Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference

2016 New J. Phys. 18 093052

(http://iopscience.iop.org/1367-2630/18/9/093052)

New Journal of Physics
The open access journal at the forefront of physics

Deutsche Physikalische Gesellschaft DPG
IOP Institute of Physics

Published in partnership with: Deutsche Physikalische Gesellschaft and the Institute of Physics

OPEN ACCESS

RECEIVED
15 January 2016

REVISED
20 May 2016

ACCEPTED FOR PUBLICATION
1 July 2016

PUBLISHED
27 September 2016

PAPER

Algorithmic independence of initial condition and dynamical law in thermodynamics and causal inference

Dominik Janzing[1,5], **Rafael Chaves**[2,3,4] **and Bernhard Schölkopf**[1]

[1] Max Planck Institute for Intelligent Systems, Spemannstr. 38, D-72076 Tübingen, Germany
[2] Institute for Physics & FDM, University of Freiburg, D-79104 Freiburg, Germany
[3] Institute for Theoretical Physics, University of Cologne, D-50937 Cologne, Germany
[4] International Institute of Physics, Federal University of Rio Grande do Norte, 59070-405 Natal, Brazil
[5] Author to whom any correspondence should be addressed.

E-mail: dominik.janzing@tuebingen.mpg.de

Keywords: arrow of time, causal inference, Kolmogorov complexity, physical entropy, algorithmic randomness

Abstract

We postulate a principle stating that the initial condition of a physical system is typically algorithmically independent of the dynamical law. We discuss the implications of this principle and argue that they link thermodynamics and causal inference. On the one hand, they entail behavior that is similar to the usual arrow of time. On the other hand, they motivate a statistical asymmetry between cause and effect that has recently been postulated in the field of causal inference, namely, that the probability distribution P_{cause} contains no information about the conditional distribution $P_{\text{effect}|\text{cause}}$ and vice versa, while P_{effect} may contain information about $P_{\text{cause}|\text{effect}}$.

1. Introduction

Drawing causal conclusions from statistical data is at the heart of modern scientific research. While it is generally accepted that *active* interventions to a system (e.g. randomized trials in medicine) reveal causal relations, statisticians have widely shied away from drawing causal conclusions from *passive* observations. Meanwhile, however, the increasing interdisciplinary field of causal inference has shown that also the latter is possible—even without information about time order—if appropriate assumptions that link causality and statistics are made [1–3], with applications in biology [4], psychology [5], and economy [6]. More recently, also foundational questions of quantum physics have been revisited in light of the formal language and paradigms of causal inference [7–13].

Remarkably, recent results from causal inference have also provided new insights about the thorny issue of the arrow of time. Contrary to a wide-spread belief, the joint distribution $P_{X,Y}$ of two variables X and Y sometimes indicates whether X causes Y or vice versa [14]. More conventional methods rely on conditional independencies and thus require statistical information of at least three observed variables [1, 2]. The intuitive idea behind the new approach is that if X causes Y, P_X contains no *information* about $P_{Y|X}$ and vice versa. Within this context it is not obvious, however, how to make precise the meaning of information. Accordingly, different formalizations of this intuitive notion have been proposed.

The *algorithmic information* approach proposed in [15, 16] gives a precise meaning to information by postulating that knowing $P_{Y|X}$ does not admit a shorter description of P_X and vice versa. This is the approach we will follow more closely here, in particular for drawing a link to thermodynamics, given that algorithmic information has already been related to the *thermodynamic entropy* [17]. Nevertheless, we should also mention an interpretation for the meaning of information recently stated in the context of *machine learning*, more precisely in *semi-supervised learning* (SSL).

SSL algorithms learn the statistical relation between two random variables X and Y from some (x, y)-pairs $(x_1, y_1), \ldots, (x_n, y_n)$ plus some unpaired instances $x_{n+1}, \ldots, x_{n+k}$. The algorithms are then supposed to predict y (or the conditional distribution $P_{Y|X=x}$) for any given input x. Without the unpaired instances, one would obtain the so-called *supervised learning* scenario [18]. There is an ongoing debate [19] in the field about in which sense and under which conditions the unpaired x-values (which, *a priori*, only tell us something about P_X) contain information about the relation between X and Y. Rephrasing this in the language used above: does P_X contain

information about $P_{Y|X}$? References [20, 21] draw the link to causality by the conjecture that the additional x-values do not help to learn $P_{Y|X}$ if X is the cause and Y the effect (the 'causal learning' scenario), while they may help when Y is the cause and X the effect (the 'anticausal learning' scenario). The hypothesis is supported by a meta-study that only found success cases of SSL in the literature for the anticausal but not for the causal learning scenario [20, 21][6]. This suggests that the 'no information' idea—despite its apparent vagueness—describes an asymmetry between cause and effect that is already relevant for scientific tasks other than causal inference. It is thus natural to explore such kind of asymmetries in the context of *physical* systems.

As a matter of fact, like the asymmetries between cause and effect, similar asymmetries between past and future are also manifest even in stationary time series [22] which can sometimes be used to infer the direction of empirical time series (e.g. in finance or brain research) or to infer the time direction of movies [23]. Altogether, these results suggest a deeper connection for the asymmetries between cause versus effect and past versus future. In particular, a physical toy model relating such asymmetries to the usual thermodynamic arrow of time has been proposed [24].

Motivated by all these insights, we propose a foundational principle for both types of asymmetries, cause versus effect and past versus future. The contributions of this paper are the following:

(1) We postulate a principle stating that the initial state of a physical system and the dynamical law to which it is subjected to should be algorithmically independent.

(2) As we show, this principle implies for a closed system the non-decrease of physical entropy if the latter is identified with algorithmic complexity (also called 'Kolmogorov complexity'). Thus, it reproduces the thermodynamic behavior for *closed* systems given earlier insights on the thermodynamic relevance of algorithmic information proposed in the literature [17].

(3) Our principle brings new insights to understand open system and we apply it to a toy model representing typical cause–effect relations.

(4) We show that the algorithmic independence of P_{cause} and $P_{\text{effect}|\text{cause}}$ stated earlier can be seen as part of this principle, if we identify cause and effect with the initial and final states of a physical system, respectively.

This paper thus links recently stated ideas from causal inference with a certain perspective of thermodynamics. To bridge such different lines of research, we start by reviewing several relevant ideas of both.

Algorithmic randomness in thermodynamics. We start by briefly introducing some basic notions of algorithmic information theory. The algorithmic randomness (also called 'algorithmic complexity', 'algorithmic information', or 'Kolmogorov complexity') $K(s)$ of a binary string s is defined as the length of its shortest compression. More precisely, $K(s)$ is the length of the shortest program on a universal Turing machine (with prefix-free encoding) that generates s and then stops [25, 26]. We call this shortest program[7] s^* the *shortest compression of* s.

The *conditional algorithmic complexity* $K(s|t)$ is defined as the length of the shortest program generating the output s from the input t. A slightly different quantity is $K(s|t^*)$ since the input t^* is slightly more valuable than the input t. This is because a Turing machine is able to convert t^* into t (by definition), while there can be in principle no algorithm that finds t^* when t is given. One can therefore show that $K(s|t)$ may be larger than $K(s|t^*)$ by a term that can grow at most logarithmically in the size of the length of t. Accounting for this kind of subtleties, several statements in Shannon information theory have nice analogues in algorithmic information theory. For instance, the information of a pair[8] is given by a sum of the information of one string and the conditional information of the other:

$$K(s, t) \stackrel{+}{=} K(s) + K(s|t^*).$$

As common in algorithmic information theory [27], the equation is not exact and therefore the equation sign is marked by the symbol $+$ indicating an error term that can be upper bounded by a constant (which does not depend of the strings involved, but does depend on the Turing machine).

As further analogue to Shannon information theory, *algorithmic mutual information* can be defined in three equivalent ways [26]:

$$I(s : t) := K(s) + K(t) - K(s, t) \stackrel{+}{=} K(s) - K(s|t^*) \stackrel{+}{=} K(t) - K(t|s^*).$$

[6] Note that success of SSL does not necessarily imply that the unpaired x-values contained information about $P_{Y|X}$. Instead, they could have helped to fit a function that is particularly precise in regions where the x-values are dense. The meta-study suggests, however, that this more subtle phenomenon does not play the major role in current SSL implementations.

[7] If there are more than one such program, we refer to the first one with respect to some standard enumeration of binary words.

[8] Algorithmic information of a *pair* of binary words can be defined by first converting the pair to one string by some fixed bijection between pairs and single binary words, which can easily be constructed by enumerating binary words and then using some fixed bijection of $\mathbb{N} \times \mathbb{N}$ and $\mathbb{N}$.

Intuitively speaking, $I(s:t)$ is the number of bits saved when s and t are compressed jointly rather than independently, or, equivalently, the number of bits that the description of s can be shortened when the shortest description of t is known, and vice versa.

There are cases where Shannon information and algorithmic information basically coincide: Let $x_1, \ldots, x_n$ be samples drawn from a fixed distribution P_X on some alphabet $\mathcal{X}$. If s_n denotes the binary encoding of the n-tuple $(x_1, \ldots, x_n)$, then the algorithmic information rate $K(s_n)/n$ converges almost surely [27] to the Shannon entropy

$$H(p_x) = -\sum_x p(x) \log p(x).$$

Here and throughout the paper, lower case letters denote probability densities (for discrete distributions the density is just the probability mass function) corresponding to the respective distributions. For instance, p_X denotes the density of P_X, and whenever this causes no confusion, we write $p(x)$ instead of p_X for sake of convenience.

The more interesting aspects of algorithmic information, however, are those where the information content of a string cannot be derived from Shannon entropy. On the one hand, the asymptotic statement on the information rate blurs the fact that the description length of a typical n-tuple is given by $nH(p_X) + K(p_X)$. Hence, the description length of the distribution also needs to be accounted for; in order to achieve the compression length $nH(p_X)$ one would need to know p_X, hence the full description of s_n involves also describing p_X [28]. In the context of causal inference it has been pointed out [15] that the description length of the joint distribution of some random variables sometimes also contains information about the underlying causal links between the variables. Therefore, in causal discovery, restricting attention to Shannon information unavoidably ignores essential aspects of information.

A second reason why Shannon information is not sufficient for our purpose is that a string may not result from independent sampling at all. If, for instance, s describes the state of a multi-particle system, the particles may have interacted and hence the particle coordinates may be correlated. Then, treating the joint state of the system as if each particle coordinate would have been drawn independently at random overestimates the description length because it ignores the correlations. In this sense, algorithmic information includes aspects of information that purely statistical notions of information cannot account for.

In a seminal paper, Bennett [29] proposed to consider $K(s)$ as the *thermodynamic entropy* of a microscopic state of a physical system when s describes the latter with respect to some standard binary encoding after sufficiently fine discretization of the phase space. This assumes an 'internal' perspective (followed in parts of this paper), where the microscopic state is perfectly known to the observer. Although $K(s)$ is in principle uncomputable, it can be estimated from the Boltzmann entropy in many-particle systems, given that the microscopic state is typical in a set of states satisfying some macroscopic constraints [17, 29]. That is, in practice one needs to rely on more conventional definitions of physical entropy.

From a theoretical and fundamental perspective, however, it is appealing to have a definition of entropy that neither relies on missing knowledge like the statistical Shannon/von-Neumann entropy [30, 31] nor on the separation between microscopic versus macroscopic states—which becomes problematic on the mesoscopic scale—like the Boltzmann entropy [32]. For imperfect knowledge of the microscopic state, Zurek [17] considers thermodynamic entropy as the sum of statistical entropy and Kolmogorov complexity [33], which thus unifies the statistical and the algorithmic perspectives of physical entropy.

To discuss how $K(s)$ behaves under Hamiltonian dynamics, notice that the dynamics on a continuous space is usually not compatible with discretization, which immediately introduces also statistical entropy in addition to the algorithmic term—particularly for chaotic systems [34]—in agreement with standard entropy increase by coarse-graining [35, 36]. Remarkably, however, $K(s)$ can also increase by applying a one-to-one map D on a *discrete* space [17]. Then $K(s) + K(D)$ is the tightest upper bound for $K(D(s))$ that holds for the general case. For a system starting in a *simple* initial state s and evolving by the repeated application of some simple map $\tilde{D}$, the description of $s_t := D(s) := \tilde{D}^t(s)$ essentially amounts to describing t and Zurek derives a logarithmic entropy increase until the scale of the recurrence time is reached [17]. Although logarithmic growth is rather weak [34], it is worth mentioning that the arrow of time here emerges from assuming that the system starts in a *simple* state. We will later argue that this is just a special case of the idea that we propose here, that is, that the initial state is independent of D. The fact that K depends on the Turing machine could arguably spoil its use in physics. However, in the spirit of Deutsch's idea that the laws of physics determine the laws of computation [37], future research may define a 'more physical version' of K by a computation model whose elementary steps directly use physically realistic particle interactions, see e.g. the computation models in [38–40]. Moreover, *quantum* thermodynamics [41] should rather rely on *quantum* Kolmogorov complexity [42].

Algorithmic randomness in causal inference. Reichenbach's principle [43] states that every statistical dependence between two variables X, Y must involve some sort of causation: either direct causation (X causes Y or vice versa) or a common cause for both X and Y. Conversely, variables without causal relation are statistically independent. However, causal relations in real life are not always inferred from *statistical* relations. Often, one

just observes similarities between single objects that indicate a causal relation. As argued in [15], two binary words x, y, representing two causally disconnected objects should be algorithmically independent, i.e.

$$I(x:y) \stackrel{+}{=} 0.$$

Depending on the context, we will here read the equation sign $\stackrel{+}{=}$ in two different ways: for theorems, symbols like x, y are considered placeholders for strings that can be arbitrarily long. Then $\stackrel{+}{=}$ means that the error term does not grow with the length of the strings (although it does depend on the Turing machine). In a concrete application where x and y are *fixed* finite strings, this is certainly meaningless. Then we interpret $\stackrel{+}{=}$ by saying that the error is 'small' compared to the complexity of the strings under consideration (provided that the latter are complex enough). The decision about what 'sufficiently complex' means is certainly difficult, but analogue issues also occur in statistics: rejecting or accepting statistical independence also depends on the choice of the significance levels (which can never be chosen by purely scientific reasons) since statistical independence actually refers to an infinite sample limit that is never reached in real-life. For sake of simplicity, we will henceforth just distinguish between *dependent* versus *independent*.

Rephrasing the ideas of [15], one could say that algorithmic independence between objects is what *typically* happens when objects are generated without causal relations, i.e., without information exchange. To elaborate on this idea, [16] considers a model where strings are created according to Solomonoff's prior [44] that is defined as the distribution of outputs obtained by uniformly randomizing all bits in the infinite input band of a Turing machine and conditioning on the case that the program halts. It can be shown [45] that this results essentially (up to factors of order 1) in the following probability distribution on the strings:

$$p(x) = c \cdot 2^{-K(x)},$$

where c is a normalization constant. Obviously, Solomonoff's prior assigns higher probability to *simple* strings. For this reason, it is often considered as a very principled implementation of *Occam's Razor* in the foundations of learning. Following this prior, if two strings x, y are generated by two independent random processes of this kind, the pair then occurs with probability

$$p(x, y) = c^2 \cdot 2^{-K(x)} \cdot 2^{-K(y)}.$$

On the other hand, it occurs with probability

$$p(x, y) = c \cdot 2^{-K(x,y)},$$

when it is generated in a joint process. Thus, $I(x:y) = K(x) + K(y) - K(x, y)$ measures the log of the probability ratio for the occurrence (after neglecting the constant c). In this sense, the amount of algorithmic information shared by two objects (or, more precisely, by the strings encoding them) can be taken as measuring the evidence for the hypothesis that they are causally related. Here, one may again object that the dependence of $I(x:y)$ on the Turing machine renders the claim at least vague if not useless: x and y can be independent with respect to one Turing machine but significantly dependent with respect to a second one. Indeed, the asymptotic statement 'equal up to a constant' does not help. Apart from our remarks from above requesting for 'natural' Turing machines for the purpose of physics, we mention that [15] discusses that the notion of being causally connected or not is also *relative*: assume, for instance, one considers the genomes of two humans. With respect to a 'usual' Turing machine we will observe significant amount of algorithmic mutual information just because both genomes are from humans. On the other hand, given a Turing machine that is specifically designed for encoding humans genomes, the mutual information is only significant if the subjects are related apart from both being humans. Certainly, the fact that they are from the same species is also a causal relation, but if we focus on causal relations on top of this (i.e., relatedness in the sense of family relations), we should only look at algorithmic dependences with respect to a Turing machine that has access to the respective background information. In other words, the fact that algorithmic mutual information is relative fits well to causality being relative as well (in the above sense).

Reference [15] further elaborates on the idea of using algorithmic dependences to obtain causal information. It develops a graphical model based framework for inferring causal relations among n objects based on *conditional* algorithmic (in)dependences in analogy to conventional causal inference which infers causal graphs among n variables from conditional statistical (in)dependences [1, 2].

More surprisingly, algorithmic information can also be used to infer whether X causes Y (denoted by $X \to Y$) or Y causes X from their joint distribution, given that exactly one of the alternatives is true[9]. If P_{cause} and $P_{\text{effect}|\text{cause}}$ are 'independently chosen by nature' and thus causally unrelated, [15] postulates that their algorithmic mutual information is negligible, formally

[9] The difficult question how well-defined causal directions emerge in physical systems whereinteractions actually imply mutual influence is discussed for a toy model in [46].

$$I(P_{\text{cause}} : P_{\text{effect}|\text{ cause}}) \stackrel{+}{=} 0. \quad (1)$$

The postulate raises, however, the following questions for practical applications: First, the joint distribution of cause and effect is not known and can only be *estimated* from finite data. The estimated distribution may show dependences between P_{cause} and $P_{\text{effect}|\text{cause}}$ that disappear in infinite sampling. Second, algorithmic mutual information is uncomputable. For these reasons, the independence postulate has only been used as an *indirect* justification of practical causal inference methods. We now describe two examples.

Causal inference for linear models with non-Gaussian noise. First, consider linear non-Gaussian additive noise models [47]: let the joint distribution $P_{X,Y}$ of two random variables X, Y be given by the linear model

$$Y = \alpha_X X + N_Y, \quad (2)$$

where $\alpha_X \in \mathbb{R}$ and N_Y is an unobserved noise term that is statistically independent[10] of X. Whenever X or N_Y is non-Gaussian, it follows that for every model of the form $X = \alpha_Y Y + N_X$, the noise term N_X and Y are statistically dependent, although they may be uncorrelated. That is, except for Gaussian variables, a linear model with independent noise can hold at most in one direction. Within that context, [47] infers the direction with additive independent noise to be the causal one. To justify this reasoning, [48] argues that whenever (2) holds, the densities of P_Y and $P_{X|Y}$ are related by the differential equation

$$\frac{\partial^2}{\partial y^2} \log p(y) = -\frac{\partial^2}{\partial y^2} \log p(x|y) - \frac{1}{\alpha_X} \frac{\partial^2}{\partial x \partial y} p(x|y).$$

Therefore, knowing $P_{X|Y}$ enables a short description of P_Y. Whenever P_Y has actually high description length (which can, of course, only be conjectured but never be proven for the specific case under consideration), we thus reject $Y \to X$ as a causal explanation. It should be emphasized that this justification does not assume that causal relations in nature are always linear. Instead, the statement reads: whenever the joint distribution is linear in one direction but not the other, the former is likely to be the causal direction. This is because it would be an implausible coincidence that P_{cause} and $P_{\text{effect}|\text{cause}}$ together generate a joint distribution that admits a linear model from effect to cause.

Information-geometric causal inference. Second, we consider the toy scenario described in [49, 50]. Assume that X and Y are random variables with values in $[0,\ 1]$, deterministically related by $Y = f(X)$ and $X = f^{-1}(Y)$, where f is a monotonically increasing one-to-one mapping of $[0,\ 1]$. If X is the cause and Y the effect then $P_{\text{effect}|\text{cause}}$ is uniquely described by f, while $P_{\text{cause}|\text{effect}}$ is given by f^{-1}. Hence, applying (1) to this special case yields

$$I(p_X : f) \stackrel{+}{=} 0. \quad (3)$$

In trying to replace (3) with a criterion that is empirically decidable, [49] postulates

$$\int_0^1 \log f'(x) p(x) \mathrm{d}x = \int_0^1 \log f'(x) \mathrm{d}x, \quad (4)$$

where f' denotes the derivative of f. In words, averaging the logarithmic slope of f over p_X is the same as averaging it over the uniform distribution. As already observed in [49, 50], (4) is equivalent to uncorrelatedness between $\log f'$ and p_X. Here, one interprets both functions $x \mapsto \log f'$ and $x \mapsto p(x)$ as *random variables* on the probability space $[0,\ 1]$ with the uniform distribution. Then the difference between the left- and the right-hand side of (4) can be written as the covariance of these random variables:

$$\mathrm{Cov}(\log f',\ p_X) = \int_0^1 \log f'(x) p(x) \mathrm{d}x - \int_0^1 \log f'(x) \mathrm{d}x \cdot \int_0^1 p(x) \mathrm{d}x.$$

To further justify (3), [50] discusses scenarios where functions f and distributions P_X are independently generated at random in a way that ensures that (4) is approximately correct with high probability. For instance, P_X can be obtained by randomly distributing some peaks across the interval $[0,\ 1]$. The same type of process can be used to generate a monotonic function f at random because the cumulative distribution function of any strictly positive probability density on $[0,\ 1]$ defines, as desired, a monotonic bijection of $[0,\ 1]$. Stating that (4) typically holds approximately always relies on strong assumptions on the generating processes for p_X and f. Therefore, (4) is just a pragmatic way to replace algorithmic independence with a computable independence condition. Intuitively, we consider (4) as stating that some of the peaks of p_X lie in regions where f has large slope, and some in regions with small slope, such that on the average the expectation of $\log f'$ over p_X does not significantly differ from the one with the uniform distribution.

One can show [49, 50] that the independence condition (4) implies a dependence for the backwards direction, i.e., the output density p_Y is positively correlated with $f^{-1'}$:

[10] Note that two variables Z, W are called statistically independent if $P_{Z,W} = P_Z P_W$, which is stronger than being uncorrelated, i.e., $\mathbb{E}[ZW] = \mathbb{E}[Z]\mathbb{E}[W]$.

$$\mathrm{Cov}(\log f^{-1'}, p_Y) \geqslant 0, \tag{5}$$

with equality if and only if f is the identity. Hence, the output density p_Y tends to be higher in regions where the function f^{-1} is steep. This is because the function 'focuses' points into regions where the derivative is small. In that sense, p_Y contains information about the mechanism relating X and Y. Moreover, [49, 50] show that (4) implies that the Shannon entropies of p_Y and p_X satisfy

$$H(p_Y) \leqslant H(p_X), \tag{6}$$

with equality if and only if f is the identify. This information theoretic implication is the main reason, among others, for stating (4) with $\log f'$ instead of just using f'. Intuitively, (6) holds because applying f to a density typically adds additional peaks, which makes the density less uniform. Only functions f that are adapted to the specific shape of the density p_X can make it smoother. As a result, [49] proposes the cause to be the variable with smaller entropy (subject, for course, to assuming a deterministic relation).

2. Results

A common root for thermodynamics and causal inference. To provide a unifying foundation connecting thermodynamics and causal inference we postulate:

Principle 1 (**Algorithmic independence between input and mechanism**). If s is the initial state of a physical system and M a map describing the effect of applying the system dynamics for some fixed time, then s and M are algorithmically independent, i.e.,

$$I(s : M) \stackrel{+}{=} 0. \tag{7}$$

In other words, knowledge of s does not enable a shorter description of M (and vice versa, with the roles of s and M interchanged). Here, we assume that the initial state, by definition, is a state that has not interacted with the dynamics before.

The last sentence requires some explanations to avoid erroneous conclusions. Below we will discuss its meaning for an intuitive example (see the end of the paragraph 'physical toy model for a deterministic non-linear cause-effect relation'). The example will also suggest that states that are independent in the sense that they 'have never seen the mechanism before' occur quite often in nature. Note that 'not seeing the mechanism' also excludes a preparation procedure for s that accounts for the length of the time interval the dynamics is active because this information is, by definition, considered as part of M.

Principle 1 is entailed by the assumption that there is no algorithmic dependence in nature without an underlying causal relation. By overloading notation, we have identified mechanism and state with their encodings into binary strings. Principle 1 needs to be taken with a grain of salt. Again, there may be some information shared by s and M that we do not account for because we call it 'background' information. Assume, for instance, we place some billiard balls on a pool table and give them randomly some momenta. In doing this, we are aware of the dynamical laws governing the balls, but we may not aware of the exact size of the table. Then, the latter is the decisive aspect of the dynamics that is algorithmically independent of the initial state. More generally, we consider the descriptions of M and s, given some background information and postulate independence conditional on the latter. Although this renders the postulate somehow tautologic, it is still useful because it has mathematical implications which are non-trivial, although they have to be taken relative to the respective background information.

To address further potential issues with principle 1, note that generalizations of algorithmic mutual information for *infinite* strings can be found in [45], which then allows to apply principle 1 to continuous physical state spaces. Here, however, we consider finite strings describing states after sufficiently fine discretizations of the state space instead, neglecting issues from chaotic systems [34] for sake of conciseness.

We should also discuss the question of how to interpret the sign $\stackrel{+}{=}$ in this context. For fixed s and M, the mutual information takes one specific value and stating that they are zero 'up to a constant term' does not make sense. A pragmatic interpretation is to replace 'up to a constant term' with 'up to a small term', where the decision of what is considered small will heavily depend on the context. A more principled interpretation is the following. In continuous space, the binaries describing state and dynamics depend on the chosen discretization. Then $\stackrel{+}{=}$ can be read as stating that the algorithmic mutual information does not increase with finer discretization.

Dynamics of closed physical systems. Principle 1 has implications that follow from the independence condition (7) regardless of *why* the independence holds in the first place. It may hold because the state has been prepared independently or because some noise has destroyed previous dependences of the state with M.

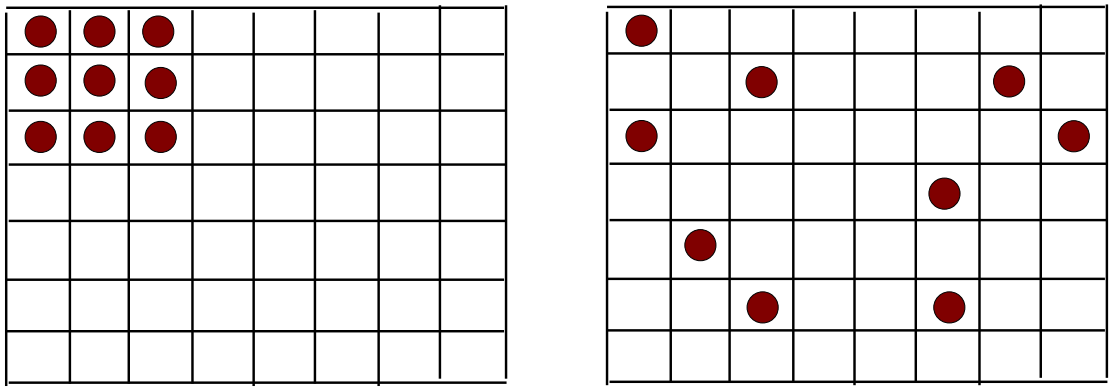

Figure 1. Left: cellular automaton starting in a state with small description length. Right: the state s' obtained from s by application of the dynamical law D.

Moreover, one could argue for a notion of 'initial state' that, by definition, implies that it has been prepared independently of M and thus, typically, shares no algorithmic information with M.

To show one immediate consequence, consider a physical system whose state space is a finite set S. Assuming that the dynamics D is a bijective map of S, it follows that the entropy cannot decrease:

Theorem 1 (**No entropy decrease**). *If the dynamics of a system is an invertible mapping D of a discrete set S of states then principle 1 implies that the algorithmic complexity can never decrease when applying D to the initial state s, i.e.*

$$K(D(s)) \stackrel{+}{\geqslant} K(s). \tag{8}$$

Proof. Algorithmic independence of s and D amounts to $K(s) \stackrel{+}{=} K(s|D^*)$. Since D is invertible, s can be computed from $D(s)$ and vice versa implying that $K(s|D^*) \stackrel{+}{=} K(D(s)|D^*)$. Thus, $K(s) \stackrel{+}{=} K(s|D^*) \stackrel{+}{=} K(D(s)|D^*) \stackrel{+}{\leqslant} K(D(s))$, concluding the proof. □

The proof is very intuitive: if $D(s)$ had a shorter description than s, knowing D would enable a shorter description of s because one could describe the latter by first describing $D(s)$ and adding the remark 'then apply D^{-1}'. The argument does not really require D to be injective for general states. Instead, it only uses that s can be uniquely reconstructed from $D(s)$.

Dynamical laws of physical systems are often simple, i.e., have small description length. Yet, principle 1 and theorem 1 are not pointless because D may also be considered as the t-fold concatenation of the same map $\tilde{D}$, where $\tilde{D}$ itself has negligible description length. Then, $I(s : D) \stackrel{+}{=} 0$ amounts to $I(s : t) \stackrel{+}{=} 0$. Theorem 1 implies that $K(\tilde{D}^t(s)) \stackrel{+}{\geqslant} K(s)$ whenever t and s are algorithmically independent. That is, while [17] derives entropy increase for a simple initial state s, we have derived it for all states s that are independent of t.

To further illustrate theorem 1, consider a toy model of a physical system consisting of $n \times m$ cells, each being occupied or not with a particle, see figure 1. Its state is described by a binary word s with nm digits. For generic s, we have $K(s) \approx nm$, while figure 1, left, shows a simple state where all particles are in the left uppermost corner containing $k \times l$ cells. A description of this state s consists essentially of describing k and l (up to a negligible amount of extra information specifying that k and l describe the size of the occupied region), which requires $\log_2 k + \log_2 l$ bits. Assume now that the dynamical evolution D transforms s into $s' = D(s)$ where s' looks 'more generic', as shown in figure 1, right. In principle, we cannot exclude that s' is equally simple as s due to some non-obvious pattern. However, excluding this possibility as unlike, theorem 1 rules out any scenario where s' is the initial state and s the final state of *any* bijective mapping D that is algorithmically independent of s'. The transition from s to s' can be seen as a natural model of a mixing process of a gas, as described by popular toy models like lattice gases [51]. These observations are consistent with standard results of statistical mechanics saying that mixing is the typical behavior, while de-mixing requires some rather specific tuning of microscopic states. Here we propose to formalize 'specific' by means of algorithmic dependencies between the initial state and the dynamics. Here, this view does not necessarily generate novel insights for *typical* scenarios of statistical physics, but it introduces a link to crucial concepts in the field of causal inference.

So far, we have avoided to discuss whether the assumption of discrete state space came from the discretization of a continuous system (which is problematic for reasons mentioned earlier) or from really focusing on discrete systems. In the former case, despite these issues, theorem 1 still shows that increase of

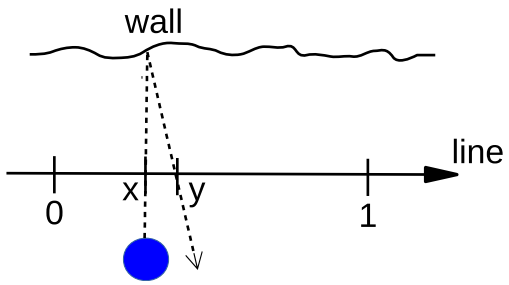

Figure 2. Physical system generating a nonlinear deterministic causal relation: a particle travelling towards a structured wall with momentum orthogonal to the wall, where it is backscattered in a slightly different direction. x and y denote the positions where P crosses a vertical line before and after the scattering process, respectively.

physical entropy does not necessarily require coarse graining effects. To argue for the latter view, one may think of a discrete quantum dynamics starting in an eigenstate with respect to some natural basis, e.g., the energy basis, and also ending up in these basis states. To satisfy principle 1, the basis must be considered as background information relative to which the independence is stated.

Dynamics of open systems. Since applying (7) to closed systems reproduces the standard thermodynamic law of non-decrease of entropy, it is appealing to state algorithmic independence for closed system dynamics only and then obtain conditions under which the independence for open system follows. We will then see that the independence of P_{cause} and $P_{\text{effect}|\text{cause}}$ can be seen as an instance of the independence principle for open systems.

Let D be a one-to-one map transforming the initial joint state (s, e) of system and environment into the final state (s', e'). For fixed e, define the open system dynamics $M : s \mapsto s'$. If s is algorithmically independent of the pair (D, e) (which is true, for instance when $K(e)$ is negligible and s and D are independent), independence of s and M follows because algorithmic independence of two strings a, b implies independence of a, c whenever c can be computed from b via a program of length $O(1)$, see e.g. [15], lemma 6.

Further, we can extend the argument above to statistical ensembles: consider n systems with identical state space S, each coupled to an environment with identical state space E (where S and E are finite sets, for simplicity). Let $(s_j, e_j) \in S \times E$ be the initial state of the jth copy and (s_j', e_j') its final state. Following the standard construction of Markovian dynamics, we assume statistical independence between the initial state s and the initial environmental state e. Further, in agreement with the general idea of this paper, we assume also that $s^n := (s_1, \ldots, s_n)$ is algorithmically independent[11] of (D, e^n) with $e^n := (e_1, \ldots, e_n)$. For our statistical ensemble, the empirical conditional distribution of final sates s', given the initial state s reads:

$$p(s'|s) := \sum_e p(e|s) \approx \sum_e p(e),$$

where the sum runs over all e with $D(s, e) = (s', e')$ for some e'. The approximate equality holds because of the statistical independence of s and e, which is approximately also true for empirical frequencies if n is large. Hence, P_S is determined by s^n and $P_{S'|S}$ is (in the limit of large n) determined by e^n and D. We thus conclude that P_S and $P_{S'|S}$ are algorithmically independent, because they are derived from two algorithmically independent objects via a program of length $O(1)$. Defining the variable 'cause' by the initial state of one copy S and 'effect' as the final state, we have thus derived the algorithmic independence of P_{cause} and $P_{\text{effect}|\text{cause}}$. Notice that it is not essential in the reasoning above that cause and effect describe initial and final states of the same physical system, one could as well consider a tripartite instead of a bipartite system.

Physical toy model for a deterministic nonlinear cause–effect relation. To describe a case where principle 1 implies thermodynamic statements that are less standard, we revisit the toy scenario of information geometric causal inference [49, 50] and observe that (3) implies

$$K(p_Y) \overset{+}{\geqslant} K(p_X). \tag{9}$$

To see this, we only need to interpret the set of probability distributions as *states* on which f defines an invertible map and (9) follows in analogy to the proof of theorem 1 because p_X can be uniquely reconstructed for p_Y when f is known. Thus, if p_Y had a shorter description than p_X, knowing f would admit a shorter description of p_X. Equation (9) matches the intuition that a distribution typically gets additional peaks by applying the nonlinear function f. Remarkably, the increase of complexity on the phenomenological level, namely the level of distributions, is accompanied by a *decrease* of Shannon entropy. To avoid possible confusion, we should

[11] Note that this is stronger than assuming independence of s_j and (D, e_j) for every single component becausewe have thus also excluded algorithmic dependences between s and e, which get only apparent when looking at the whole ensemble.

emphasize that the process f, although it is a bijection, is not a 'reversible process' in the sense of thermodynamics because the latter term refers to maps that are locally volume preserving in phase space and thus *preserve* Shannon entropy. To further clarify this point, we now describe a simple physical system whose dynamics yields the function f when restricted to a certain part of the physical phase space. The decrease of Shannon entropy is then perfectly consistent with the conservation of Shannon entropy for the entire system, in agreement with Liouville's theorem.

Figure 2 shows a simple two-dimensional system with a particle P travelling towards a wall W perpendicular to the momentum of P. P crosses a line L parallel to W at some position $x \in [0,\ 1]$. Let the surface of W be structured such that P hits the wall with an incident angle that depends on its vertical position. Then P crosses L again at some position y. Assume that L is so close to W that the mapping $x \mapsto y =: f(x)$ is one-to-one. Also, assume that 0 is mapped to 0 and 1 to 1. Let the experiment be repeated with particles having the same momenta but with different positions such that x is distributed according to some probability density p_X. Assuming principle 1, the initial distribution of momenta and positions does not contain information about the structure of W. Due to theorem 1, the scattering process thus increases the algorithmic complexity of the state. Further, this process is thermodynamically irreversible for every thermodynamic machine that has no access to the structure of W. Hence, the entire dynamical evolution is thermodynamically irreversible when the structure of W is not known, although the Shannon entropy is preserved.

Let us now focus on a restricted aspect of this physical process, namely the process that maps p_X to p_Y via the function f, so we can directly apply the information-geometric approach to causal inference [49]. Now we conclude (9) because restricting the attention to partial aspects of two objects cannot increase their mutual information, see e.g., [15]. This illustrates, again, that we can either conclude (9) by applying principle 1 directly to f, or, alternatively, we could state the principle only for the dynamics of the *closed* system and derive (9) by standard arguments of algorithmic information theory. Intuitively speaking, we expect p_Y to contain information about the wall. On the one hand, we already know that (4) implies that p_Y correlates with the logarithmic slope of f, due to (5). On the other hand, we can also prove that p_Y contains *algorithmic* information about f provided that $K(p_Y)$ is properly larger than p_Y. This is because independence of p_Y and f would imply independence of p_Y and f^{-1} and then we could conclude $K(p_X) \stackrel{+}{\geqslant} K(P_Y)$ by applying the above arguments to f^{-1} instead of f. Certainly, particles contain information about the objects they have been scattered at and not about the ones they are going to be scattered at. Otherwise a photographic image would show the future and not the past. In this sense, the observations trivially fit to the usual arrow of time. What may be unexpected according to standard thermodynamics is, as already mentioned, the *decrease* of Shannon entropy (6), which could lead to misleading conclusions such as inferring the time direction from p_Y to p_X. Thus principle 1 is of particular relevance in scenarios where simple criteria like entropy increase/decrease are inapplicable, at least without accounting for the description of the entire physical system (that often may be not available, e.g., if the momentum of the particle is not measured). The example above also suggests how the algorithmic independence could provide a new tool for the inference of time direction in such scenarios.

One could certainly time reverse the scenario where p_Y is the particle density of the *incoming* beam while p_X corresponds to the *outgoing* beam. Then, the incoming beam already contains information about the structure of the surface it is scattered at later. We now argue how to make sense of principle 1 in this case. Of course, such a beam can only be prepared by a machine or a subject that is aware of the surface structure and directs the particles accordingly. As a matter of fact, particles who were never in contact with the object cannot 'a priori' contain information about it. Then principle 1 can be maintained if we consider the process of directing the particles as part of the mechanism and reject the idea of calling the state of the hand-designed beam an 'initial' state. Instead, the initial state then refers to the time instant before the particles have been given the fine-tuned momenta and positions.

Arrow of time for an open system in the real world. So far, we have provided mainly examples that help for a theoretical understanding of the common root of thermodynamics and causal inference. Apart from discussing the foundations for both fields, the independence principle aims at describing the arrow of time for systems for which it is not obvious how to derive asymmetries between past and future from standard thermodynamics.

As one such example, reference [52] considers audio signals from a piece of music and its echo at different places of a building and addresses the task of inferring which one is the original signal and which one its echo. On the one hand, one can consider this task as part of causal inference with the echo being the effect of the original signal, as in [52]. On the other hand, the problem is arguably related to the arrow of time since the echo comes later than its original signal. Here, it would be hard to infer the time direction from *entropy* arguments: even if one manages to define a physical system like the air that carries the signal, one could hardly keep track of the entropy contained in the entire system. The independence principle, on the other hand, does not have to account for entropies of the entire system in order to infer the time direction. To show this, we first rephrase some results from [52] and then discuss future directions using the principle of algorithmic independence.

Assume the input signal is represented by the discrete time series $(X_t)_{t\in\mathbb{Z}}$ with real-valued X_t and the output signal (the echo) $(Y_t)_{t\in\mathbb{Z}}$ is obtained from the input via convolution with the impulse response function h:

$$Y_t = \sum_j h(j) X_{t-j}.$$

In the frequency space, this amounts to multiplying the Fourier transforms of X and Y with the impulse response function $\hat{h}$:

$$\hat{Y}(\nu) = \hat{h}(\nu) \cdot \hat{X}(\nu) \quad \forall \nu \in \left[-\frac{1}{2}, \frac{1}{2}\right],$$

where the Fourier transform is defined by

$$\hat{X}(\nu) := \sum_{t\in\mathbb{Z}} e^{-i2\pi\nu t} X_t$$

and likewise for $\hat{Y}$ and $\hat{h}$. In [52] it is then postulated an independence principle stating that the power spectrum $|\hat{X}|^2$ and $|\hat{h}|^2$ do not correlate, i.e., that

$$\langle |\hat{Y}^2| \rangle \approx \langle |\hat{X}|^2 \rangle \cdot \langle |\hat{h}|^2 \rangle, \tag{10}$$

where $\langle \cdot \rangle = \int_{-1/2}^{1/2} \cdot \, d\nu$ denotes the expectation over all frequencies. A hypothetical model where $(Y_t)_{t\in\mathbb{Z}}$ is the cause and $(X_t)_{t\in\mathbb{Z}}$ the effect then assumes a mechanism whose impulse response function is $1/\hat{h}$. The absolute square of the latter is necessarily negatively correlated with $\hat{Y}$ as one can easily show (apart from non-generic exceptions) [52]. Intuitively, this is because Y tends to have high amplitudes for those frequencies for which $|\hat{h}(\nu)|^2$ is large and hence $1/|\hat{h}(\nu)|^2$ is small. In this sense, *independence* between cause and mechanism in causal direction implies *dependence* between effect and mechanism, in close analogy to the Information Geometric setting above.

Although the so-called spectral independence criterion (SIC) in (10) turned out to be helpful for the experiments performed in [52], it is not hard to see its limitations: assume, for instance, that both $\hat{X}(\nu)$ and $\hat{h}(\nu)$ mainly contain power in the region of frequencies that are close to zero. This fact alone does not show that they contain significant amount of information about each other. After all, it is the typical behavior of any time series that is not fractal that the power decays quickly for frequencies far away from zero (for sufficiently fine time discretization). Future research for more sophisticated independence criteria than (10), which are not mislead by those kind of dependences that occur without causal dependence between input and mechanism, could be guided by principle 1.

Accordingly, the basic postulate then amounts to

$$I(\hat{X} : \hat{h}) \stackrel{+}{=} 0.$$

We then obtain

$$K(\hat{X}) \stackrel{+}{\leqslant} K(\hat{Y}) \stackrel{+}{\leqslant} K(\hat{X}) + K(\hat{h}),$$

where the first inequality follows, again in analogy to the proof of theorem 1 and the proof of (9), and the second one holds because $\hat{Y}$ can be computed from $\hat{X}$ and $\hat{h}$ by a program of length $O(1)$. In the generic case, $K(\hat{Y})$ will thus be larger than $K(\hat{X})$ and we infer the more complex signal to be the echo, which provides a well-defined arrow of time. As a further aspect of this asymmetry between past and future, $1/\hat{h}$ and $\hat{Y}$ cannot be independent because applying theorem 1 in backwards direction would then imply $K(\hat{X}) \stackrel{+}{\geqslant} K(\hat{Y})$. Although algorithmic complexity is uncomputable, it is not hopeless to approximate it by compression schemes that are appropriate for specific tasks, see e.g. [53].

3. Discussion

Already Reichenbach linked asymmetries between cause and effect to the arrow of time when he argued that the statistical dependence patterns induced by causal structures $X \leftarrow Z \rightarrow Y$ (common cause) versus $X \rightarrow Z \leftarrow Y$ (common effect) naturally emerge from the time direction of appropriate mixing processes [43]. In this work we provide a new foundational principle describing additional asymmetries that appear when algorithmic rather than only statistical information is taken into account. As a consequence it follows naturally the non-decrease of entropy (if the latter is identified with algorithmic complexity) and the independence between P_{cause} and $P_{\text{effect}|\text{cause}}$, thus providing further relations between thermodynamics and causal inference.

Non-Markovian dynamics. Intuitively, our principle resembles the standard way to obtain Markovian dynamics of open systems, coupling a system to a statistically independent environment [54]. In this sense, our principle can be understood as a notion of Markovianity that is stronger in two respects: first, the initial state of

Figure 3. Left: pure ause-effect relation without common cause. Right: cause–effect relation that is confounded by a common cause.

the system is not only *statistically* but also *algorithmically* independent of the environment, and second, it is also algorithmically independent of the dynamical law. It thus provides a useful new rationale for finding the most plausible causal explanation for given observations arising in study of open systems. It is known, however, that *non*-Markovian dynamics is ubiquitous. As argued in [55], for instance, the dynamics of a quantum system interacting strongly with the environment is not Markovian because it does not start in a product state. Instead, initial state of system and environment already share information. At least for these cases, we will also expect violations of principle 1. It should be emphasized, however, that also for non-Markovian systems (for which the initial state has not been prepared independently of the environment) one is sometimes interested in the question of *what would happen to an input state if it was prepared independently*. This perspective becomes particularly clear by discussing analogies between non-Markovian dynamics and the phenomenon of *confounding* in the world of statistics and causal inference [2].

To explain this, consider just two variables X and Y where the statistical dependence is entirely due to the causal influence of X on Y. The corresponding causal relation is visualized in figure 3, left. For this relation, the observed conditional distribution $P_{Y|X}$ can be interpreted as describing also the behavior of Y under interventions on X. Explicitly, $P_{Y|X=x}$ is not only the distribution of Y after we have *observed* that X attains the value x. Instead, it also decribes the distribution of Y given that we *set* X to the value x by an external intervention. Using similar language as in [2], we write this coincidence of observational and interventional probabilities as

$$P_{Y|X=x} = P_{Y|do(X=x)}.$$

On the other hand, if the dependence between X and Y is only partly due to the influence of X on Y but also due to the common cause Z as in figure 3, right, setting X to the value x yields a different distribution than observing the value x, i.e.

$$P_{Y|X=x} \neq P_{Y|do(X=x)}.$$

On can show [2] that the interventional probability can then be computed via

$$p(y|do(x)) = \sum_z p(y|x, z)p(z) \neq \sum_z p(y|x, z)p(z|x) = p(y|x). \tag{11}$$

Assume, for instance, one observes a correlation between taking a medical drug (variable X) and recovery from a disease (variable Y). Let say, the correlation is partly because the drug helps and partly because women take the drug more often than men and are, at the same time, more likely to recover. The question of whether it is worth to take the drug needs to be based on $P_{Y|do(X=x)}$, not on $P_{Y|X}$. If the data base contains information on the gender Z, we can adjust for this confounder using (11) and obtain the interventional probabilities from the observational ones. Otherwise, finding $P_{Y|do(X=x)}$ requires randomized experiments. This example shows that although the input x is in fact not independent of the mechanism relating X and Y, we are interested in the question what would happen if we made it independent. Markovian and non-Markovian systems can be seen as the physical analogs of figure 3, left and right, respectively: a system is non-Markovian because the future state of the system is not only influenced by the present state but also by some common history or state of the environment. Like in the case of random variables, for a non-Markovian system we may be interested in what would happen for a 'typical' input state, that is, one that is prepared independently of the state of the environment and the dynamics.

Going back to the causal inference world, we should emphasize that algorithmic independence of P_X and $P_{Y|X}$ has only been postulated for the causal relation in figure 3, left, and not for the confounded scenario on the right hand side. Accordingly, confounding may be detected by dependences between P_X and $P_{Y|X}$ [15]. Likewise, for physical systems, dependences between a state and dynamics may indicate non-Markovian dynamics[12].

More generally speaking, algorithmic information has also attracted interest for the foundations of physics recently. For instance, given the recent connections between the phenomenon of quantum nonlocality [56] with

[12] For quantum systems, note that [13] discusses a condition that can also indicate common causes.

algorithmic information [57, 58] and causality [7–13], our results may also point new directions for research in the foundations of quantum physics.

Acknowledgments

Johan Åberg and Philipp Geiger made helpful remarks on an earlier version of the manuscript. RC acknowledges financial support from the Excellence Initiative of the German Federal and State Governments (Grants ZUK 43 & 81), the US Army Research Office under contracts W911NF-14-1-0098 and W911NF-14-1-0133 (Quantum Characterization, Verification, and Validation), the DFG (GRO 4334 & SPP 1798).

References

[1] Spirtes P, Glymour C and Scheines R 1993 *Causation, Prediction, and Search* (*Lecture Notes in Statistics*) (New York: Springer)
[2] Pearl J 2000 *Causality* (Cambridge: Cambridge University Press)
[3] Pearl J 2014 *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (San Mateo, CA: Morgan Kaufmann)
[4] Friedman N 2004 Inferring cellular networks using probabilistic graphical models *Science* **303** 799–805
[5] Jackson A and Scheines R 2005 Single mothers' self-efficacy, parenting in the home environment, and children's development in a two-wave study *Soc. Work Res.* **29** 7–20
[6] Moneta A, Entner D, Hoyer P and Coad A 2013 Causal inference by independent component analysis: theory and applications* *Oxford Bull. Econ. Stat.* **75** 705–30
[7] Fritz T 2012 Beyond Bell's theorem: correlation scenarios *New J. Phys.* **14** 103001
[8] Leifer M and Spekkens R 2013 Towards a formulation of quantum theory as a causally neutral theory of Bayesian inference *Phys Rev.* A **88** 052130
[9] Wood C and Spekkens R 2015 The lesson of causal discovery algorithms for quantum correlations: causal explanations of Bell-inequality violations require fine-tuning *New J. Phys.* **17** 033002
[10] Chaves R, Majenz C and Gross D 2015 Information-theoretic implications of quantum causal structures *Nat. Commun.* **6** 5766
[11] Chaves R, Kueng R, Brask J B and Gross D 2015 Unifying framework for relaxations of the causal assumptions in Bell's theorem *Phys. Rev. Lett.* **114** 140403
[12] Henson J, Lal R and Pusey M 2014 Theory-independent limits on correlations from generalized Bayesian networks *New J. Phys.* **16** 113043
[13] Ried K, Agnew M, Vermeyden L, Janzing D, Spekkens R and Resch K 2015 A quantum advantage for inferring causal structure *Nat. Phys.* **11** 414–20
[14] Mooij J M, Peters J, Janzing D, Zscheischler J and Schölkopf B 2016 Distinguishing cause from effect using observational data: methods and benchmarks *J. Mach. Learn. Res.* **17** 1–102
[15] Janzing D and Schölkopf B 2010 Causal inference using the algorithmic Markov condition *IEEE Trans. Inf. Theory* **56** 5168–94
[16] Lemeire J and Janzing D 2012 Replacing causal faithfulness with algorithmic independence of conditionals *Minds Mach.* **23** 227–49
[17] Zurek W 1989 Algorithmic randomness and physical entropy *Phys. Rev.* A **40** 4731–51
[18] Vapnik V 1998 *Statistical Learning Theory* (New York: Wiley)
[19] Chapelle O, Schölkopf B and Zien A 2010 *Semi-Supervised Learning* (Cambridge, MA: MIT Press)
[20] Schölkopf B, Janzing D, Peters J, Sgouritsa E, Zhang K and Mooij J 2012 On causal and anticausal learning *Proc. 29th Int. Conf. on Machine Learning (ICML)* ed J Langford and J Pineau (New York: ACM) pp 1255–62
[21] Schölkopf B, Janzing D, Peters J, Sgouritsa E, Zhang K and Mooij J 2013 Semi-supervised learning in causal and anticausal settings *Empirical inference, Festschrift in honor of Vladimir Vapnik* ed B Schölkopf, Z Luo and V Vovk (Berlin: Springer) pp 129–41
[22] Peters J, Janzing D, Gretton A and Schölkopf B 2009 Detecting the direction of causal time series *Proc. 26th Int. Conf. on Machine Learning ACM (Montreal) International Conference Proceeding Series* (New york: ACM) pp 801–8
[23] Pickup L C, Pan Z, Wei D, Shih Y, Zhang C, Zisserman A, Schölkopf B and Freeman W T 2014 Seeing the arrow of time *2014 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)* pp 2043–50
[24] Janzing D 2010 On the entropy production of time series with unidirectional linearity *J. Stat. Phys.* **138** 767–79
[25] Kolmogorov A 1965 Three approaches to the quantitative definition of information *Problems Inform. Transm.* **1** 1–7
[26] Chaitin G 1975 A theory of program size formally identical to information theory *J. Assoc. Comput. Mach.* **22** 329–40
[27] Li M and Vitányi P 1997 *An Introduction to Kolmogorov Complexity and Its Applications* 3rd edn (New York: Springer) 2008
[28] Barron A and Cover T 1991 Minimum complexity density estimation *IEEE Trans. Inf. Theory* **37** 1034–54
[29] Bennett C 1982 The thermodynamics of computation—a review *Int. J. Theor. Phys.* **21** 905–40
[30] Cover T and Thomas J 1991 *Elements of Information Theory* (*Wileys Series in Telecommunications*) (Hoboken, NK: Wiley)
[31] Nielsen M and Chuang I 2000 *Quantum Computation and Quantum Information* (Cambridge: Cambridge University Press)
[32] Jaynes E T 1965 Gibbs vs. Boltzmann entropies *Am. J. Phys.* **33** 391–8
[33] Caves C M 1993 Information and entropy *Phys. Rev.* E **47** 4010–7
[34] Caves C 1994 Information, entropy and chaos *Physical Origins of Time Asymmetry* ed J Halliwell *et al* (Cambridge: Cambridge University Press) pp 47–89
[35] Balian R 1992 *From Microphysics to Macrophysics* (Berlin: Springer)
[36] Zeh H D 2001 *The Physical Basis of the Direction of Time* (Berlin: Springer)
[37] Deutsch D 1997 *The Fabric of Reality* (London: Penguin Books)
[38] Janzing D and Beth T 2001 Complexity measure for continuous time quantum algorithms *Phys. Rev.* A **64** 022301
[39] Wocjan P, Rötteler M, Janzing D and Beth T 2002 Simulating Hamiltonians in quantum networks: efficient schemes and complexity bounds *Phys. Rev.* A **65** 042309
[40] Janzing D 2007 Spin-1/2 particles moving on a 2D lattice with nearest-neighbor interactions can realize an autonomous quantum computer *Phys. Rev.* A **75** 012307
[41] Gemmer J, Michel M and Mahler G 2005 *Quantum Thermodynamics: Emergence of Thermodynamic Behavior Within Composite Quantum Systems* (Berlin: Springer)
[42] Mora C, Kraus B and Briegel H 2007 Quantum Kolmogorov complexity and its applications *Int. J. Quantum Inform.* **5** 729–50

[43] Reichenbach H 1956 *The Direction of Time* (Berkeley, CA: University of California Press)
[44] Solomonoff R 1960 A preliminary report on a general theory of inductive inference *Technical report V-131* ZTB-138 Zator Co.
[45] Levin L 1974 Laws of information conservation (non-growth) and aspects of the foundation of probability theory *Problems Inf. Trans.* **10** 206–10
[46] Allahverdyan A and Janzing D 2008 Relating the thermodynamic arrow of time to the causal arrow *J. Stat. Mech.* **4** P04001
[47] Kano Y and Shimizu S 2003 Causal inference using nonnormality *Proc. Int. Symp. on Science of Modeling—The 30th Anniversary of the Information Criterion (Tokyo, Japan)* pp 261–70
[48] Janzing D and Steudel B 2010 Justifying additive-noise-based causal discovery via algorithmic information theory *Open Syst. Inf. Dyn.* **17** 189–212
[49] Daniusis P, Janzing D, Mooij J M, Zscheischler J, Steudel B, Zhang K and Schölkopf B 2010 Inferring deterministic causal relations *Proc. 26th Annual Conf. on Uncertainty in Artificial Intelligence (UAI)* pp 143–50
[50] Janzing D, Mooij J, Zhang K, Lemeire J, Zscheischler J, Daniušis P, Steudel B and Schölkopf B 2012 Information-geometric approach to inferring causal directions *Artif. Intell.* **182–183** 1–31
[51] Frisch U, d'Humiéres D and Hasslacher B 1987 Lattice gas hydrodynamics in two- and three-dimensions *Complex Syst.* **1** 649–707
[52] Shajarisales N, Janzing D, Schölkopf B and Besserve M 2015 Telling cause from effect in deterministic linear dynamical systems *Proc. 32th Int. Conf. on Machine Learning (ICML)* pp 285–94
[53] Steudel B, Janzing D and Schölkopf B 2010 Causal Markov condition for submodular information measures *Proc. 23rd Annual Conf. on Learning Theory (COLT)* pp 464–76
[54] Lindblad G 1976 On the generators of quantum dynamical semigroups *Commun. Math. Phys.* **48** 119–30
[55] Pechukas P 1994 Reduced dynamics need not be completely positive *Phys. Rev. Lett.* **73** 1060–2
[56] Brunner N, Cavalcanti D, Pironio S, Scarani V and Wehner S 2014 Bell nonlocality *Rev. Mod. Phys.* **86** 419–78
[57] Poh H, Markiewicz M, Kurzyński P, Cerè A, Kaszlikowski D and Kurtsiefer C Probing quantum–classical boundary with compression software *New J. Phys.* **18** 035011
[58] Bendersky A, de la Torre G, Senno G, Figueira A and Acin S 2016 Algorithmic pseudorandomness in quantum setups *Phys. Rev. Lett.* **116** 230402