Free energy of a general computation

Ämin Baumeler[1,2] and Stefan Wolf[2,3]

[1]Institute for Quantum Optics and Quantum Information (IQOQI), Austrian Academy of Sciences, Boltzmanngasse 3, 1090 Vienna, Austria

[2]Facoltà indipendente di Gandria, Lunga scala, 6978 Gandria, Switzerland

[3]Faculty of Informatics, Università della Svizzera italiana, Via Buffi 13, 6900 Lugano, Switzerland

(Received 4 July 2019; published 15 November 2019)

Starting from Landauer's slogan "information is physical," we revise and modify Landauer's principle stating that the erasure of information has a minimal price in the form of a certain quantity of free energy. We establish a direct link between the erasure cost and the work value of a piece of information and show that the former is essentially the length of the string's best compression by a reversible computation. We generalize the principle by deriving bounds on the free energy to be invested for—or gained from, for that matter—a general computation. We then revisit the second law of thermodynamics and compactly rephrase it (assuming the Church-Turing-Deutsch hypothesis that physical reality can be simulated by a universal Turing machine): Time evolutions are logically reversible—"the future fully remembers the past (but not necessarily vice versa)." We link this view to previous formulations of the second law, and we argue that it has a particular feature that suggests its "logico-informational" nature, namely, *simulation resilience*: If a computation faithfully simulates a physical process violating the law, then that very computation procedure violates it as well.

DOI: 10.1103/PhysRevE.100.052115

I. INTRODUCTION

In 1961, Rolf Landauer famously stated "Information is physical" [1]: Despite the success of Shannon's making information an abstract concept (that can be viewed and understood independently of its particular physical realization), Landauer—while not questioning the power of that abstract view—recalls that all information storing, treatment, and transmission is ultimately a physical process and, thus, subject to physical laws. A specific law relevant in this context is the second law of thermodynamics. Its consequence for information processing has been called *Landauer's principle* [2]: "The erasure of N bits of information costs at least an amount of $NkT \ln 2$ (k being Boltzmann's constant) of free energy that must be dissipated as heat into the environment of temperature T." (Note that this heat dissipation is crucial for the argument: It represents the compensation required for avoiding a violation of the second law despite an entropy decrease in the memory device through the erasure process.) Conversely, erased strings have a *work value* (see, e.g., Refs. [3,4]): By, e.g., encoding an erased bit string of length N in the particle's position of a gas within a box, where the particle's position is on the left half for the value 0 and on the right half otherwise, and by placing a piston in the center, $NkT \ln 2$ of free energy can be extracted from the environment, "randomizing" the original string.

In this article, we modify and generalize Landauer's principle in the following respects: First, it is claimed that the erasure cost is proportional not to the length of the string to be erased, but of its best compression—given the entire knowledge of the erasure device (Sec. IV). We obtain these results from the bounds we acquire on the work value of information (Sec. III) and a direct connection between erasure cost and work value of any piece of information. Second, we generalize these results to a lower bound on the free-energy cost, or value, of a general computation (Sec. V). Our findings are modifications of known results (see e.g., Refs. [4–7]) to the constructive setting—where all involved processes are imagined to be carried out by a Turing machine. Furthermore, we give a lower bound on the free-energy gain possible from certain computations—a bound that matches the cost of the inverse computation. We look at the use of the erasure cost as an intrinsic randomness measure in the context of quantum correlations (Sec. VI). Having these results at hand, we finish by proposing a computational version of the second law of thermodynamics (Sec. VII). This comes with a speculation about which of its trait is the reason why such a version exists in the first place. Candidates are its "encoding independence" and "simulation resilience:" If a computation simulates a process violating the second law, then that computing procedure cannot be closed but must dissipate "junk" bits onto the other parts of the tape or heat into the environment. Thus, the violation of the law by a process carries over to its simulation. The reason is that a degree of freedom is represented by a degree of freedom.

II. WORK VALUE: STATE OF THE ART

While other results (see, e.g., Refs. [8–13]) on the work value of information focus on using information reservoirs to generate energy flows, the below described results and this article focus on the work value of information—being in form of random variables or bit strings—*per se*. As opposed to discussing the role of information in thermodynamic processes, we discuss the thermodynamics processes of information.

A. Bennett's view

Bennett [5] claimed the work value of a string S, $\mathcal{W}(S)$, to be proportional to the difference between its length, $\ell(S)$, and the length of the shortest program that produces S. The latter is called the Kolmogorov complexity of S, denoted by $K(S)$ [14]. Expressed mathematically, this amounts to

$$\mathcal{W}(S) = [\ell(S) - K(S)]kT \ln 2. \tag{1}$$

Bennett's argument is that S can be logically, hence, thermodynamically [15] reversibly mapped to the string $P||000\cdots 0$, where the symbol $||$ denotes concatenation, and P is the shortest program generating S. The length of the generated 0-string is $\ell(S) - K(S)$.

It was already pointed out by Zurek [16] that while 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 (if the Church-Turing hypothesis is true); the Kolmogorov complexity is an uncomputable value. We will see, however, that Bennett's value is an upper bound on the work value of S. Bennett also links the string's erasure cost to its probabilistic entropy [17].

B. View of Dahlsten *et al.*

Dahlsten *et al.* [4,6] follow Szilárd [3] in putting the knowledge of the demon extracting the work at the center of their attention. More precisely, they claim $\mathcal{W}(S) = \ell(S) - D(S)$, 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 its ignorance. Building on these results and in the same probabilistic spirit, the cost of erasure [18] as well as of general computations [7] have been linked to entropic expressions of (conditional) probability distributions.

III. WORK EXTRACTION AS DATA COMPRESSION

In the following, we model work extraction to be an algorithm executed by a "demon with knowledge."

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 0s 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 $0^{\ell(S)-\ell(P)}$, whereas the rest of the tape is (again) the same as before the work extraction. The operations are logically reversible and can, hence, be carried out thermodynamically reversibly [15]. 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 $\mathcal{W}(S|X)$ the maximal length of an all-0-string extractable logically reversibly from S, given the knowledge X, times $kT \ln 2$, i.e.,

$$\mathcal{W}(S|X) := [\ell(S) - \ell(P)]kT \ln 2 \tag{2}$$

if P's length is minimal.

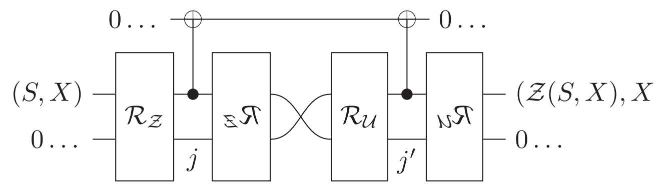

FIG. 1. Schematic circuit of thermodynamically neutral compression with helper. The circuit $\mathcal{R}_{\mathcal{Z}}$ implements the compression algorithm $\mathcal{Z}$ with Toffoli gates only, $\reflectbox{$\mathcal{R}_{\mathcal{Z}}$}$ is the same circuit in reverse order. Then again, the circuit $\mathcal{R}_{\mathcal{U}}$ implements the corresponding decompression algorithm with Toffoli gates only, and $\reflectbox{$\mathcal{R}_{\mathcal{U}}$}$ is its reverse. The symbols j and j' represent the "junk" that arises from implementing the circuits with Toffoli gates only.

B. Lower bound

We show that every specific data-compression algorithm leads to a lower bound on the extractable work: Let $\mathcal{Z}$ be a computable function

$$\mathcal{Z} : \{0, 1\}^* \times \{0, 1\}^* \longrightarrow \{0, 1\}^* \tag{3}$$

such that

$$(A, B) \mapsto (\mathcal{Z}(A, B), B) \tag{4}$$

is injective.[1] We call $\mathcal{Z}$ a data-compression algorithm with helper. Then we have

$$\mathcal{W}(S|X) \geqslant [\ell(S) - \ell(\mathcal{Z}(S, X))]kT \ln 2. \tag{5}$$

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

$$A||B \;\mapsto\; \mathcal{Z}(A, B)||0^{\ell(A)-\ell(\mathcal{Z}(A,B))}||B \tag{6}$$

is computable and bijective. From the two (possibly irreversible) circuits which compute the compression and its inverse, one can obtain a reversible circuit for 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" [19] in both circuits. The resulting two circuits now still have 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 (see Fig. 1). The combined circuit has only the compressed data plus the 0s 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.) A typical example for an algorithm that can be used here is universal data compression *à la* Ziv-Lempel [20].

C. Upper bound

We have the following upper bound on the extractable work:

$$\mathcal{W}(S|X) \leqslant [\ell(S) - K_{\mathcal{U}}(S|X)]kT \ln 2, \tag{7}$$

[1] The set $\{0, 1\}^*$ is the set of all finite but arbitrarily long bit strings.

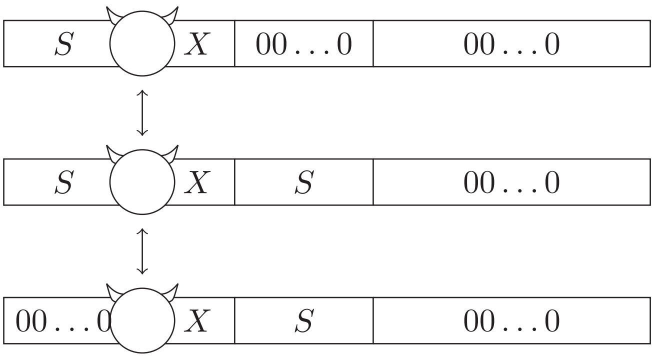

FIG. 2. The demon uses X as program to produce a second copy of S, which thereafter is used to generate $0^{\ell(S)}$ via the reversible operation $S \oplus S$ (bitwise addition modulo 2).

where $K_{\mathcal{U}}(S|X)$ is the conditional Kolmogorov complexity (with respect to the universal Turing machine $\mathcal{U}$) 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 extraction demon is able only to carry out the computation in question (logically, hence, thermodynamically) reversibly if it 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 $K_{\mathcal{U}}(S|X) \approx 0$. The latter means that X itself can be seen as a program for generating an additional copy of S. The demon can then bitwisely XOR this extra S to the original S (to be work-extracted) on the tape, hereby producing $0^{\ell(S)}$ reversibly to replace the original S, at the same time saving the new one, as reversibility demands (see Fig. 2). When Bennett's "uncomputing trick" is used—allowing any computation by a Turing machine to be made logically reversible [19]—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 $0^{\ell(S)}$: This results in a maximal work value matching the (in that case trivial) upper bound.[2]

[2]Let us compare our bounds with the entropy-based results of Refs. [4,6]: According to the latter, a demon knowing S entirely is able to extract maximal work: $\mathcal{W}(S) \approx \ell(S)kT \ln 2$. What does it mean to "know S"? The knowledge can consist of (a) a copy of S or of (b) the 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. The constructive as opposed to the entropic groups of results are in accordance in cases (a) and (b) but 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," where, Ω is the so-called halting probability [21] 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: $K_{\mathcal{U}}(\Omega_n\,|$ "It is bits 1–n of TM $\mathcal{A}$'s halting probability"$)\approx n$: No work is extractable according to our upper bound. Intuitively, this gap opens up whenever the "description complexity" is smaller than the Kolmogorov complexity. (Note that a self-reference argument, called the Berry paradox, shows that the notion of "description complexity" is problematic and can never be defined consistently for all strings.)

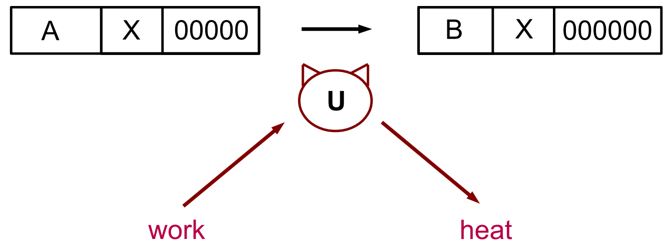

FIG. 3. The energy cost of a general computation.

IV. REVISING LANDAUER'S PRINCIPLE

Here we revise Landauer's principle to give a lower and an upper bound on the erasure cost.

A. Connection to work value

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

$$\mathcal{W}(S|X) + \mathcal{C}_{\text{erasure}}(S|X) = NkT \ln 2. \tag{8}$$

To see this, consider first the combination extract-then-erase. In the extraction process we gain $\mathcal{W}(S|X)$ of free energy and consequently have to erase N bits. Since this is one specific way of erasing, we have

$$\mathcal{C}_{\text{erasure}}(S|X) \leqslant NkT \ln 2 - \mathcal{W}(S|X). \tag{9}$$

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

$$\mathcal{W}(S|X) \geqslant NkT \ln 2 - \mathcal{C}_{\text{erasure}}(S|X). \tag{10}$$

We spend $\mathcal{C}_{\text{erasure}}(S|X)$ of free energy to erase the string and use all of the string as "fuel."

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 $\mathcal{Z}$ be a computable function, $\mathcal{Z} : \{0, 1\}^* \times \{0, 1\}^* \longrightarrow \{0, 1\}^*$, such that $(A, B) \mapsto (\mathcal{Z}(A, B), B)$ is injective. Then we have

$$\begin{aligned} K_{\mathcal{U}}(S|X)kT \ln 2 &\leqslant \mathcal{C}_{\text{erasure}}(S|X) \\ &\leqslant \ell(\mathcal{Z}(S, X))kT \ln 2. \end{aligned} \tag{11}$$

The first inequality follows from Eq. (8) in combination with the upper bound (7), the second from Eq. (8) and the lower bound (5).

V. GENERALIZING LANDAUER'S PRINCIPLE

Erasure as well as work extraction can be seen as special cases of a computation with a given input and an output. Here we generalize Landauer's principle and discuss the work cost and work value of a general computation; i.e., we generalize the already obtained bounds on the cost (minimal amount of free energy that has to be used) and value (maximal amount of free energy that can be gained) to general computation. Assume that a (universal) Turing machine performs a computation such that the initial content of the tape is A and X (plus a corresponding finite number of 0s) and the final state is B and X (where X can be seen, again, as a "catalyst"). Depending on A, B, and X, this computation can have a work cost or value, respectively. If it has some work cost, then the party performing the computation has to invest free energy that will be dissipated as heat to the environment during the computation. In the case that the computation has some value, heat from the environment is transformed to free energy.

A. The energy cost of a general computation

The following result is an algorithmically constructive modification of entropic results [7] and a generalization of less constructive but also complexity-based claims [22].

Work cost of a general computation. (see Fig. 3) Let $\mathcal{Z}$ be a computable function, $\mathcal{Z}:\{0,1\}^*\times\{0,1\}^*\to\{0,1\}^*$, such that $(V,W)\mapsto(\mathcal{Z}(V,W),W)$ is injective. Assume that the Turing machine $\mathcal{U}$ carries out a computation such that A is its initial state, C_1 the first intermediate state, C_2 the second, etc., up to C_n, and B is the final state. Then the energy cost of this computation with side information X (always on the tape), $\mathcal{C}_\mathcal{U}(A\to_{\{C_i\}}B\,|\,X)$, is at least

$$\mathcal{C}_\mathcal{U}\big(A\to_{\{C_i\}}B\,\big|\,X\big)\ \geqslant\ kT\ln 2\Big\{K_\mathcal{U}(A|X)-\sum_{i=1}^{n}[\ell(\mathcal{Z}(C_i,X))-K_\mathcal{U}(C_i|X)]-\ell(\mathcal{Z}(B,X))\Big\}. \quad (12)$$

Proof. Let us consider the computation from (A,X) to (C_1,X). According to the above [see expression (11)], the erasure cost of A, given X, is at least $K_\mathcal{U}(A|X)kT\ln 2$. One possibility of realizing this complete erasure of A is to first transform it to C_1 (given X), and then erase C_1—at cost at most $\ell(\mathcal{Z}(C_1,X))kT\ln 2$. Therefore, the cost to get from A to C_1 given X cannot be lower than the difference between $K_\mathcal{U}(A|X)kT\ln 2$ and $\ell(\mathcal{Z}(C_1,X))\cdot kT\ln 2$. The statement follows by summing all contributions of the individual computing steps. *QED.*

Note that if no intermediate results are specified, the bound simplifies to

$$\mathcal{C}_\mathcal{U}(A\to B\,|\,X)\geqslant[K_\mathcal{U}(A|X)-\ell(\mathcal{Z}(B,X))]kT\ln 2 \quad (13)$$

(see also Ref. [23]).

B. The energy value of a general computation

We consider the work value of a computation from A to B, given X. More specifically, this is, a computation that starts with (A,X) and finishes with (B,X), where B is freely choosable by the computation among all strings with a given complexity $K_\mathcal{U}(B|X)$. The work value is denoted by $\mathcal{W}_\mathcal{U}(A\to B\,|\,X)$, and it is bounded from below as follows.

Work value of a general computation. Let $\mathcal{Z}$ be a computable function, $\mathcal{Z}:\{0,1\}^*\times\{0,1\}^*\to\{0,1\}^*$, such that $(V,W)\mapsto(\mathcal{Z}(V,W),W)$ is injective. The work value of a computation from A to B, given X, is bounded from below by

$$\mathcal{W}_\mathcal{U}(A\to B\,|\,X)\ \geqslant\ [K_\mathcal{U}(B|X)-\ell(\mathcal{Z}(A,X))]\,kT\ln 2. \quad (14)$$

Proof. The cost of erasing A, given X, is at most $\ell(\mathcal{Z}(A,X))\,kT\ln 2$ (see expression (11)). We use a stretch of the resulting all-0-string of some length N for gaining $NkT\ln 2$ free energy. The resulting string of length N is then used as a program for the universal Turing machine $\mathcal{U}$, with additional input X, and where the computation is made logically reversible using Bennett's "uncomputing" trick [19]; let B be the resulting string. Then $K_\mathcal{U}(B|X)\leqslant N$. *QED.*

C. Combination

Consider the following "circular computation," given X:

$$A\longrightarrow B\longrightarrow A. \quad (15)$$

The free-energy gain of computing B from A is at least

$$[K_\mathcal{U}(B|X)-\ell(\mathcal{Z}(A,X))]\,kT\ln 2, \quad (16)$$

whereas the cost for computing A back from B is at least this same amount. The identity of the two bounds is not very surprising; it implies that the bound on the work value is "at least as tight" as the one for the cost of the inverse computation, since otherwise a *perpetuum mobile* of the second kind results.

VI. "RANDOMNESS" AND QUANTUM CORRELATIONS

Landauer's revised principle suggests that the erasure cost of a piece of information is an intrinsic, context-free, physical measure for its randomness independent of probabilities and counterfactual statements (that "some value could just as well have been different").[3] This can be tested in a context in which randomness is central: Bell correlations [24] predicted by quantum theory. In a proof of principle, it was shown [25] that in essence, a similar mechanism as in the probabilistic setting arises: If the correlation is nonlocal, the inputs are incompressible, and nonsignaling holds, then the outputs must be highly complex as well.

Before we describe some of the findings of Ref. [25] in more detail, we introduce the required notation. For an infinite string $a=(a_1,a_2,\dots)$, we define its "truncation" $a_{[n]}:=(a_1,a_2,\dots,a_n,0,0,\dots)$: the string a where all symbols after the nth are set to 0. The expressions $K(a)$ and $K(a\mid b)$, where b is an infinite string as well, denote the functions

$$K(a):\mathbb{N}\to\mathbb{N},\text{ with }\quad n\mapsto K(a_{[n]}), \quad (17)$$

[3] Moreover, such a point of view allows one to discuss randomness on operational grounds.

$$K(a \mid b) : \mathbb{N} \to \mathbb{N}, \text{ with } \quad n \mapsto K(a_{[n]} \mid b_{[n]}). \tag{18}$$

An incompressible string a has the property

$$K(a) \approx n :\Longleftrightarrow \lim_{n\to\infty} \frac{K(a_{[n]})}{n} = 1, \tag{19}$$

and a computable string a the property

$$K(a) \approx 0 :\Longleftrightarrow \lim_{n\to\infty} \frac{K(a_{[n]})}{n} = 0. \tag{20}$$

Intuitively, the shortest program that prints an incompressible string consists of that very same string, and the shortest program that prints a computable string has a constant length. Moreover, we say two functions f and g mapping natural numbers to natural numbers, where $g \not\approx 0$, satisfy $f \approx g$ if and only if

$$\lim_{n\to\infty} \frac{f(n)}{g(n)} = 1. \tag{21}$$

Having this notation at hand, the result stated above is the following. Let (a, b, x, y) be a tuple of four infinite binary strings, where

(1) the PR-box condition [26] is satisfied:

$$x_i \oplus y_i = a_i b_i \text{ for all } i \in \mathbb{N}, \tag{22}$$

(2) both "input" strings a and b are independent and incompressible: $K(a, b) = K(a) + K(b) \approx 2n$,

(3) the no-signaling condition is satisfied:

$$K(x \mid a) \approx K(x \mid a, b), \text{ and} \tag{23}$$

$$K(y \mid b) \approx K(y \mid a, b). \tag{24}$$

It follows from these conditions that the "output" strings are not computable—even if conditioned on the "input":

$$K(x \mid a) = \Theta(n) \text{ and } K(y \mid b) = \Theta(n). \tag{25}$$

While this stated result assumes the existence of correlations not attainable by quantum means [27], the same article proves an analogous statement for quantum correlations, e.g., for the quantum violations of the chained Bell inequalities [28,29].

These results allow 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 potentially opens the door to novel functionalities, namely, complexity amplification and expansion [30]. What results is an all-or-nothing flavor of the Church-Turing hypothesis [31]: 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.

VII. THE SECOND LAW AS LOGICAL REVERSIBILITY

In Landauer's principle, the price for the logical irreversibility of the erasure transformation comes in the form of a thermodynamic effort. (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.) In an attempt to harmonize this somewhat hybrid picture, we invoke Wheeler's [32] "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 antithesis to Landauer's slogan, and we propose the following synthesis of the two: If Wheeler motivates us to look at the environment as being a computation 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, logico-computational version. Time evolutions of closed systems are injective: Nature computes with Toffoli, but no AND or OR gates.

Note that this fact is *a priori* asymmetric 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 called deterministic, and randomized otherwise.

Logical reversibility is a simple computational version of a discretized second law, and it has implications resembling the traditional versions of the law: First, 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 $K(R_t) + O(1)$ if R_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 an 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 while the lighter gets lighter.[4] A function accentuating imbalance, instead of lessening it, is not injective, as the following counting argument shows.

"Clausius" toy 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 1s in one of the halves (which equals the deficit of 1s in the other). We observe that the probability (with respect to the uniform distribution over all strings of some Hamming-weight couple $(wn, (1 - w)n)$, where the first half has wn 1s and the

[4]The Hamming weight of a binary string S is the number of 1s in S.

second $(1-w)n)$ 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 $1/2 \leqslant w < 1$ and $0 < \Delta \leqslant 1-w$, is

$$\frac{\binom{n}{(w+\Delta)n}\binom{n}{(1-w-\Delta)n}}{\binom{n}{wn}\binom{n}{(1-w)n}} = 2^{-\Theta(n)}. \tag{26}$$

The example suggests that logical reversibility might be the "Church-Turing manifestation" of the second law: If reality is computed by a Turing machine, then physical laws correspond to properties of such computations—as in the case of the second law: logical reversibility. If we assume for a moment that the second law of thermodynamics has indeed such a simple Church-Turing manifestation, it is a natural question in how far this already makes the law special. In fact, the law does have a peculiar related property, *encoding independence*: Since the second law deals with degrees of freedom, and a degree of freedom will correspond, in another encoding, to a degree of freedom again, it is either respected in both encodings or violated in both. (In comparison: It cannot be decided by just looking at a running program whether the simulated system "violates or respects Kepler's laws"—that would crucially depend on how masses and their position are represented by the code.) Hand in hand with this comes the property of simulation resilience. Let us take again the example of Kepler's laws: If a program simulates planets moving in, say, square orbits, this does not mean that the program execution, viewed as a physical process, is by itself problematic—the laws of gravity are not simulation resilient. If, in sharp contrast to this, we simulate a system, e.g., the microstate sequence of a steam engine, then that simulation process violates the second law just as the simulated system does: With respect to the second law of thermodynamics, the simulation of "reality" is just as good as "reality" itself.

VIII. CONCLUSION

We start from Landauer's principle, stating that the erasure of information requires an amount of free energy, to be dissipated as heat into the environment, proportional to the number of independent binary degrees of freedom of that information. Specifically, the use of reversible data compression, imagined to be carried out by Fredkin and Toffoli's ballistic computer, implies that the necessary amount is proportional to the length of the best compression of the information into a binary string (and not to the length of the original string, as often stated). We generalize and broaden the scope of the principle, and its converse, to lower bounds on the free-energy cost of—or gain from—a general computation: the bounds on cost versus gain are in accord.

Landauer derived, in 1961, his principle from the second law of thermodynamics. We close the circle by formulating a simple "Church-Turing version" of that law: logical reversibility of the overall computation, including the environment. This fact alone implies variants of the historical versions of the second law, due to Boltzmann, Clausius, and also Kelvin; it is perhaps equivalent to theirs, certainly simpler. The arising belief that the law is rather *logical* than physical in its nature is nourished by two properties of the second law: its encoding independence and its *simulation resilience*.

Confronted with the relevance of the second law of thermodynamics in computation, and with its simulation resilience, let us close with the (provocative) question whether Landauer's "Information is physical" should be replaced by *"The second law of thermodynamics is not physical."*

ACKNOWLEDGMENTS

We thank Charles Bédard, Claus Beisbart, Sophie Berthelette, Paul Boes, Gilles Brassard, Časlav Brukner, Xavier Coiteux-Roy, Borivoje Dakić, Paul Erker, Jürg Fröhlich, and Arne Hansen for enlightening discussions, and anonymous reviewers for helpful comments. Ä.B. is supported by the Swiss National Science Foundation (SNSF) under Grant No. 175860, the Erwin Schrödinger Center for Quantum Science & Technology (ESQ), and the Austrian Science Found (FWF): Z3. S.W. is supported by the Swiss National Science Foundation (SNSF) under Grant No. 182452, the NCCR QSIT under Grant No. 185902, and the Hasler Foundation.

[1] R. Landauer, Information is inevitably physical, in *Feynman and Computation: Exploring the Limits of Computers*, edited by A. J. G. Hey (Perseus Books, Reading, MA, 1998), Chap. 8, pp. 77–92.

[2] R. Landauer, Irreversibility and heat generation in the computing process, IBM J. Res. Dev. **5**, 183 (1961).

[3] L. Szilard, Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen, Z. Phys. **53**, 840 (1929).

[4] O. C. O. Dahlsten, R. Renner, E. Rieper, and V. Vedral, The work value of information, arXiv:0908.0424.

[5] C. H. Bennett, The thermodynamics of computation—A review, Int. J. Theor. Phys. **21**, 905 (1982).

[6] O. C. O. Dahlsten, R. Renner, E. Rieper, and V. Vedral, Inadequacy of von Neumann entropy for characterizing extractable work, New J. Phys. **13**, 053015 (2011).

[7] P. Faist, F. Dupuis, J. Oppenheim, and R. Renner, The minimal work cost of information processing, Nat. Commun. **6**, 7669 (2015).

[8] M. B. Plenio and V. Vitelli, The physics of forgetting: Landauer's erasure principle and information theory, Contemp. Phys. **42**, 25 (2001).

[9] T. Sagawa and M. Ueda, Fluctuation Theorem with Information Exchange: Role of Correlations in Stochastic Thermodynamics, Phys. Rev. Lett. **109**, 180602 (2012).

[10] S. Deffner and C. Jarzynski, Information Processing and the Second Law of Thermodynamics: An Inclusive, Hamiltonian Approach, Phys. Rev. X **3**, 041003 (2013).

[11] N. Merhav, Relations between work and entropy production for general information-driven, finite-state engines, J. Stat. Mech.: Theory Exp. (2017) 023207.

[12] A. B. Boyd, D. Mandal, and J. P. Crutchfield, Thermodynamics of Modularity: Structural Costs Beyond the Landauer Bound, Phys. Rev. X **8**, 031036 (2018).
[13] M. P. Frank, Physical foundations of Landauer's principle, in *Reversible Computation*, edited by J. Kari and I. Ulidowski (Springer, Cham, 2018), pp. 3–33.
[14] A. N. Kolmogorov, Three approaches to the quantitative definition of information, Int. J. Comput. Math. **2**, 157 (1968).
[15] E. Fredkin and T. Toffoli, Conservative logic, Int. J. Theor. Phys. **21**, 219 (1982).
[16] W. H. Zurek, Algorithmic randomness and physical entropy, Phys. Rev. A **40**, 4731 (1989).
[17] C. H. Bennett, Notes on Landauer's principle, reversible computation, and Maxwell's demon, Stud. Hist. Philos. Sci. B **34**, 501 (2003).
[18] L. del Rio, J. Åberg, R. Renner, O. C. O. Dahlsten, and V. Vedral, The thermodynamic meaning of negative entropy, Nature (London) **474**, 61 (2011).
[19] C. H. Bennett, Logical reversibility of computation, IBM J. Res. Dev. **17**, 525 (1973).
[20] J. Ziv and A. Lempel, Compression of individual sequences via variable-rate coding, IEEE Trans. Inf. Theory **24**, 530 (1978).
[21] G. J. Chaitin, A theory of program size formally identical to information theory, J. ACM **22**, 329 (1975).
[22] W. H. Zurek, Thermodynamic cost of computation, algorithmic complexity and the information metric, Nature (London) **341**, 119 (1989).
[23] S. Wolf, Second thoughts on the second law, in *Adventures between Lower Bounds and Higher Altitudes*, edited by H. J. Böckenhauer, D. Komm, and W. Unger, Lecture Notes in Computer Science book series, LNCS Vol. 11011 (Springer, Cham, 2018), pp. 463–476.
[24] J. S. Bell, On the Einstein Podolsky Rosen paradox, Phys. Phys. Fiz. **1**, 195 (1964).
[25] S. Wolf, Nonlocality without counterfactual reasoning, Phys. Rev. A **92**, 052102 (2015).
[26] S. Popescu and D. Rohrlich, Quantum nonlocality as an axiom, Found. Phys. **24**, 379 (1994).
[27] B. S. Cirel'son, Quantum generalizations of Bell's inequality, Lett. Math. Phys. **4**, 93 (1980).
[28] J. Barrett, L. Hardy, and A. Kent, No Signaling and Quantum Key Distribution, Phys. Rev. Lett. **95**, 010503 (2005).
[29] R. Colbeck and R. Renner, Free randomness can be amplified, Nat. Phys. **8**, 450 (2012).
[30] Ä. Baumeler, C. A. Bédard, G. Brassard, and S. Wolf, Kolmogorov amplification from Bell correlation, in *2017 IEEE International Symposium on Information Theory (ISIT)* (IEEE, New York, 2017), pp. 1544–1558.
[31] S. Wolf, An all-or-nothing flavor to the Church-Turing hypothesis, in *Theory and Applications of Models of Computation. TAMC 2017*, edited by T. Gopal, G. Jäger, and S. Steila, Lecture Notes in Computer Science book series, LNCS Vol. 10185 (Springer, Cham, 2017), pp. 39–56.
[32] J. A. Wheeler, Information, physics, quantum: The search for links, in *Proceedings of the third International Symposium Foundations of Quantum Mechanics in the Light of New Technology, Central Research Laboratory, Hitachi, Ltd., Kokubunji, Tokyo, Japan, August 28–31, 1989*, edited by S.-i. Kobaysahi, H. Ezawa, Y. Murayama, and S. Nomura (Physical Society of Japan, Tokyo, Japan, 1990), pp. 354–368, reprinted in Ref. [1].