —, then a history string H is written to the tape during the computation of S from X such that after XORing, the demon can, in a (reverse) stepwise manner, *uncompute* the generated copy of S and end up in the tape's original state — except that the original S is now replaced by $0^{\mathrm{len}(S)}$: This results in a maximal work value matching the (in that case trivial) upper bound.

Discussion. Let us compare our bounds with the entropy-based results of [7]: According to the latter, a demon *knowing S entirely* is able to extract maximal work: $\mathrm{WV}(S) \approx \mathrm{len}(S)$. What does it mean to "know S"? The knowledge can consist of (a) a *copy* of S, or of (b) its ability to *compute* such a copy with a given program P, or (c) it can determine S uniquely *without* providing the ability to compute it (see Figure 3).

The constructive *versus* the entropic results are in accordance in the cases (a) and (b), but are *in conflict* in case (c): For instance, assume the demon's knowledge about S is: *"S equals the first N bits Ω_N of the binary expansion of Ω."* Here, Ω is the so-called halting probability [6] of a fixed universal Turing machine $\mathcal{A}$ (*e.g.*, the demon $\mathcal{U}$ itself). Although there is a short *description* of S in this case, and S is thus uniquely determined in an entropic sense, it is still incompressible, even given that knowledge:



(a)

(b)

(c)

Figure 3. Ways of knowing S

$$K_{\mathcal{U}}(\Omega_n \,|\, \text{"It is bits 1–}n\text{ of TM }\mathcal{A}\text{'s halting probability"}) \approx n :$$

No work is extractable according to our upper bound. This gap opens whenever the *"description complexity"* is smaller than the *Kolmogorov complexity*. (Note that a self-referential argument, called *Berry paradox*, shows that the notion of "description complexity" is problematic and can never be defined consistently for all strings.)

IV. Modifying Landauer's Principle

A. Connection to Work Value

For a string $S \in \{0,1\}^N$, let $\mathrm{WV}(S|X)$ and $\mathrm{EC}(S|X)$ be its work value and erasure costs, respectively, given an additional string X (a "catalyst" which remains unchanged, as above). Then

$$\mathrm{WV}(S|X) + \mathrm{EC}(S|X) = N\ .$$

To see this, consider first the combination extract-then-erase. Since this is *one specific way* of erasing, we have

$$\mathrm{EC}(S|X) \leq N - \mathrm{WV}(S|X)\ .$$

If, on the other hand, we consider the combination erase-then-extract, this leads to

$$\mathrm{WV}(S|X) \geq N - \mathrm{EC}(S|X)\ .$$

B. Bounds on the Erasure Cost

Given the results on the work value above, as well as the connection between the work value and erasure cost, we obtain the following bounds on the thermodynamic cost of *erasing a string* S by a demon, modeled as a universal Turing machine $\mathcal{U}$ with initial tape content X.

Landauer's principle, revisited. *Let* C *be a computable function* $C : \{0,1\}^* \times \{0,1\}^* \longrightarrow \{0,1\}^*$ *such that* $(A, B) \mapsto (C(A, B), B)$ *is injective. Then we have*

$$K_{\mathcal{U}}(S|X) \leq \mathrm{EC}(S|X) \leq \mathrm{len}(C(S, X)) \ .$$

V. Randomness and Quantum Correlations; Reversibility and the Second Law

Landauer's revised principle puts forward two ideas: First, the erasure cost is an *intrinsic, context-free, physical measure for randomness* (entirely independent of probabilities and counterfactual statements of the form "some value *could* just as well have been *different*"). The idea that the erasure cost — or the Kolmogorov complexity related to it — is a measure for randomness (independent of probabilities) can be tested in a context in which randomness has been paramount: *Bell correlations* [2] predicted by quantum theory. In a proof of principle, it was shown [16] that in essence, a similar mechanism as in the probabilistic setting arises: *If* the correlation is non-local and the inputs are incompressible and non-signaling holds, *then* the outputs must be highly complex as well. This allows for a discussion of quantum correlations without the usual counterfactual arguments used in derivations of *Bell inequalities* (combining in a single formula results of different measurements that cannot actually be carried out together). Furthermore, this opens the door to novel functionalities, namely *complexity amplification and expansion* [1]. What results is an *all-or-nothing flavor of the Church-Turing hypothesis*: Either no physical computer exists that is able to produce non-Turing-computable data — or even a "device" as simple as a single photon can.

The second idea starts from the observation that the price for the *logical* irreversibility of the erasure transformation comes in the form of a *thermodynamic* effort.[2] In an attempt to harmonize this somewhat *hybrid* picture, we invoke Wheeler's [15] *"It from Bit*: Every *it* — every particle, every field of force, even the spacetime continuum itself — derives its function, its meaning, its very existence entirely [...] from the apparatus-elicited answers to yes-or-no questions, binary choices, *bits*." This is an anti-thesis to Landauer's slogan, and we propose the following synthesis of the two: If Wheeler motivates us to look at the environment as being *information* as well, then Landauer's principle may be read as: The necessary environmental compensation for the logical irreversibility of the erasure of S is such that *the overall computation, including the environment, is logically reversible: no information ever gets completely lost.*

Second law, logical-computational version. *Time evolutions are injective: Nature computes with Toffoli, but no AND or OR gates.*

(Note that this fact is *a priori a*symmetric in time: The future must uniquely determine the past, not necessarily *vice versa*. In case the condition holds also for the reverse time direction, the computation is *deterministic*, and *randomized* otherwise.)

If logical reversibility is a simple computational version of a discretized second law, does it have implications resembling the traditional versions of the law? First of all, it leads to a "Boltzmann-like" form, *i.e.*, the existence of a quantity essentially monotonic in time. More specifically, the logical reversibility of time evolution implies that the Kolmogorov complexity of the global state at time t can be smaller than the one at time 0 only by at most $K(C_t)+O(1)$ if C_t is a string encoding the time span t. The reason is that one possibility of describing the state at time 0 is to give the state at time t, plus t itself; the rest is exhaustive search using only a constant-length program simulating forward time evolution (including possible randomness).

Similarly, logical reversibility also implies statements resembling the version of the second law due to *Clausius*: "Heat does not spontaneously flow from cold to hot." The rationale here is explained with a toy example: If we have a circuit — the time evolution — using only (logically reversible) Toffoli gates, then it is *impossible* that this circuit computes a transformation mapping a pair of strings to another pair such that the Hamming-heavier of the two becomes even heavier whilst the lighter gets lighter. A function accentuating such imbalance instead of lessening it is not injective, as a basic counting argument shows.

Example. Let a circuit consisting of only Toffoli gates map an $N(=2n)$-bit string to another. We consider the map separately on the first and second halves and assume the computed function to be conservative, *i.e.*, to leave the Hamming weight of the full string unchanged at n (conservativity can be seen as some kind of *first* law, *i.e.*, the preservation of a quantity). We look at the excess of 1's in one of the halves (which equals the deficit of 1's in the other). We observe that the probability (with respect to the uniform distribution over all strings of some Hamming-weight couple $(wn, (1-w)n)$) of the *imbalance substantially growing* is exponentially weak. The key ingredient for the argument is the function's injectivity. Explicitly, the probability that the weight couple goes from $(wn, (1-w)n)$ to $((w+\Delta)n, (1-w-\Delta)n)$ — or more extremely —, for $1/2 \leq w < 1$ and $0 < \Delta \leq 1-w$, is

$$\frac{\binom{n}{(w+\Delta)n}\binom{n}{(1-w-\Delta)n}}{\binom{n}{wn}\binom{n}{(1-w)n}} = 2^{-\Theta(n)} \ .$$

[2] Since the amount of the required free energy (and heat dissipation) is proportional to the length of the best compression of the string, the latter can be seen as a *quantification* of the erasure transformation's irreversibility.

Note here that we even get the correct, exponentially weak "error probability" with which the traditional second law can be "violated."

Finally, logical reversibility also implies statements resembling Kelvin's version of the second law: "A single heat bath alone has no work value." This, again, follows from a simple counting argument. There exists no reversible circuit that, for general input environments (with a fixed weight — intuitively: *heat energy*), extracts redundancy, *i.e.*, work value, and concentrates it in some pre-chosen bit positions: *Concentrated* redundancy is *more* of it.

Example. The probability that a fixed circuit maps a "Hamming bath" of length N and Hamming weight w to another such that the first n positions contain all 1's and such that the Hamming weight of the remaining $N - n$ positions is $w - n$ (again, we are assuming conservation here) is

$$\frac{\binom{N-n}{w-n}}{\binom{N}{w}} = 2^{-\Theta(n)} .$$

Discussion. We propose a logical view of the second law of thermodynamics: *the injectivity or logical reversibility of time evolution*. (This is somewhat ironic as the second law has often been related to its exact opposite: *irreversibility*.) It implies, within the Church-Turing view, Clausius-, Kelvin-, and Boltzmann-like statements — including their "failure probabilities."

VI. Conclusion

Using new, constructive results on work extraction and a direct connection between extraction and erasure, we propose *a reformulation of Landauer's principle*, essentially stating that the erasure cost of a string is *not* proportional to its length, but to the one of its *best compression*. We have taken into account the case where the erasing demon possesses initial knowledge about the string to be deleted, in the form of some fixed additional string. We have argued, in the spirit of the *Church-Turing hypothesis*, that the constructive-algorithmic, as opposed to entropic, power of the knowledge is relevant. In this same view, time evolutions are considered to be computed by a Turing machine, and Landauer's principle, when combined with Wheeler's "It from Bit," naturally leads to a simple formulation of the second law of thermodynamics *as a property of this very computation, namely logical reversibility*: It alone implies historical versions (due to Boltzmann, Clausius, or Kelvin) of the law.

Acknowledgments. We thank Mateus Araújo, Veronika Baumann, Ämin Baumeler, Charles Bédard, Gilles Brassard, Harvey Brown, Časlav Brukner, Bora Dakic, Paul Erker, Jürg Fröhlich, Nicolas Gisin, Arne Hansen, Marcus Huber, Alberto Montina, Benno Salwey, Martin Schüle, Andreas Winter, and Magdalena Zych for enlightening discussions.

The author is supported by the Swiss National Science Foundation (SNF) and the National Centre of Competence in Research "Quantum Science and Technology" (QSIT), and the *Hasler Foundation*.

References

[1] Ä. Baumeler, C. Bédard, G. Brassard, and S. Wolf, Kolmogorov amplification from Bell correlation, submitted to *International Symposium in Information Theory (ISIT) 2017*, 2017.

[2] J. S. Bell, On the Einstein-Podolsky-Rosen paradox, *Physics*, Vol. 1, pp. 195–200, 1964.

[3] C. H. Bennett, The thermodynamics of computation, *International Journal of Theoretical Physics*, Vol. 21, No. 12, pp. 905–940, 1982.

[4] C. H. Bennett, Logical reversibility of computation, *IBM J. Res. Develop.*, Vol. 17, No. 6, pp. 525–532, 1982.

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

[6] G. Chaitin, A theory of program size formally identical to information theory, *Journal of the ACM*, Vol. 22, pp. 329–340, 1975.

[7] O. Dahlsten, R. Renner, E. Rieper, and V. Vedral, The work value of information, *New J. Phys.*, Vol. 13, 2011.

[8] L. del Rio, J. Åberg, R. Renner, O. Dahlsten, and V. Vedral, The thermodynamic meaning of negative entropy, *Nature*, 2011.

[9] P. Faist, F. Dupuis, J. Oppenheim, and R. Renner, The minimal work cost of information processing, *Nature Communications*, 2015.

[10] E. Fredkin and T. Toffoli, Conservative logic, *International Journal of Theoretical Physics*, Vol. 21, No. 3–4, pp. 219-253, 1982.

[11] A. Kolmogorov, Three approaches to the quantitative definition of information, *Problemy Peredachi Informatsii*, Vol. 1, No. 1, pp. 3–11, 1965.

[12] R. Landauer, Irreversibility and heat generation in the computing process, *IBM Journal of Research and Development*, Vol. 5, pp. 183–191, 1961.

[13] R. Landauer, Information is inevitably physical, *Feynman and Computation 2*, 1998.

[14] L. Szilárd, Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen (On the reduction of entropy in a thermodynamic system by the intervention of intelligent beings), *Zeitschrift für Physik*, Vol. 53, pp. 840–856, 1929.

[15] J. A. Wheeler, Information, physics, quantum: the search for link, *Proceedings III International Symposium on Foundations of Quantum Mechanics*, pp. 354–368, 1989.

[16] S. Wolf, Non-locality without counterfactual reasoning, *Phys. Rev. A*, Vol. 92, No. 052102, 2015.

[17] J. Ziv and A. Lempel, Compression of individual sequences via variable-rate coding, *IEEE Transactions on Information Theory*, Vol. 24, No. 5, p. 530, 1978.

[18] W. H. Zurek, Algorithmic randomness and physical entropy, *Phys. Rev. A*, Vol. 40, No. 8, 1989.