On Bits As Fuel

Stefan Wolf[1]

[1]*Faculty of Informatics, Università della Svizzera italiana (USI), Via G. Buffi 13, 6900 Lugano, Switzerland*

The converse of Landauer's principle states that (the physical representation of) certain bit-strings (*e.g.*, 0^N) can be used to extract work from the environment. To the general question what the work value of a classical string $S \in \{0,1\}^N$ is, there have been different answers invoking algorithmic *complexity* and conditional *entropy* about S from the extractor's viewpoint, respectively. We harmonize the results, in particular by making explicit what "knowing S" can mean. We find that the work value of S does not depend on some probability distribution around it, but only on two fixed strings: S as well as the initial state of the device carrying out the work-extraction computation. The context-free view extends to the notion of a macrostate, the second law of thermodynamics, randomness, and quantum non-locality.

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 a bit of information costs an amount of at least $kT \ln 2$ of free energy which is then dissipated as heat to the environment (of temperature T). Here, "erasing a bit" stands for "forcing the corresponding binary degree of freedom into the state 0;" the fact has been derived by Landauer from the second law of thermodynamics: The reduction of entropy due to the loss of the binary degree of freedom must be compensated by an increase, of at least the same amount, in environmental entropy.

Landauer's principle led the way to the solution of the famous problem of *Maxwell's demon* [4]: If the "demon's brain" is taken into account, then the order she creates *outside* is reflected by the disorder appearing *inside* her — the latter in the form of complex data depending on the demon's observations and in the end represented within her internal state since they are required for guiding her actions throughout the sorting procedure. If we assume that the demon's (N-bit) memory is in the initial state 0^N, then one can regard this 0-string, which is "used up" in the end, as the resource required by the demon for her order-creating action. Indeed, the *converse of Landauer's principle* states that the string 0^N — more precisely, a physical representation of it[1] — has a *fuel value* of $kTN \ln 2$: It allows for transforming this amount of environmental heat into work. The situation has also been discussed by Szilárd [14] if the memory cell is a single-molecule gas.

We address the question what in general the fuel value is of (a physical representation of) a classical string S. Since the *reversible* extraction of the string 0^N from S is equivalent to the gain of free energy of $kTN \ln 2$, we have a first answer: *Work extraction is data compression.*

II. WORK EXTRACTION: STATE OF THE ART

A. The Results by Bennett and by Zurek

Bennett [3] claimed the fuel 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]:

$$W(S) = (\mathrm{len}(S) - K(S))kT \ln 2\,.$$

Bennett's argument is that (the physical representation of) S can be — logically, hence, thermodynamically [9] — 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).

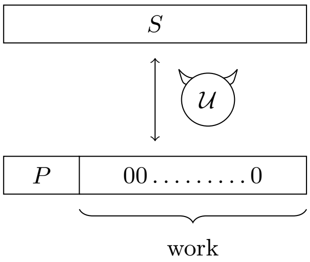


Figure 1. Bennett's argument

It was already pointed out by Zurek [17] that whereas 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 *not.* This means that the "demon" that can carry out the work-extraction computation on S from scratch does not exist under the Church-Turing hypothesis. We will see, however, that Bennett's value represents an *upper bound* on the fuel value of S.

[1] The string 0^N is, physically, special here in the sense that there exists a "constant-size" machine, including the program, that can actually generate it. This is important to note since *a priori*, the semantics — which of the two states is 0, which is 1? — of the different bit positions is arbitrary.

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

Dahlsten *et al.* [8] follow Szilárd [14] in putting the *knowledge* of the demon extracting the work to the center of their attention. More precisely, they claim

$$W(S) = (\mathrm{len}(S) - D(S))kT \ln 2\,,$$

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.

The work [8] does not consider the algorithmic aspects of the demon's actions extracting the free energy, but the effect of the demon's *a priori knowledge on S*. If we model the demon as an algorithmic apparatus, then we should specify the *form* of that knowledge explicitly. For instance, vanishing conditional entropy means that S is *uniquely determined* from the demon's viewpoint. Does this now mean that the demon possesses a *copy* of S, or the *ability* to produce such a copy, or pieces of *information* that uniquely determine S? This question 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 complexity even given her internal state (an example see below). In this case, the first result claims $W(S)$ to be 0, whereas $W(S) \approx \mathrm{len}(S)$ according to the second. The gap vanishes if, *e.g.*, "knowing S" is understood in a constructive — as opposed to entropic — sense, meaning that "the demon has a copy of S represented in her internal state:" If that copy is included in Bennett's reasoning, then his result reads

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

III. WORK EXTRACTION IS DATA COMPRESSION (WITH HELPER)

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 we assume to be sufficiently long for the tasks and inputs in question, but *finite*. The tape initially contains S, the string the fuel 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 [9]. Logical reversibility in our model is the ability of the same demon to carry out the backward computation, *i.e.*, from $P||X$ to $S||X$.[2] We denote by $E(S|X)$ the *maximal amount of* 0*-bits extractable logically reversibly from S given the knowledge X*, *i.e.*,

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

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

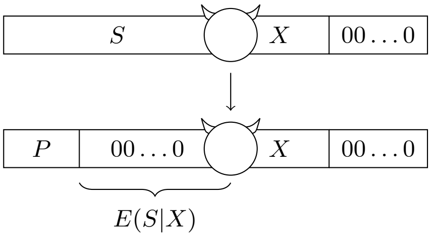


Figure 2. The model

According to the above, the work value of any physical representation of S for a demon knowing X is

$$W(S|X) = E(S|X)kT \ln 2\ .$$

We derive bounds on $E(S|X)$.

B. Lower Bound

Let C be a computable function

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

with $\mathrm{len}(C(A,B)) \le \mathrm{len}(A)$ and such that

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

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

$$E(S|X) \ge \mathrm{len}(S) - \mathrm{len}(C(S,X))\,.$$

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 and where no further input or output bits are involved. This can be achieved by first implementing all logical gates with Toffoli gates and uncomputing all

[2] Note that this is the natural way of defining logical reversibility in our setting with a *fixed* input and output but *no sets nor bijective maps* between them.

junk [4] in both of the circuits. The resulting two circuits have now both still the property that the input is part of the output. As a second step, we can simply combine the two, where the first circuit's first output becomes the second's second input, and *vice versa.* 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 output, on the bit positions carrying the input previously. (This circuit is roughly as efficient as the less efficient of the two irreversible circuits for the compression and for the decompression, respectively.) We assume this reversible circuit to be hard-wired in the demon's head. A typical example for a compression algorithm that can be used is Ziv-Lempel [16].

C. Upper Bound

We have the following upper bound on $E(S|X)$:

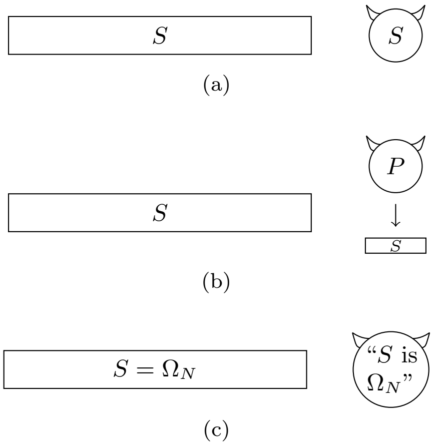
$$E(S|X) \leq \text{len}(S) - K_{\mathcal{U}}(S|X) \, ,$$

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, our 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. Then demon can then bit-wisely XOR this new copy of S to the original S on the tape, hereby producing $0^{\text{len}(S)}$ *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 the 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 $0^{\text{len}(S)}$: This results in a maximal fuel value matching the (in this case trivial) upper bound. Note that this is in harmony with [8] if this is how vanishing conditional entropy is established.

D. Description Complexity and Ω

We contrast our bounds with the entropy-based results of [8]: According to the latter, a demon *having complete knowledge of S* is able to extract maximal work: $E(S) \approx \text{len}(S)$. *What means "knowing S"?* (see Figure 3). The results are in accordance with ours if the demon's *knowledge* consists of (a) a *copy* of S, or at least of (b) its *ability to algorithmically reconstruct* S, based on a known program P, as discussed above. It is, however, possible (c) that the demon's knowledge is of different nature, merely *determining S uniquely without providing the ability to build S*. For instance, let the demon's knowledge about S be: "S equals the first N bits Ω_N of the binary expansion of Ω." Here, Ω is the so-called halting probability [5] of a fixed universal Turing machine (*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, there is no *set of instructions shorter than S enabling the demon to generate S* — which would be required for work extraction from S according to our upper bound. In short, the gap reflects the gap between the *"unique-description complexity"*[3] and the *Kolmogorov complexity.*



(a)

(b)

(c)

Figure 3. Knowing S

E. Work Value and Macrostate

Let us investigate the connection between work extraction from a microstate S and the corresponding *macrostate* $M(S)$. Whereas the meaning of the latter notion is clear in thermodynamics in the case of equilibrium states or with respect to a fixed coarse-graining, it is less obvious how to define it for general S.

It has already been observed that the notion of Kolmogorov complexity can allow, in principle, for *thermodynamics independent of probabilities or ensembles.* Zurek [17] defines physical entropy H_p to be

$$H_p(S) := K(M) + H(S|M) \ ,$$

where M stands for the collected data at hand, and $K(M)$ for their most compressed description, while

[3] Note that a diagonal argument, called *Berry paradox*, shows that the notion of "description complexity" cannot be defined generally for all strings.

$H(S|M)$ is the remaining conditional Shannon entropy of the microstate S, given M. That definition of a macrostate it *subjective* since it depends on the *available* data. How instead can the macrostate — and *entropy*, for that matter — be defined objectively? We propose to use the *Kolmogorov sufficient statistics* [10] of the microstate: For any $k \in \mathbf{N}$, let M_k be the smallest set such that $S \in M_k$ and $K(M_k) \leq k$ hold. Let further k_0 be the value of k where the function $\log|M_k|$ becomes linear with slope -1. Intuitively speaking, k_0 is the point from which on there is no more "structure" to be exploited for describing S within M_{k_0}. We define $M(S) := M_{k_0}$ to be *S's macrostate.* It yields a program generating S of minimal length

$$K(S) = k_0 + \log|M_{k_0}| = K(M(S)) + \log|M(S)|\,.$$

The fuel value of a string $S \in \{0,1\}^N$ is now related to the macrostate $M(S) \ni S$ by

$$E(S) \leq N - K(M(S)) - \log|M(S)| = \log(|S|/|M(S)|) - k_0$$

(see Figure 4): Decisive is neither the complexity of the macrostate nor its log-size *alone*, but their *sum.*

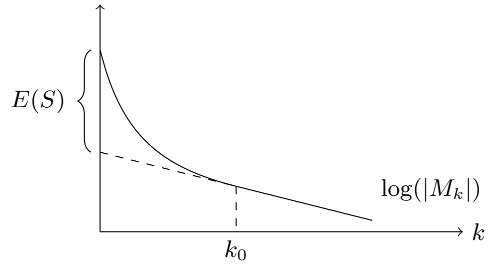


Figure 4. Kolmogorov sufficient statistics, macrostate, and fuel value

A notion defined in a related way is the *sophistication* or *interestingness* as discussed by Aaronson [1] investigating the process where milk is poured into coffee (see Figure 5). Whereas the initial and final states are "simple" and "uninteresting," the intermediate (non-equilibrium) states display a rich (coarse-grained) structure; here, the sophistication — and also $K(M)$ for our macrostate M — becomes maximal.

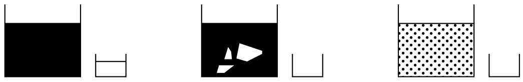


Figure 5. Coffee and milk

During the process under consideration, neither the macrostate's complexity nor its size is monotonic in time: Whereas $K(M)$ has a *maximum* in the non-equilibrium phase of the process, $\log|M|$ has a *minimum* there (see Figure 6). On the other hand, the complexity of the *microstate* itself,

$$K(S) = K(M) + \log|M|\;,$$

is a candidate for a monotonically nondecreasing quantity: Is this *entropy*, and is its monotonicity the *second law of thermodynamics* in that view? That law is guaranteed to hold under the assumption of *logical reversibility*: The future contains (essentially) all the complexity of the past if we can, step by step, reconstruct the latter from the former.

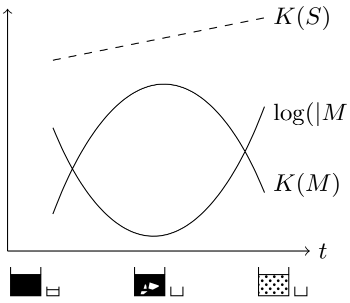


Figure 6. The complexity and the size of the macrostate

IV. CONCLUSIONS

In an attempt to reconcile two groups of statements on the fuel value of a string S, we suggest that this quantity be given by the *difference of the length of S and of its compression*, given the complete knowledge (initial state) of the extraction device. We understand the work-extraction process algorithmically, to be carried out by a Turing machine. The *Church-Turing hypothesis* stating that *all* natural processes can be so seen motivates this view; a similar perspective can been used with respect to Bell inequalities [15].

When one replaces the *entropy* of a probability distribution by a context-independent *complexity*, the second law of thermodynamics reads: *The complexity of the microstate of a closed system does not decrease.* This is equivalent to saying that the *fuel value in a closed system does not increase.* Here, the circle is closed; we find the Landauer principle we started from: If a complex substring of the microstate becomes simple, then in another part of it, there must be redundancy (*i.e.*, low complexity or: free energy) disappearing into complexity (*i.e.*, "heat").

If the second law is the fact that the complexity of the microstate is non-decreasing, then the law automatically holds for all logically *reversible* processes: The past cannot have been significantly more complex than the future if the latter allows for reconstructing the former in principle.

While we propose to view the second law as a statement about microstates, it does have an implication for macrostates *as long as their "descriptions" are simple* — which is the case, *e.g.*, if the macrostate is of the usual thermodynamic kind such as "a gas of N particles is in

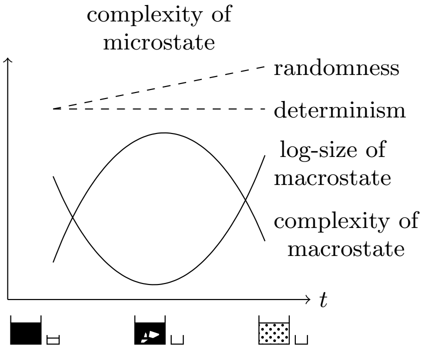


Figure 7. Determinism vs. randomness

a volume V at temperature T," etc.: Then, logical reversibility implies that the macrostate does (virtually) *not shrink* in time. In this, we can recognize the traditional view of the second law. Note, however, that under the reversibility condition, this consequence holds *always*, while the second law is compatible with the entropy shrinking by N with probability 2^{-N}. The gap reflects the fact that the thermodynamical and computational notions of a macrostate can differ: A gas state may look unsuspicious (large thermodynamic microstate) in the traditional view although, *e.g.*, the positions of its particles are simply encoded in the expansion of π (very small computational macrostate). Which law is now fundamental, which is emergent? (It is fair to assume that the one that is valid *without exception* underlies the other?)

Ironically, the computational second law follows from *reversibility* of the computation, whereas the thermodynamical second law is usually linked to its exact opposite: *irreversibility.*

Given logical reversibility and, hence, the validity of the second law, *determinism* denotes the fact that the complexity is essentially constant in time, whereas *randomness* is an increase of that quantity (see Figure 7). Alternative definitions of (freeness of) randomness are based on a given spacetime structure, such as in [7]. (Note that this definition is compatible with full determinism in the probabilistic picture. The additional condition that "something which did not happen could have" is difficult to formalize.) It has been proposed to start from freeness as fundamental instead, and to find a resulting causal structure from there [2].

ACKNOWLEDGMENTS

We thank Mateus Araújo, Ämin Baumeler, Harvey Brown, Časlav Brukner, Bora Dakic, Paul Erker, Jürg Fröhlich, Nicolas Gisin, Arne Hansen, Marcus Huber, Alberto Montina, Benno Salwey, Andreas Winter, and Magdalena Zych for enlightening discussions.
The author is supported by the Swiss National Science Foundation (SNF), the National Centre of Competence in Research "Quantum Science and Technology" (QSIT), and the COST action on Fundamental Problems in Quantum Physics. We thank all members of the *Facoltà indipendente di Gandria* for their warm support.

[1] S. Aaronson, http://www.scottaaronson.com/blog/?p=762, 2011.

[2] Ä. Baumeler, A. Feix, and S. Wolf, Maximal incompatibility of locally classical behavior and global causal order in multi-party scenarios, *Phys. Rev. A*, Vol. 90, No. 042106, 2014.

[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.

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

[6] R. Cilibrasi and P. Vitányi, Clustering by compression, *IEEE Transactions on Information Theory*, Vol. 51, No. 4, 1523–1545, 2005.

[7] R. Colbeck and R. Renner, No extension of quantum theory can have improved predictive power, *Nature Communications*, Vol. 2, 411, 2011.

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

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

[10] P. Gàcs, J. T. Tromp, and P. M. B. Vitányi, Algorithmic statistics, *IEEE Transactions on Information Theory*, Vol. 47, No. 6, 2001.

[11] A. N. 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] S. Wolf, Non-locality without counterfactual reasoning, arXiv:1505.07037, 2015.

[16] 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.

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