# A Mathematical Theory of Knowledge, Science, Bias and Pseudoscience

Daniele Fanelli[1],

**1 Meta-Research Innovation Center at Stanford (METRICS), Stanford University, Palo Alto, CA, USA.**

*** email@danielefanelli.com**

## Abstract

This essay unifies key epistemological concepts in a consistent mathematical framework built on two postulates: 1-information is finite; 2-knowledge is information compression. Knowledge is expressed by a function $K(Y;X)$ and two fundamental operations, $\oplus, \otimes$. This $K$ function possesses fundamental properties that are intuitively ascribed to knowledge: it embodies Occam's razor, has one optimal level of accuracy, and declines with distance in time. Empirical knowledge differs from logico-deductive knowledge solely in having measurement error and therefore a "chaos horizon". The $K$ function characterizes knowledge as a cumulation and manipulation of patterns. It allows to quantify the amount of knowledge gained by experience and to derive conditions that favour the increase of knowledge complexity. Scientific knowledge operates exactly as ordinary knowledge, but its patterns are conditioned on a "methodology" component. Analysis of scientific progress suggests that classic Popperian falsificationism only occurs under special conditions that are rarely realised in practice, and that reproducibility failures are virtually inevitable. Scientific "softness" is simply an encoding of weaker patterns, which are simultaneously cause and consequence of higher complexity of subject matter and methodology. Bias consists in information that is concealed in ante-hoc or post-hoc methodological choices. Disciplines typically classified as pseudosciences are sciences expressing extreme bias and therefore yield $K(Y;X) \leq 0$. All knowledge-producing activities can be ranked in terms of a parameter $\Xi \in (-\infty, \infty)$, measured in bits, which subsumes all quantities defined in the essay.

## Author Summary

- Knowledge is just information compression.
- Science is just knowledge conditioned upon methodology.
- "Soft" science is just relatively "weak" science.
- Bias is just misplaced information.
- Pseudoscience is just extreme bias.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v1 | CC-BY 4.0 Open Access | rec: 19 Apr 2016, publ: 19 Apr 2016

# Contents

0.1 Postulate 1: Information is finite . . . . . 4
0.2 Postulate 2: Knowledge is information compression . . . . . 4
0.3 Mathematization of knowledge . . . . . 6
0.4 Statistical interpretation . . . . . 9
0.5 Alternative forms of K . . . . . 11
0.6 Operations on Information . . . . . 12
0.7 Shannon Entropy with ⊗ and ⊕ . . . . . 16
0.8 Properties of K . . . . . 20
0.9 Knowledge . . . . . 24
0.10 Knowledge growth . . . . . 26
0.11 Science . . . . . 28
0.12 Scientific progress . . . . . 34
0.13 Reproducibility . . . . . 37
0.14 Soft science . . . . . 38
0.15 Bias . . . . . 40
0.16 Pseudoscience . . . . . 44
0.17 Hierarchy of the Sciences (and Pseudosciences) . . . . . 46
0.18 Summary of findings . . . . . 47
0.19 Predictions and conclusions . . . . . 49

# Introduction

A science of science is flourishing in all disciplines and promises to boost discovery on all research fronts [1]. This growing literature of empirical studies, intervention experiments and theoretical models intensifies interest in a cross-disciplinary, quantitative and operationalizable theory of ”good” and ”bad” science.

Textbook philosophy of science offers little guidance for meta-researchers. Despite a century of debate, no consensus has been reached on the criteria that demarcate genuinely scientific knowledge from metaphysics or pseudoscience (Table 1). Indeed, the very existence of such universal criteria has been dismissed as a pseudo-problem. A popular view amongst contemporary philosophers postulates that “science” is just a word indicating a variety of practices. These practices bear a “family resemblance” to each other but do not share a single universal property. To determine what is valid science, according to this approach, we must use multiple criteria e.g. [2–4].

The multi-criterial solution to the demarcation problem is of limited theoretical and practical utility. It merely shifts the question from identifying a single property common to all the sciences to identifying many properties that are common to some. Inevitably, there is little consensus on what these properties consist in, and most proposals include normative or behavioural (i.e. subjective) principles such as ”rigorously assessing evidence”, ”openness to criticism”, etc... Furthemore, since the minimum number of characteristics that a legitimate science should possess is arbitrary, virtually any practice can be considered a ”science” according to one scheme or another (e.g. intelligent design [13]).

In addition to potentially legitimizing pseudoscience, the family-resemblance approach does little to explain the diversity of scientific disciplines. There is ample evidence that scientific practices vary gradually and almost linearly if disciplines are arranged according to the complexity of phenomena they are concerned with (i.e. broadly speaking, mathematics, physical, biological, social sciences and humanities) [24–27]. This suggests that general principles underlie and explain at least

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v1 | CC-BY 4.0 Open Access | rec: 19 Apr 2016, publ: 19 Apr 2016

**Table 1. Demarcation theories**

| principle | science | non-/pseudoscience | author, ref |
|---|---|---|---|
| positivism | reached the positive stage: builds knowledge on empirical data | still in teological or meta-physical stages: phenomena are explained by deities or non-observables | Comte 1830 [5] |
| methodologism | follows rigorous methods for selecting hypotheses, acquiring data and drawing conclusions | fails to follow the scientific method | e.g. Pearson 1900, Poincare 1914 [6, 7] |
| verificationism | builds upon verified statements | relies on non-verifiable statements | Wittgenstein 1922 [8] |
| falsificationism | builds upon falsifiable, non falsified statements | produces explanations devoid of verifiable counterfactuals | Popper 1959 [9] |
| methodological falsificationism | generates theories of increasing empirical content, which are accepted when surprising predictions are confirmed | protects its theories with a growing belt of auxilliary hypotheses, creating "degenerate" research programs | Lakatos 1970 [10] |
| norms | follows four fundamental norms: universalism, communism, disinterestedness, organized scepticism | operates on other, if not opposite, sets of norms | Merton 1942 [11] |
| paradigm | is post-paradigmatic: solves puzzles defined and delimited by the rules of an accepted paradigm | is pre-paradigmatic: lacks a unique and unifying intellectual framework or is fragmented into multiple competing paradigms | Kuhn 1974 [12] |
| multi-criterial approaches | bears a sufficient "family resemblance" to other activities we call "science" | shares too few characteristics with activities that we consider scientific | e.g. Laudan 1983, Dupre 1993, Pigliucci 2013 [2–4] |

**Table 2. Properties variably possessed by sciences**

| principle | property | author, ref |
|---|---|---|
| scientific hierarchy | simplicity, generality, quantifiability, recency, human relevance | Comte 1830 [5] |
| consilience | ability to subsume disparate phenomena under general principles | Whewell 1840 [14] |
| lawfulness | nomoteticity, i.e. interest in uncovering general laws, as opposed to idioteticity, i.e. interest in individuality | Windelband 1894 [15] |
| data hardness | data that resist the solvent influence of critical reflection | Russel 1914 [16] |
| empiricism | ability to calculate in advance the results of an experiment | Conant 1951 [17] |
| rigour | rigour in relating data to theory | Storer 1967 [18] |
| maturity | ability to produce and test mechanistic hypotheses, as opposed to mere fact collection | Bunge 1967 [19] |
| cumulativity | cumulation of knowledge in tightly integrated structures | Price 1970 [20] |
| codification | consolidation of empirical knowledge into succinct and interdependent theoretical formulations | Zuckerman and Merton 1973 [21] |
| consensus | levels of consensus on the significance of new knowledge and the continuing relevance of old | Zuckerman and Merton 1973 [21] |
| core cumulativity | rapidly growing core of unquestioned general knowledge | Cole 1983 [22] |
| invariance | contextual invariance of phenomena | Humphreys 1990 [23] |

some of disciplines' diversity. The source and nature of these differences have been insightfully discussed in the literature (Table 2) but no consensual view has emerged.

This essay proposes a quantitative theory of science, which unifies in a coherent and consistent mathematical framework a variety of fundamental concepts including knowledge, bias, reproducibility, soft-science and pseudoscience. The theory rests on two postulates: 1) information is finite; 2) knowledge is information compression. The following section will introduce and justify these postulates. Subsequently, the Methods section will give mathematical expression to the concept of knowledge and knowledge progress. The Results section will first show that the proposed functions embody properties expected of knowledge, and will then proceed with a brief analysis of each concept in turn. The aim of this essay is to present the logic of the approach proposed and to illustrate its potential uses. Complete analyses of specific issues are left to future work.

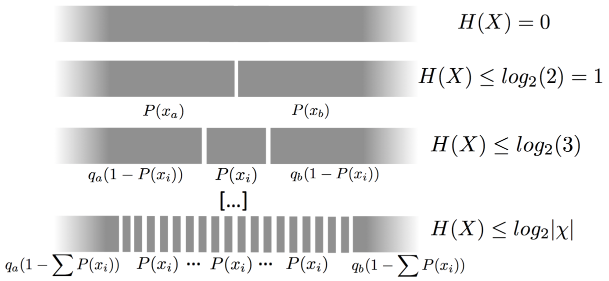

**Figure 1.** Pictorial representation of the nature of information, which can be of arbitrary size and yet always be finite. Attributes are represented by bars that have no defined ends. Information emerges from discontinuities which are always countable and finite

.

## 0.1 Postulate 1: Information is finite

The first postulate appears to reflect a simple but easily overlooked fact of nature. The universe - at least, the portion of it that we can see and have causal connection to - contains finite amounts of matter and energy, and therefore cannot contain infinite amounts of information. If each quantum state represents a bit, and each transition between (orthogonal) states represents an operation, then the universe has performed circa $10^{120}$ operations on $10^{90}$ bits since the Big Bang [28].

Even if our understanding of physics turned out to be incorrect, there would be still little doubt that information is finite for living organisms, because sensory organs are finite structures. The sense of vision is constructed out of a limited number of cone and rod cells; the sense of hearing is uses information from a limited number of hair cells, each of which responds to a narrow band of acoustic frequencies, etc.

But even if we ignored both physical and biological arguments, the finitude of information would be a reasonable assumption to make for scientific knowledge, because measurement, by definition, is finite. Measurement is technically defined as the partitioning of attributes in a set of mutually exclusive categories [29]. In principle this partitioning can be infinite, but in practice it never is. Real measurements of empirical phenomena always contemplate a range of plausible values and are delimited at one or both ends by extreme values that capture all residual probability.

We can picture the process of measurement, and in general the generation of information, as a progressive quantization of a unidimensional attribute. This quantization operates ”from the inside out” so to speak, and by necessity always leaves two ”open ends” of finite probability (Fig 1).

## 0.2 Postulate 2: Knowledge is information compression

The second postulate claims that the essence of any manifestation of what we call ”knowledge” consists in the encoding of a pattern, which reduces the amount of information required to navigate the world successfully. Patterns are simply dependencies between attributes, in other words relationships that makes one event (the manifestation of one of possible instantiations of an attribute) more or less likely depending on another event. By encoding patterns, an organism reduces the uncertainty it confronts about its environment - in other words it *adapts*. Therefore, just like postulate 1, postulate 2 also is likely to reflect an elementary fact of nature, which is arguably the essence not just of human knowledge, but of life itself.

The idea that knowledge, or at least scientific knowledge, is information compression is far from new. Physicist and philosopher Ernst Mach famously argued, for example, that the value of physical laws lied in the “economy of thought” that they permitted [30]. Other famous scientists such as mathematician Henri Poincaré (1854-1912) expressed similar ideas [7]. Following the discovery of information theory, scientific knowledge and other cognitive activities have been examined in quantitative terms e.g. [31]. Nonetheless, the equivalence between scientific knowledge and information compression has been dismissed as a principle of secondary importance by later philosophers (e.g. Karl Popper 1902–1994 [32]), and today clearly does not occupy the foundational role that it arguably deserves [33].

Philosophical resistance to equating science with information compression might partially be explained by two common misconceptions. The first one is an apparent conflation of lossy and lossless compression. Modern proponents of the hypothesis seem to adopt a lossless compression model and therefore debate over whether empirical data truly are compressible e.g. [34]. Clearly, however, science is a lossy form of compression: the laws and relations that scientists discover typically include error terms and tolerate large portions of unexplained variance.

The second, and most important, source of misplaced skepticism is a lack of appreciation for the universality of the equivalence between knowledge and information compression. As already mentioned, the encoding of patterns underlies not just scientific knowledge but all forms of knowledge and all forms of biological adaptation. Changes in any species’ population genetic frequencies in response to environmental pressures can be seen as a form of adaptive learning, in which a signal ”reinforces” (by selecting favourably) certain responses, i.e. specific genomic structures, and ”weakens” (selects out) others. Endocrine, immune and nervous systems are simply more advanced pattern-encoding structures, which operate on a faster scale than natural (genetic) selection.

Human cognition is just another higher-order manifestation of biological pattern encoding, not qualitatively but quantitatively superior to other forms. When we say that we “know” something, we are claiming that we have fewer uncertainties about it. We “know a city”, for example, if and in proportion to how we are able to navigate in it. We “know a song” when we are able to reproduce with no error or hesitation the sequence of words and intonations that will recreate it. We “know a person” in proportion to how many patterns about them we have encoded. We might, for example, only be able to relate their facial features to their name. When we know them better, however, we can predict how they might respond to the question “where are you from?”. When we know them well, we can predict their behaviour rather accurately and foretell what will make them happy, interested, etc.

Scientific knowledge is just another expression of human knowledge, and as such is again nothing more than a pattern-encoding activity that reduces uncertainty about one phenomenon by relating it to information about another phenomenon. The knowledge produced by all fields of scientific research is structured in this way. A mathematical theorem consists in drawing a logical connection between two seemingly unrelated theoretical constructs. The laws of physics are obviously describing patterns, but even research that appears purely descriptive, for example the measurement of values of physical constants, consists in mapping and connecting properties of known objects. Most biological and biomedical research consists in identifying correlations, causes and/or describing properties of natural phenomena. Research in taxonomy and systematics might appear to be an exception, but it is not: organizing a multitude of species into a succint taxonomical tree is the most elementary form of data compression. Quantitative social and behavioural sciences operate in a similar fashion to biological sciences; and even qualitative, ethnographic, purely descriptive social research consists in data compression, because it presupposes that there are general facts about human experiences, individuals or groups that can be described and synthesized.

# Analysis

## 0.3 Mathematization of knowledge

Information theory offers a straightforward and universal measure of the pattern linking two variables $X, Y$: the Mutual Information function

$$I(X;Y) = H(Y) - H(Y|X) \qquad (1)$$

in which H(X) is Shannon's entropy function:

$$H(X) = -\sum_{x} p(x)log(p(x)) \qquad (2)$$

with $X$ being a discrete random variable. A mathematically equivalent formulation to 4 is

$$I(X;Y) = H(Y) + H(X) - H(Y,X) \qquad (3)$$

To quantify and compare knowledge across the sciences we need to modify this equation, allowing the function to be:

1. Standardized: In order to allow meaningful comparisons between different domains of knowledge and/or of knowledge growth, we need our quantities to be scale-free.

2. Accuracy-dependent: Following the first postulate, the objects of knowledge (both terms in the equation) should be quantizable to arbitrary degrees.

3. Multi-dimensional: For similar reasons to the above, the objects of knowledge need not be confined to two variables. Knowledge always connects two sets of attributes, but each of these can include an arbitrary number of dimensions, i.e. attributes of different classes.

4. Time-dependent: Following the second postulate, if knowledge is proportional to the ability to anticipate events, then its function must accommodate varying levels of "distance" between events. This distance is probably spatio-temporal. However, for the purposes of this essay, we will consider this separation to be exclusively on a time dimension, which lends itself to an intuitive interpretation.

These three considerations lead to define a function K(Y;X) ("K" for "Knowledge"), which in its most complete form is written as

$$K(Y;X) \equiv K(\underline{Y}_t^{m;\underline{\alpha_y}}, \underline{X}_{t_0}^{n,\underline{\alpha_x}}) = \frac{H(\underline{Y}_t^{m,\underline{\alpha_y}}) - H(\underline{Y}_t^{m,\underline{\alpha_y}}|\underline{X}_{t_0}^{n,\underline{\alpha_x}})}{H(\underline{Y}_t^{m,\underline{\alpha_y}}) + H(\underline{X}_{t_0}^{n,\underline{\alpha_x}})} \qquad \text{(knowledge function)}$$

in which the underlined terms represent vectors, and subscripts and superscripts indicate fixed properties of these vectors. For example, $\underline{Y}_t^{m,\underline{\alpha_y}}$ represents a random vector of length $m$, whose elements correspond to the joint distribution of $m$ random variables (attributes), each of which was measured at a time $t$ with accuracies represented by the m-length vector $\underline{\alpha_y} : \{\alpha_{y_1}, \alpha_{y_2}, ...\alpha_{y_m}\}$ , i.e. $\underline{Y}_t^{m,\underline{\alpha_y}} = \{Y_{1,t}^{\alpha_{y_1}}, Y_{2,t}^{\alpha_{y_2}}, ...Y_{m,t}^{\alpha_{y_m}}\}$.

This notation is burdensome, and for practical purposes will always be reduced to the minimum necessary. Unless specified otherwise, we will assume

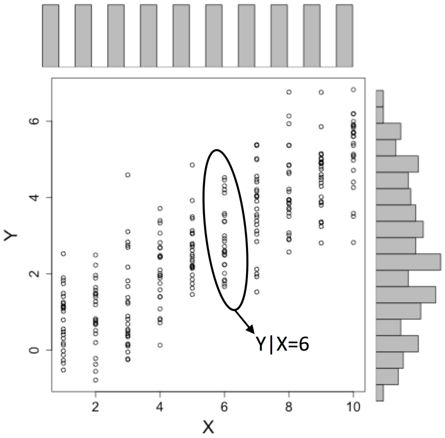

**Figure 2.** Visual representation of the conditioning of a random variable, $Y$, on a specific value of a second variable $X = x$, what in the essay we represent as $Y|x$.

$m = 1, n = 1, \alpha_x = constant, \alpha_y = constant, t_0 = 0, d_0 = 0, t > 0, d = constant$ and thus whenever possible we will use $Y$ in place $\underline{Y}_t^{m,\underline{\alpha_y}}$ and $X$ in place of $\underline{X}_{t_0}^{n,\underline{\alpha_x}}$, leading to:

$$K(Y;X) = \frac{H(Y) - H(Y|X)}{H(Y) + H(X)} \tag{4}$$

Since $H(X) + H(Y|X) = H(YX)$, equation 4 (henceforth referred to as ”K function”) can also be written as $1 - \frac{H(YX)}{(H(Y)+H(X)}$. The term $H(Y)$ will be referred to as the *explanandum*, latin for ‘what is to be explained’, in other words the phenomenon about which there is uncertainty. The $H(X)$ term will be referred to as the *explanans*, i.e. ‘what explains’, and represents the cue, the signal, the model, and any other entity that reduces the uncertainty of the explanans. The combination of $YX$ constitutes what we will refer to as a *system*. The denominator of equation 4, i.e. the sum of the entropies of explanans and explanandum, will be called *uncertainty space*. The pattern linking explanandum and explanans, which determines the extent to which the system is knowable, is quantified by the conditional term $H(Y|X)$ which, depending on the context, will be referred to as the *pattern*, the *law* or the *hypothesis*.

Explanans and explanandum need not be classic random variables. Indeed, it is crucial that they may consist in individual entities of non-zero probability, which are best imagined as single instantiations (events, outcomes) of classic random variables. We will refer to such entities as ”events” or ”objects” and represent them with lower-case letters. The distinction between random variables and events or objects is crucial to the mathematics of the K function. If explanans $X$ and explanandum $Y$ are classic random variables, then $I(Y;X) \leq \min(H(Y);H(X))$ and therefore $0 \leq K(Y;X) \leq \min(H(Y);H(X)/(H(Y)+H(X))$. However, if $X = x$ and $Y = y$, with $x$ and $y$ representing individual outcomes with probability $P(x)$ and $P(y)$, then their mutual information may assume any value, $I(y;x) \in (-\infty, \infty)$ and therefore $K(y;x) \in (-1, 1)$. Combined cases may also occur: $K(Y;x)$ and $K(y;X)$. A pattern $H(Y;x)$, for example, can be pictured as the conditional distribution of a random variable $Y$ conditioned on $X = x$ (Fig 2).

Note, however, that if $H(y) = 0$ or $H(x) = 0$ or $H(y) = \infty$ or $H(x) = \infty$, then $K(y,x) = 0$. Knowledge, in other words, requires uncertainty in both explanans and explanandum to be finite and different from zero. The first postulate guarantees that this is always the case. From a mathematical point of view, imposing a quantization for the objects of the $K$ function implies that we cannot use Shannon's differential entropy function $h(X) = \int f(x) log f(x) dx$. Although it superficially looks like the continuous equivalent of equation (2), differential entropy has a fundamentally different meaning and different properties from its discrete counterpart [35]. In particular, the differential entropy of a continuous distribution is *not* the limit of this quantization for infinitesimally small values. Such limit actually equals $-\infty$, revealing the existence of an unbridgeable gap between discrete and differential entropy functions. Moreover, the differential entropy of a continuous density function is scale-dependent and can assume negative values. Plugged into the K function, differential entropy yields incongruous results that further justify its exclusion from the theory - at least in its present form.

**Complexity as information** Shannon's entropy is a measure of average uncertainty or, equivalently, of the rarity of events. It is the expectation of the function $log\frac{1}{p(x)}$, which quantifies the rarity of the outcome $x$ of $X$, and therefore the information that event $x$ conveys by means of its probability.

The nature of event $x$, in particular its structure, is irrelevant to its Shannon's information, but it too is a source of information. This latter information, quantifiable as the minimum information necessary to reproduce the object itself, is measured by Kolmogorov complexity [36,37]. Let $x$ be a binary string (e.g. a sequence 010010110...), $U$ a universal Turing machine (i.e. akin to a general-purpose computer), $\pi(x)$ a computer program that recreates $x$ and $l(\pi)$ is the length of the program. The Kolmogorov complexity of $x$ is defined as:

$$C(x) = \min_{\pi:U(\pi)=x} l(\pi) \tag{5}$$

i.e. it is the length of the shortest program that prints $x$ and halts. Postulate n.1 guarantees that any object or event has a finite description. Therefore, since objects can be described by computer programs, since programs can be translated into one another and since Turing machines can simulate any other computer, Kolmogorov's complexity is a universal quantity. It is non-computable, but is approximated by data compression algorithms.

Albeit conceived independently, Kolmogorov complexity and Shannon's entropy turn out be closely related. The latter represents the lower limit of the former:

$$\lim_{n\to\infty} \frac{E(C(x))}{n} = H(X) \tag{6}$$

For example, a binary string of length $n$ will have Kolmogorov complexity approximately equal to $n$ times the entropy of the probability distribution of $0's$ and $1's$ in the string, i.e. $H(X)$. Even more remarkably, Kolmogorov complexity turns out to be also definable in ways that are mathematically analogous to Shannon's entropy, because:

$$C(x) \approx \log \frac{1}{P_U(x)} \tag{7}$$

in which $P_U(x)$ is the universal probability of $x$ [36], defined as the probability that a randomly drawn program (e.g. a string of 1's and 0's compiled by flipping a coin) would print the string $x$ and halt:

$$P_U(x) = \sum_{\pi:U(\pi)=x} 2^{-l(\pi)} = Pr(U(\pi) = x) \tag{8}$$

This mathematical equivalence between information and algorithmic complexity can made fully explicit by introducing the notion of "Total Information".

**Total information** Let $X : \{x_1, x_2, ..., x_n\}$ be a set of $n$ objects of Kolmogorov complexities $C(x_1), C(x_2)...C(x_n)$. We define as "Total Information Content" of $X$, $T(X)$, the sum of the complexities of its components: $T(X) \equiv \sum C(x)$. Clearly:

$$\sum_x C(x) \approx \sum_x log \frac{1}{P_U(x)} = log \frac{1}{P_U(x_1) \times P_U(x_1) \times .... \times P_U(x_n)} \equiv log \frac{1}{P_T(X)} \quad (9)$$

in which $P_T(X)$ is just a new probability value. Therefore, the total information contained in a set of objects can be represented as a single object or event of non-zero and finite probability. At the same time, however, indicating with $E[\;]$ the expectation function and with $P(x)$ the frequency-derived probability of each class of objects in $X$:

$$\sum_x C(x) \approx n \times E[C(x)] = n \times \sum_x P(x) log \frac{1}{P_U(x)} =$$
$$= n \times (\sum_x P(x) log \frac{1}{P(x)} + \sum_x P(x) log \frac{P(x)}{P_U(x)}) \equiv n \times (H(X) + D(X||X_U)) \quad (10)$$

where $H(X)$ is Shannon's entropy and $D(X||X_U)$ is the Kullback-Leibler distance between the probability distribution $P(x)$ and the universal probability $P_U(x)$. Since Shannon's entropy is the limit of compressibility of $X$, $P_U(x) \leq P(x)$ and thus $D(X||X_U) \geq 0$. Therefore, the same quantity, the total information of a set of objects, can be represented as a single entity of non-zero probability and as a combination of Shannon entropy and "residual complexity" terms, multiplied by a constant.

In sum, virtually identical properties and mathematical treatments underlie the concepts of information and complexity. Since the $K$ function is a standardized quantity, its calculation is identical whether its terms are based on Shannon's entropy, Kolmogorov's complexity, or Total Information (if we plug 10 in the K function, the $n$ terms cancel out). This is true for *all the results presented in this essay*. For practical purposes most analyses in the essay will be based on Shannon's entropy notation as it applies to classic random variables. Explicit reference to Kolmogorov complexity (henceforth, simply complexity) or Total information will be made only when necessary.

## 0.4 Statistical interpretation

The $K$ function is compatible with both a subjectivist (Bayesian) and a frequentist interpretation of probability.

**Bayesian interpretation** Bayes' theorem is a simple equivalence stating that:

$$P(Y|X) = \frac{P(Y) \times P(X|Y)}{P(X)} \quad (11)$$

Bayesian interpretation of the $K$ function would posit that $H(Y)$ is the prior belief about the probability (or probability distribution) of an event, $H(Y|X)$ is the posterior belief and $K(Y;X)$ quantifies the knowledge gained after new information $X$ is obtained. We can plug Bayes' equivalence in the conditional entropy function and obtain:

$$H(Y|X) = -\sum P(yx) log P(y|x) = -\sum P(yx) log \frac{P(y) \times P(x|y)}{P(x)} =$$
$$= -\sum P(yx)[log P(y) + log P(x|y) - log P(x)] = H(Y) + H(X|Y) - H(X) \quad (12)$$

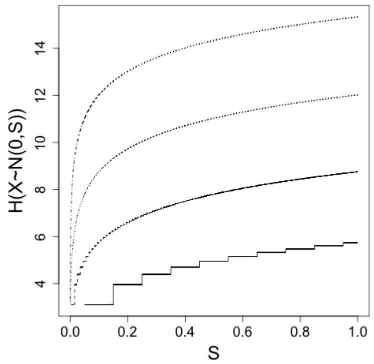

**Figure 3.** Relation between the standard deviation of a normal distribution and the corresponding Shannon entropy. The latter is a monotonic function of the former, but its values vary depending on the accuracy with which entropy is measured.

Therefore, plugging Bayes' theorem into the $K$ function yields:

$$\begin{aligned} K(Y;X) &= \frac{H(Y) - H(Y|X)}{H(Y) + H(X)} = \\ &= \frac{H(Y) - H(Y) - H(X|Y) + H(X)}{H(Y) + H(X)} = \frac{H(X) - H(X|Y)}{H(Y) + H(X)} = K(X;Y) \end{aligned} \tag{13}$$

Hence, the equivalence contained in Bayes' theorem is also contained in the K function, and is expressed in K's property of symmetry: $K(Y;X) = K(X;Y)$. The knowledge that we gain about a hypothesis given some data is equivalent to the knowledge gained about that data given a hypothesis.

**Frequentist interpretation** The $K$ function lends also itself to a non-subjectivist interpretation, in which both explanandum and explanans are classic random variables (or instantiations of them). For example, the explanandum may correspond to a dependent variable and the explanans to the independent variable in regression analysis, and the pattern (conditional entropy term) would quantify the residual (unexplained) variance. Indeed, if we removed the $H(X)$ term from the denominator of the K function, the resulting equation would resemble a generalized non-directional measure of statistical effect size such as Pearson's $\eta^2$:

$$\eta^2 = \frac{\sigma^2_Y - \sigma^2_{Y|X}}{\sigma^2_Y} \tag{14}$$

in which $\sigma^2_Y$ is the total variance of the dependent variable and $\sigma^2_{Y|X}$ is its conditional counterpart.

Unlike $\eta^2$, K is based on the entropy function, which gives it distinctive mathematical properties (Fig 3). However, the structural similarity suggests that $\eta^2$ (and similar non-directional measures of statistical association) can be conceptualized as analogous to the $K$ function proposed here, and in particular analogous to special forms of it in which the information costs of the predictor variables and the theoretical appratus (e.g. the specification of a statistical model) are ignored. This particular form of K is rather useful and will recur in this essay. We will refer to it as "uncorrected K", and indicate it with a lower-cap $k$:

$$k(Y;X) \equiv \frac{H(Y) - H(Y|X)}{H(Y)} \tag{Uncorrected K}$$

## 0.5 Alternative forms of K

All manifestations of what we call knowledge can be quantified by a function with the structure of the K function. In particular,

**Explanatory Knowledge** An event or a phenomenon is mysterious in proportion to how improbable or unpredictable it is, and is said to be explained in proportion to how much more likely it becomes once other circumstances (explanans) are invoked. Translating into our notation, let $y$ be an explanandum that occurs at time $t_0$. An explanation of it would consist in postulating the subsistence of events (or objects) that occurred at time $t \leq t_0$, whose explanatory power is measured as:

$$K(y_{t_0}; x_t) = \frac{H(y_{t_0}) - H(y_{t_0}|x_t)}{H(y_{t_0}) + H(x_t)} \qquad \text{(Explanatory K)}$$

This is simply the function 4 applied to individual objects or events, and with particular time values. Whilst $K(Y;X)$ quantifies knowledge, the overall ability to explain one or a class of phenomena, $K(y;x)$ quantifies the *understanding* about a specific event or phenomenon (see section 0.9).

It might be objected that explanations require causes, and therefore that the pattern $H(Y|X)$ is of a special kind. Causes, however, as argued below, are just patterns, and are quantifiable as such.

**Causal Knowledge** The word “cause” indicates an event that will give rise to another event, a concept that is as intuitive to human cognition as it is hard to define logically and quantify empirically. A necessary condition for causation, recognized by all definitions, is temporal asymmetry: causes always precede their effects. However, temporal asymmetry alone is an insufficient to distinguish causation from correlation or indeed from mere coincidence [38, 39].

In systems that can be extensively manipulated, causal knowledge is achieved by stabilizing or randomizing all non-relevant factors, and altering the state of the explanans of interest - what we call ”experiment”. Fields of knowledge that cannot rely on direct experimentation, such as many medical and social sciences, have developed additional criteria to identify causality. These criteria include strength, consistency, specificity, plausibility, coherence, analogy [40] - all factors that, in essence, require the pattern to be large and relatively independent from contingent aspects of the system.

Whether causation can be proven or even conceived of without recurring to experimental evidence is the subject of a growing debate [41]. However, the difference between experimental and non-experimental approaches to proving causation might be less substantive than it seems. A causal link is just a peculiarly reliable pattern; one that entails that, if the state of an explanans is known at time $t_0$, then the state of the explanandum is expected to occur with higher probability at a time $t > t_0$. The manipulations entailed by experimental evidence are deemed superior to observational inferences simply because they greatly reduce the likelihood of false positives and false negatives (it would be very unlikely to observe event $y$ occur after manipulation of event $x$ by chance, especially if the experiment is repeated), thereby yielding more reliable and accurate quantification of the pattern at hand.

Therefore, whilst distinguishing causal patterns from mere correlations is a complex, fascinating and still unsolved problem, operating such distinction is not strictly necessary for the purpose of quantifying knowledge. At the root of any definition and any approach to causation lies a the same, intuitive principle: a causal relation is a relatively robust and universal pattern which makes a particular state of an explanandum more probable in response to an earlier state of an explanans. The state

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v1 | CC-BY 4.0 Open Access | rec: 19 Apr 2016, publ: 19 Apr 2016

of the explanans could be imposed or could simply be known. We can translate the former condition using a "do" notation [39], i.e. $X|do(X=x)$ and we translate the latter simply as $X=x$. In either case, the uncertainty of the explanans disappears, i.e. $H(X|do(X=x))=0$ and $H(X=x)=0$. Therefore, in terms of the $K$ function, causal knowledge is quantified as

$$\begin{aligned}K(Y_t;X|do(X_{t_0}=x)) &\equiv \frac{H(Y_t)-H(Y_t|X_{t_0}=x)}{H(Y_t)+H(X_{t_0}|do(X_{t_0}=x)} = \\ &= \frac{H(Y_t)-H(Y_t|X_{t_0}=x)}{H(Y_t)+H(X_{t_0}=x)} = \\ &= \frac{H(Y_t)-H(Y_t|X_{t_0}=x)}{H(Y_t)} \equiv k(Y_t;X_{t_0}) \quad \text{(Causal K)}\end{aligned}$$

Note that $K(Y_t;X|do(X_{t_0}=x)) > k(Y_t;X_{t_0})$ for all systems, which is consistent with the fact that causal knowledge is generally valued more highly than non-causal or correlational knowledge.

**Theoretical Knowledge** Theories, hypotheses, models play, in our functions, exactly the same role as empirical events or objects, and are quantified in the same way (further details are given in section 0.11). When taking the role of explanans, in particular, theoretical constructs are just devices that reduces the uncertainty of an explanandum. To illustrate the role of theories and models in acting as explanantia, let explanandum $Y$ be the joint distribution of two variables, i.e. $\underline{Y}:\{Y_1,Y_2\}$. Since $H(\underline{Y})=H(Y_1)+H(Y_2|Y_1)$, this explanandum is a system in itself, the knowledge of which depends on the size of conditional term $H(Y_1|Y_2)$. Let $\underline{X}\equiv T:\{\tau_1,\tau_2,...\}$ be a set of alternative models or theories that determine the value of $H(Y_1|Y_2)$. Each $\tau_i$ has an associated entropy $H(\tau_i)$ which might express either the theory's plausibility relative to competing theories or its complexity measured in a relative or absolute sense (see sections 0.3 and 0.11). In either case, we get:

$$K(\underline{Y};\tau_i)=\frac{H(Y_1)+H(Y_2|Y_1)-(H(Y_1|\tau_i)+H(Y_2|Y_1,\tau_i))}{H(Y_1)+H(Y_2|Y_1)+H(\tau_i)} \tag{15}$$

Since $\tau_i$ is only relevant to the relation between $Y_1$ and $Y_2$ then we can assume $H(Y_1|\tau_i)=H(Y_1)$ for every $\tau_i$, and the equation simplifies to:

$$K(\underline{Y};\tau_i)=\frac{H(Y_2|Y_1)-H(Y_2|Y_1,\tau_i))}{H(Y_1)+H(Y_2|Y_1)+H(\tau_i)} \quad \text{(Theoretical K)}$$

which expresses the explanatory power of a single $\tau_i$. The average over all alternative theories is given by $K(\underline{Y};T)$. The similarity of this function with the standard $K$ function illustrates how a theory (model, etc...) $T$ is just a device that modulates the relative uncertainty of elements within a system.

## 0.6 Operations on Information

We now define two simple operations that, combined with their respective inverses, capture the essence of information processing. We will always refer to classic random variables, with corresponding Probability Mass Function (PMF) and Probability Mass Vector (PMV). A random variable $X$ with alphabet $|\chi|=\{x_1,x_2...x_n\}$ will have:

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v1 | CC-BY 4.0 Open Access | rec: 19 Apr 2016, publ: 19 Apr 2016

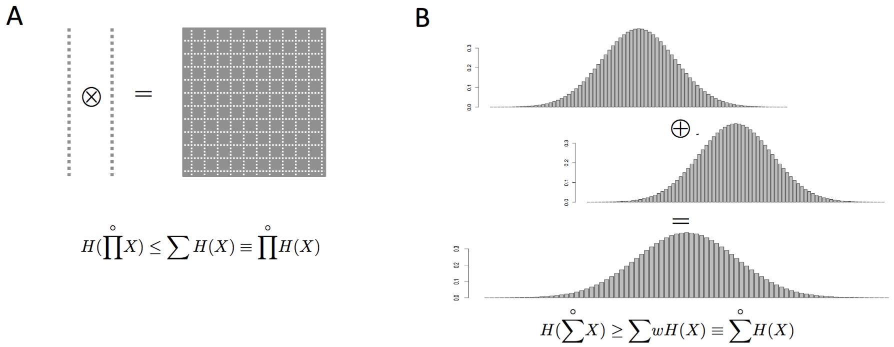
A
⊗
=
$H(\overset{\circ}{\prod} X) \leq \sum H(X) \equiv \overset{\circ}{\prod} H(X)$
B
0.0
0.1
0.2
0.3
⊕
=
$H(\overset{\circ}{\sum} X) \geq \sum wH(X) \equiv \overset{\circ}{\sum} H(X)$

**Figure 4.** Visual representation of the two fundamental operations defined in the text. A) The ⊗ operation joins attributes along separate dimensions. Here attributes are symbolized by dotted lines which could represent quantized random variable with a uniform distribution. B) The ⊕ operation joins attributes along the same dimension. Here attributes are represented by quantized random variables with a normal distribution. See text for further details.

$$F_X = \begin{cases} P_X(x_1) \text{ for } X = x_1 \\ P_X(x_2) \text{ for } X = x_2 \\ \quad ... \\ \quad ... \\ P_X(x_n) \text{ for } X = x_n \\ \quad 0 \quad \text{otherwise} \end{cases} \qquad \text{(Probability Mass Function)}$$

And will have a Probability Mass Vector:

$$\overrightarrow{F}_X = \begin{pmatrix} P_X(x_1) \\ P_X(x_2) \\ ... \\ ... \\ P_X(x_n) \end{pmatrix} \qquad \text{(Probability Mass Vector)}$$

⊗**: Expanding** This operation expands the number of attributes that are processed about a phenomenon or event, thereby augmenting the information about it. Let $X$ and $Z$ be two random variables with alphabets $\chi = \{x_1, x_2, ..., x_n\}$ and $\zeta = \{z_1, z_2, ..., z_m\}$, and PMFs $F_X$ and $F_Z$. The o-product $X \otimes Z$ is a new random variable, the alphabet of which is the outer product of the two alphabets and the PMV is the corresponding joint probability:

$$X \otimes Z \to \{X, Z\} \quad :$$

$$\overrightarrow{F}_X \times \overrightarrow{F}_{Z|X} = \begin{pmatrix} P_X(x_1) * P_{Z|X}(z_1|x_1) \\ P_X(x_1) * P_{Z|X}(z_2|x_1) \\ ... \\ P_X(x_2) * P_{Z|X}(z_1|x_2) \\ ... \\ P_X(x_n) * P_{Z|X}(z_m|x_n) \end{pmatrix} \equiv \begin{pmatrix} P(x_1, z_1) \\ P(x_1, z_2)) \\ ... \\ P(x_2, z_1) \\ ... \\ P(x_n, z_m) \end{pmatrix} \qquad (16)$$

The joining of two random variables can be effectively visualized as yielding a matrix (Fig 4), which is an intuitive and useful representation. However, a joint distribution is

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v1 | CC-BY 4.0 Open Access | rec: 19 Apr 2016, publ: 19 Apr 2016

for all means and purposes just a new variable, and can be equivalently thought of as a vector (e.g. a column vector). The extension of the $\otimes$ operation to multiple dimensions is straightforward:

$$\overset{\circ}{\prod_{i \in n}} X_i = X_1 \otimes X_2 \otimes ... \otimes X_n :$$
$$\overrightarrow{F}_{\underline{X}} = \overrightarrow{F}_{X_1} \otimes \overrightarrow{F}_{X_2|X_1} \otimes \overrightarrow{F}_{X_3|X_2X_1} \otimes ... \otimes \overrightarrow{F}_{X_n|X_{n-1}...X_1} \qquad (17)$$

The conditional probability distributions underlying the $\otimes$ operation may be calculated theoretically (as is done, for example, in quantum mechanics) but is more commonly derived empirically. An important condition correspond to the case in which these values are unknown and/or the variables are assumed to be independent i.e. $F_{Z|X} = F_Z$, which simplifies the $\otimes$ operation to:

$$\overrightarrow{F}_X \times \overrightarrow{F}_Z = \begin{pmatrix} P_X(x_1) * P_Z(z_1) \\ P_X(x_1) * P_Z(z_2) \\ ... \\ P_X(x_2) * P_Z(z_1) \\ ... \\ P_X(x_n) * P_Z(z_m) \end{pmatrix} \qquad (18)$$

When, as in this case, independent of the random variables is assumed, the $\otimes$ operation is effectively equivalent to an outer product (Kronecker product) of PMVs, and the notation used is intended to recall this fact.

**$\oslash$: Reducing** The $\oslash$ operation reverses $\otimes$. Let 1 and 2 be two different attributes, and let them be represented by the random variables $X_1$ and $X_2$ with alphabets $\chi_1 = \{1, 2, ..., n\}$ and $\chi_2 = \{1, 2, ..., m\}$ and PMFs $F_{X_1}$ and $F_{X_2}$, and let $\underline{X} = X_1 \otimes X_2$. Then we define $\oslash$ as:

$$\underline{X} \oslash X_2 \rightarrow \{X_1\} : \overrightarrow{F}_{\underline{X} \oslash X_2} = \overrightarrow{F}_{\underline{X}} \times \frac{1}{\overrightarrow{F}_{X_2|X_1}} =$$
$$= \begin{pmatrix} \underline{P}(x_1x_1) * \frac{1}{P_{2|1}(x_1|x_1)} + \underline{P}(x_1x_2) * \frac{1}{P_{2|1}(x_2|x_1)} + ... + \underline{P}(x_1x_m) * \frac{1}{P_{2|1}(x_m|x_1)} \\ \underline{P}(x_2x_1) * \frac{1}{P_{2|1}(x_1|x_2)} + \underline{P}(x_2x_2) * \frac{1}{P_{2|1}(x_2|x_2)} + ... + \underline{P}(x_2x_m) * \frac{1}{P_{2|1}(x_m|x_2)} \\ ... \\ \underline{P}(x_nx_1) * \frac{1}{P_{2|1}(x_1|x_n)} + \underline{P}(x_nx_2) * \frac{1}{P_{2|1}(x_2|x_n)} + ... + \underline{P}(x_nx_m) * \frac{1}{P_{2|1}(x_m|x_n)} \end{pmatrix} \qquad (19)$$

The $\oslash$ operation re-extracts marginal probabilities. When applied to the joint distribution of two variables, as above, it re-creates one of them, e.g. $X_1$. In the general case of a multidimensional random vector $\underline{X} = \{X_1, X_2, ...X_n\}$, the operation extracts a subset of dimensions, which we will often indicate as a "complementary" random vector $\underline{X}^c = \underline{X} \ominus X_n$.

**$\oplus$: Cumulating** This operation cumulates attributes that are processed about a phenomenon or event, thus generating and updating the total information contained in it. Let $X_1$ and $X_2$ be two random variables with alphabets $\chi_1 = \{x_1, x_2, ..., x_n\}$ and $\chi_2 = \{x_1, x_2, ..., x_m\}$ and PMFs $F_{X_1}$ and $F_{X_2}$, the operation $X_1 \oplus X_2$ yields a new random variable whose alphabet is the union set of the alphabets after appropriate re-scaling, i.e. $\chi_1 \cup \chi_2 = \{x'_1, x'_2, ..., x_u\}$ and whose PMV is the weighted sum of the corresponding re-scaled PMVs $\overrightarrow{F}_{X_1}$ and $\overrightarrow{F}_{X_2}$:

$X_1 \oplus X_2$ :

$$w_1 \overrightarrow{F}'_{X_1} + w_2 \overrightarrow{F}'_{X_2} = \begin{pmatrix} w_1 P_{x_1}(x'_1) + w_2 P_2(x'_1) \\ w_1 P_{x_1}(x'_2) + w_2 P_{x_2}(x'_2) \\ ... \\ w_1 P_{x_1}(x_u) + w_2 P_{x_2}(x_u) \end{pmatrix} = \begin{pmatrix} P_{\tilde{X}}(x'_1) \\ P_{\tilde{X}}(x'_2) \\ ... \\ ... \\ P_{\tilde{X}}(x_u) \end{pmatrix} \tag{20}$$

in which $u$ is the alphabet size of the union set, $|\underline{\chi}| = |\chi_1 \cup \chi_2|$ and $w_1$ and $w_2$ correspond to the weights:

$$w_1 = \frac{w_1^*}{w_1^* + w_2^*} \text{ and } w_2 = \frac{w_2^*}{w_1^* + w_2^*}, w_1 + w_2 = 1 \tag{21}$$

It is straightforward to generalize the example above to the case of $k$ random variables:

$$\overset{\circ}{\sum_{i\in k}} X_i = X_1 \oplus X_2 \oplus ... \oplus X_k = \sum_{i\in k} w_i \overrightarrow{F}_{X_i} \tag{22}$$

in which $w_i$ is an element of a vector weights $\underline{w} = \{w_1, w_2, ..., w_k\}$ in turn derived from a vector $\underline{w}* = \{w*_1, w*_2, ..., w*_k\}$ through the same calculations indicated above.

Therefore, the $\oplus$ operation always subtends vectors of weights that are pre-determined empirically or theoretically. The simplest and most natural condition, however, is one in which $w* = 1$, and the cumulated weights $w$ directly reflect the frequencies experienced.

The o-plus operation applies also in the case of individual events and absence of information. Let $x_1$ and $x_2$ be experienced attributes of two separate objects or event, with $x_1 \neq x_2$ . $x_1$ and $x_2$ can be represented as random variables, $X_1$ and $X_2$ with PMFs:

$$F_{X_1} = \begin{cases} 1 \text{ for } X_1 = x_1 \\ 0 \quad \text{otherwise} \end{cases} \quad F_{X_2} = \begin{cases} 1 \text{ for } X_2 = x_2 \\ 0 \quad \text{otherwise} \end{cases} \tag{23}$$

The PMF of their o-sum will be:

$$F_{X_1 \oplus X_2} = \begin{cases} w_1 \times 1 + w_2 \times 0 \text{ for } \tilde{X} = x_1 \\ w_1 \times 0 + w_2 \times 1 \text{ for } \tilde{X} = x_2 \\ \quad 0 \quad \text{otherwise} \end{cases} = \begin{cases} w_1 \text{ for } \tilde{X} = x_1 \\ w_2 \text{ for } \tilde{X} = x_2 \\ 0 \quad \text{otherwise} \end{cases} \tag{24}$$

In which, as before, $w_1 = \frac{w_1^*}{\sum_{i\in k} w_i}$ and $w_2 = \frac{w_2^*}{\sum_{i\in k} w_i}$. As argued before, whilst the weights can in principle have any value, the default assumption is $w* = 1$, which is a case particularly fitting when information about attributes is acquired at the most elementary level. With such weights, the resulting PMV is:

$$\overrightarrow{F}_{X_1 \oplus X_2} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} \tag{25}$$

Now, to illustrate how frequency distributions of an attribute are built by the cumulation of experiences, let's o-sum to the previous two events a third event with attribute equivalent to one of the previous two, i.e $x'_2 = x_2$. As before, we can represent $x'_2$ as a single-valued random variable $X_3$:

$$F_{X_3} = \begin{cases} 1 \text{ for } X_3 = x_2 \\ 0 \quad \text{otherwise} \end{cases} \tag{26}$$

Using equal weights, the operation $\overset{\circ}{\sum} X \equiv X_1 \oplus X_2 \oplus X_3$ will yield a new PMF:

$$\overrightarrow{F}_{\overset{\circ}{\sum} X} = \begin{cases} w_1 \times 1 + w_2 \times 0 + w_3 \times 0 \\ w_1 \times 0 + w_2 \times 1 + w_3 \times 1 \\ 0 \quad \text{otherwise} \end{cases} = \begin{cases} \frac{1}{3} \text{ for } \tilde{X} = x_1 \\ \frac{2}{3} \text{ for } \tilde{X} = x_2 \\ 0 \quad \text{otherwise} \end{cases} \tag{27}$$

And this of course produces the corresponding PMV:

$$\overrightarrow{F}_{\overset{\circ}{\sum} X} = \begin{pmatrix} w_1 \\ w_2 + w_3 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \end{pmatrix} \equiv \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} \tag{28}$$

In which $\nu$ is used to indicate a frequency-derived probability. A mathematically equivalent operation would consist in o-summing not each variable anew, as we did above, but to o-sum the cumulative variable $\tilde{X}$ to the new variable $X_3$, using appropriate weights. Let $\tilde{X} \equiv X_1 \oplus X_2$ and let $\underline{w}* = \{1, 2\}$ such that
$w_1 = 2/3, w_2 = 1/3$, then

$$\overrightarrow{F}_{\tilde{X} \oplus X_3} \equiv w_1 \overrightarrow{F}_{\tilde{X}} + w_2 \overrightarrow{F}_{X_3} = \frac{2}{3} \begin{pmatrix} \frac{1}{2} \\ \frac{1}{2} \end{pmatrix} + \frac{1}{3} \begin{pmatrix} 0 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{1}{3} \\ \frac{2}{3} \end{pmatrix} \equiv \begin{pmatrix} \nu_1 \\ \nu_2 \end{pmatrix} \tag{29}$$

**$\ominus$: Removing** This operation reverses of the $\oplus$ operation. Let $\tilde{X} = X_1 \oplus X_2$ be a cumulated random variable of alphabet $\underline{\chi} = \{x_1, x_2, ..., x_u\}$ and PMV
$\overrightarrow{F}_{\tilde{X}} \equiv w_1 \overrightarrow{F}_{X_1} + w_2 \overrightarrow{F}_{X_2}$. The $\ominus$ operation is defined as:

$$\tilde{X} \ominus X_2 = X_1 : \quad \overrightarrow{F}_{\tilde{X} \ominus X_2} = \frac{F_{\tilde{X}} - w_2 F_{X_2}}{1 - w_2} = \begin{pmatrix} \frac{P_{\tilde{X}}(1) - w_2 P_2(1)}{1 - w_2} \\ \frac{P_{\tilde{X}}(2) - w_2 P_2(2)}{1 - w_2} \\ ... \\ \frac{P_{\tilde{X}}(u) - w_2 P_2(u)}{1 - w_2} \end{pmatrix} \tag{30}$$

It is theoretically possible to extend this operation to cases in which none of the variables involved is a cumulated variable, but this would require setting restrictions on the applicable PMVs or alternatively allow probabilities to be negative, a concept of unclear meaning. The possibility of extending this operation should be explored in future work.

## 0.7 Shannon Entropy with $\otimes$ and $\oplus$

In this section we examine how the two operations defined above affect the entropy and K functions.

**Lemma 0.1.** *Let $1, 2, ...n$ be attributes, let $x_1, x_2...x_n$ be corresponding events (i.e. instantiations of the attributes), and let $X_1, X_2, ..., X_n$ be corresponding random variables, i.e. random variables whose probability distributions $P_1, P_2...P_n$ are the result of cumulations of instantiations of each attribute.*

$$H(\overset{\circ}{\prod}_n x) = log \frac{1}{P_{X_1}(x_1)} + log \frac{1}{P_{X_2|X_1}(x_2|x_1)} + ... + log \frac{1}{P_{X_n|X_{n-1}X_{n-2}...X_1}(x_n|x_{n-1}x_{n-2}...x_1)}$$

$$H(\overset{\circ}{\prod}_n X) = H(X_1) + H(X_2|X_1) + ... + H(X_n|X_{n-1}X_{n-2}...X_1)... = \sum_n H(X|X_{n-1}, X_{n-2}, ...X_1)$$

$$H(\overset{\circ}{\prod}_n X \oslash X_m) = \sum_{n \neq m} H(X_n|X_{n-1}, X_{n-2}, ...X_1)$$

*Proof.* Follows directly from our definition of the operation $\otimes$ and $\oslash$ and the chain rule of entropy, according to which if $X$ and $Y$ are two random variables and $XY$ is their joint distribution, then $H(XY) = H(X) + H(Y|X)$. □

As mentioned previously, a case of particular relevance occurs when the attributes are assumed to be independent (Fig 5). To indicate this condition we will use a special notation, in which the operation sign is exterior to the entropy symbol:

$$H(\overset{\circ}{\prod}_n X) \equiv \overset{\circ}{\prod}_n H(X) \iff \sum_n H(X|X_{n-1}, X_{n-2}, ...X_1) \equiv \sum_n H(X) \tag{31}$$

**Lemma 0.2.** *Let $X$ be an attribute and let $x_1, x_2...x_m$ be events, i.e. instantiations of the attribute $X$, then*

$$H(\overset{\circ}{\sum}_m x) \equiv H(\overrightarrow{F}_X) \equiv H(X)$$

*in which $\overrightarrow{F}_X$ is the PMV obtained by cumulation. Let $X_1, X_2...X_k$ be separate cumulations of events of the same attribute then*

$$H(\overset{\circ}{\sum}_k X) = \sum_{j\leq k} w_j H(X_j) + \sum_{j\leq k} w_j D(X_j||\tilde{X}) \equiv \overset{\circ}{\sum} H(X) + \overset{\circ}{\sum} D(X||\tilde{X}) \tag{32}$$

*In which $\tilde{X} \equiv \overset{\circ}{\sum} X$, $w_j$ are the weights underlying the $\oplus$ operation and $D$ is Kullback-Leibler's distance.*

*Proof.* The first claim follows directly from the definition of entropy and of the $\oplus$ operation, which from a cumulation of individual events gives rise to a random variable with PMV $\overrightarrow{F}_X$. The second claim follows similarly:

$$H(\overset{\circ}{\sum} X) = -\sum_x \sum_{j\leq k} w_j P_j(x) log(\sum_{j\leq k} w_j P_j(x)) = -\sum_{j\leq k} w_j \sum_x P_j(x) log(\sum_{j\leq k} w_j P_j(x)) =$$

$$= -\sum_{j\leq k} w_j \sum_x P_j(x) log(w_j P_j(x) \frac{\sum_{j\leq k} w_j P_j(x)}{w_j P_j(x)}) = -\sum_{j\leq k} w_j \sum_x P_j(x)[log(w_j) + log(P_j(x)) -$$

$$-log(\frac{w_j P_j(x)}{\sum_{j\leq k} w_j P_j(x)})] = -\sum_{j\leq k} w_j \sum_x P_j(x) log(P_j(x)) + \sum_{j\leq k} w_j \sum_x P_j(x) log(\frac{P_j(x)}{\sum_{j\leq k} w_j P_j(x)}) \equiv$$

$$\equiv \sum_{j\leq k} w_j H(X_j) + \sum_{j\leq k} w_j D(X_j||\tilde{X}) \tag{33}$$

□

**Lemma 0.3.** *Let $X$ be a random variable with alphabet $\chi : \{x_1, x_2...x_m\}$ and PMV $\overrightarrow{F}_X$ and let $x_i$ be a new event which is cumulated to $X$ with a weight $w_i$.*

$$\text{If } x_i \notin \chi, \text{ then} \quad H(X \oplus x_i) = C \times H(X) + H(\underline{C}) \tag{34}$$

*in which $C = 1 - w_i$ and $\underline{C} : \{C, 1 - C\}$ is the PMV of binary probability $C$.*

$$\text{If } x_i = x_j, x_j \in \chi, \text{ then} \quad H(X \oplus x_i) = CH(X) + H(\underline{C}) - \delta(i,j)_X \tag{35}$$

*in which* $\delta(i,j)_X \equiv -CP(i=j)log\frac{CP(i=j)}{CP(i=j)+(1-C)} - (1-C)log\frac{(1-C)}{CP(i=j)+(1-C)}$ *is the information distance in* $X$ *between the two events* $i, j$ *given the probability distribution* $P(x)$ *of* $X$.

*Let* $X_1, X_2, ..., X_k, X_{k+1}$ *be random variables, then*

$$H(\overset{\circ}{\sum}_{k+1}X) = \\ C \times (H(\overset{\circ}{\sum}_{k}X) + D(\overset{\circ}{\sum}_{k}X||\overset{\circ}{\sum}_{k+1}X)) + (1-C)(H(X_{k+1}) + D(X_{k+1}||\overset{\circ}{\sum}_{k+1}X)) \tag{36}$$

*Proof.* We prove the latter statement and derive the former two as special cases.

$$H(\overset{\circ}{\sum}_{k+1}X) = -\sum_x \sum_{i\leq k+1} w_iP_i(x)log\sum_{i\leq k+1} w_iP_i(x) = \\ = -\sum_x \sum_{i\leq k+1} w_iP_i(x)log\sum_{i\leq k+1} w_iP_i(x)\frac{\sum_{i\leq k} z_iP_i(x)}{\sum_{i\leq k} z_iP_i(x)} \tag{37}$$

in which $z_i$ are the weights of the previous cumulation, i.e. that of the first $k$ elements. Since $w_i \equiv \frac{w_i^*}{\sum_{i\leq k+1} w_i^*}$ and $z_i \equiv \frac{w_i^*}{\sum_{i\leq k} w_i^*}$, then $w_i = C \times z_i$, therefore the above becomes:

$$-\sum_x \sum_{i\leq k+1} w_iP_i(x)log\sum_{i\leq k} z_iP_i(x) - \sum_y \sum_{i\leq k+1} w_iP_i(x)\frac{\sum_{i\leq k+1} w_iP_i(x)}{\sum_{i\leq k} z_iP_i(x)} = \\ = -\sum_x \sum_{i\leq k} Cz_iP_i(x)log\sum_x \sum_{i\leq k} z_iP_i(x) - \sum_x w_{k+1}P_{k+1}(x)log\sum_{i\leq k} z_iP_i(x) + \\ + \sum_x \sum_{i\leq k} Cz_iP_i(x)log\frac{\sum_{i\leq k} z_iP_i(x)}{\sum_{i\leq k+1} w_iP_i(x)} - \sum_x w_{k+1}P_{k+1}(x)log\frac{\sum_{i\leq k+1} w_iP_i(x)}{\sum_{i\leq k} z_iP_i(x)} \tag{38}$$

which simplifies to the above, with $w_{k+1} = 1 - C$. When the cumulation regards individual events $x$, then the probability of $x_i$ is one, which removes the second entropy term and simplifies to the equation above to $C(H(\underline{z}) - logC) - (1-C)log(1-C)$ or equivalently $CH(X_m) + H(\underline{C})$, plus eventual distance terms when
$x_{m+1} = x_j \in |\chi_m|$. □

Lemma 0.3 shows that all terms that add or subtract entropy are directly dependent on $C$ and therefore on the weighting underlying the operation. If weights are just an inverse function of the number of elements added, then the entropy of a cumulated random variable will always converge to a stable value.

**Lemma 0.4** (Cumulation of completely overlapping variables)**.** *Let* $X_1, X_2, ..., X_k$ *be random variables.*

$$H(\overset{\circ}{\sum}_{j\leq k}X_j) = \sum_{j\leq k} w_jH(X_j) \iff X_m = X_n \forall m \leq k, n \leq k \tag{39}$$

*Proof.* Follows from (19), since $w_1 \sum_x P_1(x)log(\frac{P_1(x)}{\sum_j w_jP_j(x)}) = 0$ and
$w_2 \sum_x P_2(x)log(\frac{P_2(x)}{\sum_j w_jP_j(x)}) = 0$ implies that $P_1(x) = \sum_j w_jP_j(x)$ and
$P_2(x) = \sum_j w_jP_j(x)$ and therefore $P_1(x) = P_2(x) \forall x \in \tilde{\chi}$. □

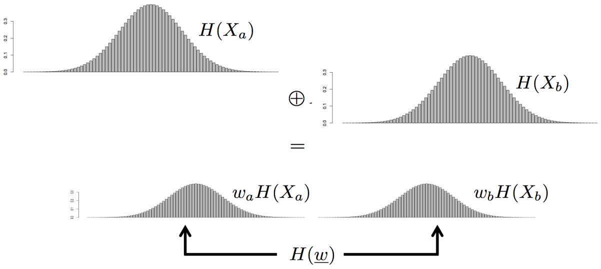

**Figure 5.** Visual representation of the ⊕ operation between random variables whose alphabets are completely non-overlapping (see Figure 4B for comparison and text for further details).

**Lemma 0.5** (Cumulation of completely non-overlapping variables)**.** *Let $X_1, X_2, ..., X_k$ be random variables, let $\underline{w}$ be the vector of weights underlying the operation $\overset{\circ}{\sum}_{j\leq k} X_j$, such that $\sum_{w\in\underline{w}} w = 1$.*

$$H(\overset{\circ}{\sum}_{j\leq k}X_j) = \overset{\circ}{\sum}H(X) + H(\underline{w}) \iff \chi_i \cup \chi_j = \emptyset \forall i \leq k, j \leq k, i \neq j \tag{40}$$

*Proof.* When $X_1, X_2$ do not overlap, then

$$\begin{aligned} H(\overset{\circ}{\sum}_{j\leq k}X_j) &= -\sum_{j\leq k} w_j \sum_x P_j(x) log(w_j P_j \frac{\sum_{j\leq k} w_j P_j(x)}{w_j P_j(x)}) = \\ &= -\sum_{j\leq k} w_j \sum_x P_j(x) log(w_j P_j(x)) = -\sum_{j\leq k} w_j \sum_x P_j(x) log(P_j(x)) - \sum_{j\leq k} w_j log(w_j) \end{aligned} \tag{41}$$

□

The lemma yields an important and intuitive result. When completely independent objects or events are joined in one, each of them retains its original amount of information, but this gets reduced by a constant because each object now contains information about the other (if the instantiation of one of them occurs, then no instantiation of the other can occur). Therefore, the information in the new object is a weighted average of the information of its components, plus the information needed to distinguish them. The latter information is contained in the entropy of the weights. With equal weights and just two distributions, for example, each entropy term is halved and augmented by one bit (Fig 5).

**Lemma 0.6.** *Let $X_1, X_2, ..., X_k$ be random variables, with $H(\overset{\circ}{\sum}_k X) = \overset{\circ}{\sum}_k H(X) + \overset{\circ}{\sum}_k D(X||\tilde{X})$, then*

$$0 \leq \overset{\circ}{\sum}_k D(X_j || \overset{\circ}{\sum}_{j\leq k} X_j) \leq H(\underline{w}) \leq log(k) \tag{42}$$

*Proof.* Follows from all preceding lemmas. □

The analogy between conclusion of lemma 0.5 and the definition of Total Information Content given previously should be clear, and can be made explicit:

**Lemma 0.7.** *Let $x$ be an event or object with universal probability $P_U(x)$. Then, with $n = 1$,*

$$T(x) = H(x) + D(x||P_U(x))$$

$$T(\overset{\circ}{\sum} x) = H(X) + D(X||U)$$

$$T(\overset{\circ}{\sum} X) = \overset{\circ}{\sum} H(X) + \overset{\circ}{\sum} D(X||\tilde{X}) + \overset{\circ}{\sum} D(X||U)$$

*Proof.*

$$\begin{aligned} T(\overset{\circ}{\sum} x) &= -\sum_{x}\sum_{j\leq k} w_j P_j(x) log \sum_{j\leq k} w_j P_{Uj}(x) = -\sum_{x}\sum_{j\leq k} w_j P_j(x) log \sum_{j\leq k} w_j \frac{P_j(x)P_{Uj}(x)}{P_j(x)} = \\ &= -\sum_{x}\sum_{j\leq k} w_j P_j(x) log \sum_{j\leq k} w_j P_j(x) + \sum_{x}\sum_{j\leq k} w_j P_j(x) log \frac{P_j(x)}{P_{Uj}(x)} \equiv \\ &\equiv H(\overset{\circ}{\sum}_{j\leq k} X_j) + \overset{\circ}{\sum} D(X||U) \end{aligned} \quad (43)$$

The other cases follow similarly. □

# Results

We will first show that the $K$ function expresses three properties that knowledge is intuitively expected to have.

## 0.8 Properties of K

**Property 1: Occam's razor** The principle popularly known as "Occam's razor" states that when two equally plausible explanations are available for the same phenomenon, the simpler explanation should be preferred. Although it makes intuitive sense, this principle has proven difficult to justify on philosophical and logical grounds. Several aesthetic, probabilistic and pragmatic arguments have been proposed, but Occam's razor remains a notion that, rather than being demonstrated, is postulated as a desideratum of knowledge. Intriguingly, the present theory does not need to postulate this principle nor invoke additional arguments to justify it. Occam's razor is an intrinsic property of the K function.

Translated into our notation, Occam's razor posits the existence of a single explandum $y$ and two candidate explanantia $a$ and $b$:

$$K(y;a) = \frac{H(y) - H(y|a)}{H(y) + H(a)} \quad vs. \quad K(y;b) = \frac{H(y) - H(y|b)}{H(y) + H(b)} \quad (44)$$

Both explanations are assumed to be equally effective, i.e. $H(y) - H(y|a) = H(y) - H(y|b)$, but one of the explanation is "simpler" than the other. As discussed in section 0.3, simplicity can be expressed in terms of Kolmogorov complexity (e.g. the length of the minimum description of the explanation), Shannon entropy (i.e. the frequency-derived probability or the subjective weight or plausibility attributed to each explanation) or a combination of both (i.e. the Total information). Using Shannon's entropy notation:

$$H(a) < H(b) \quad (45)$$

which leads to

$$K(y;a) > K(y;b) \tag{46}$$

The simpler explanation yields a higher K, which justifies Occam's choice.

The remarkable aspect of this result is that the $K$ function was not constructed with the explicit purpose of accommodating Occam's razor, but was derived from the mutual information function following a postulated equivalence of knowledge with pattern-detection. The finding that Occam's razor is intrinsic to K is a striking support for the notion that knowledge is information compression and that simplicity and elegance are not an arbitrary aesthetic values that people (including scientists) choose to impose on knowledge, as scholars have argued [33]. To the extent that it underlies the encoding of patterns, simplicity *is* knowledge.

**Property 2: Optimal accuracy** Another property of K that, like Occam's razor, is commonly and intuitively associated with knowledge is its dependence on optimal accuracy. When accuracy is suboptimal, information is lost; when accuracy is excessive, information becomes redundant, wasting resources and reducing knowledge. Again, with no need to postulate resource costs, we find this property to be intrinsic to our definition of knowledge as expressed in the K function.

**Definition: accuracy** Let $X$ be a quantized attribute, and let $a \in \mathbb{N}$ be the number $\alpha_x$ of partitions of the attribute, or in other words the size of the alphabet of $X$. We define as the accuracy of measurement of $X$ the size of the partition $\alpha_x = \frac{1}{a_x}$. The corresponding random variable, which associates probabilities to each partition of $X$, will represent a quantization of X indicated as $X^{\alpha_x}$ or, when the value is clear from the context,$X^{\alpha}$ . An increase of measurement accuracy can be represented as a progressive sub-partitioning of the attribute being measured (Fig 1).

**Lemma 0.8.** *Let $X^{\alpha}$ be a quantized variable of accuracy a, let $n \in \mathbb{N}$ with $n \geq 2$, and let $\alpha' = \alpha/n$ represent a higher accuracy. Then:*

$$0 < H(X^{\alpha'}) - H(X^{\alpha}) \leq log(n) \tag{47}$$

*Proof.* The proof follows from the effects that the partitioning has on the entropy. If $H(X^{\alpha}) = -\sum_1^a p(x)log(p(x))$, with $x$ corresponding to the value of attribute in each of the $a$ partitions, then $H(X^{\alpha'}) = -\sum_1^{a\times n} p(x')log(p(x')) = -\sum_1^a \sum_1^n p(x')log(p(x'))$. Known properties of entropy tells us that the entropy of the n-partition of $\alpha$ is smaller or equal to the logarithm of the number $n$ of partitions with equality if and only if the n-partitions of $\alpha$ have all the same probability, i.e. $P(x') = \frac{1}{n}\forall x'$. □

**Definition: measurement error** Let $X^{\alpha}$ be a quantized random variable with accuracy $\alpha$, and let $\alpha' = \alpha/n$ represent a higher accuracy. The measurement error of $X^{\alpha}$ is a quantity $e \in \mathbb{Q}$ such that:

$$H(X^{\alpha'}) - H(X^{\alpha}) = log(n), \forall \alpha' \leq e \tag{48}$$

**Definition: empirical system** A system $YX$ is said to be empirical if its quantization has measurement error. Equivalently, a non-empirical, (i.e. logico-deductive) system is a system YX for which measurement error $e = 0$.

The effect that a change in accuracy has on K depends on the underlying properties of the system, and in particular on the speed with which the entropy of the explanandum and/or explanans increase relative to their joint distribution.

**Theorem 0.9** (Knowledge of empirical systems has an optimal accuracy)**.** *For every system $Y^{\alpha_y}, X^{\alpha_x}$ that is empirical (has non-zero measurement error), there are optimal values of accuracy $\alpha*_y$ and $\alpha*_x$ such that:*

$$K(Y^{\alpha*_y}, X^{\alpha*_x}) > K(Y^{\alpha_y}, X^{\alpha_x}) \forall \alpha_y \neq \alpha*_y, \alpha_x \neq \alpha*_x \tag{49}$$

*Proof.* Limiting for simplicity the case to the accuracy of $Y$ and indicating, as before, with $\alpha' = \alpha/n$, we have:

$$K(Y^{\alpha'}; X) > K(Y^{\alpha}; X) \iff \frac{H(Y^{\alpha'}X)}{H(Y^{\alpha'}) + H(X)} < \frac{H(Y^{\alpha}X)}{H(Y^{\alpha}) + H(X)} \tag{50}$$

From lemma 0.8 we know that $H(Y^{\alpha'}) - H(Y^{\alpha}) \leq log(n)$, i.e. $H(Y^{\alpha'}) \leq H(Y^{\alpha}) + log(n)$, and re-arranging we get the condition:

$$H(Y^{\alpha'}|X) - H(Y^{\alpha}|X) < (1 - K(Y^{\alpha}; X))log(n) \tag{51}$$

which can only be true if $H(Y^{\alpha'}|X) - H(Y^{\alpha}|X) << log(n)$ for any alphabet size of $Y$. In other words, $K$ will only grow at every new sub-partitioning of $Y$ if the resulting sub-partitions are never equally probable when conditioned on the explanans. Therefore, inequality 51 is only satisfied for systems that are measurable to infinite accuracy. The opposite condition, $K(Y^{\alpha'}; X) > K(Y^{\alpha}; X)$ is not necessarily true. In particular, it is always true when $K(Y^{\alpha}; X)$ has reached its maximum, i.e. when $H(Y^{\alpha}X) = H(X)$, but not necessarily otherwise. □

Theorem 0.9 guarantees that, for any system composed of two quantized random variables $Y^{\alpha_y}, X^{\alpha_x}$, if accuracy is increased arbitrarily then $K(Y^{\alpha_y}, X^{\alpha_x})$ will reach a maximum and subsequently decline, unless $\lim_{\alpha \to 0} H(Y^{\alpha_y}|X^{\alpha_x}) = 0$. The latter condition is only and always satisfied when $Y = X$, i.e. when the system is composed of an identity. As will be discussed in section 0.11, identities are the defining property of logico-deductive knowledge (e.g. mathematics), which is a special case that can nonetheless be analysed and quantified with the K function.

**Property 3: Ignorance about the future** This is also an intrinsic property of the K function, derivable mathematically with no need to postulate physical or physiological constraints. As will be shown, this also implies that empirical knowledge is limited by a "chaos horizon".

**Theorem 0.10** (K is a non-increasing function of time)**.** *Let $Y_t, X_{t_0}$ be an explanandum and an explanans, let $H(Y_t|X_{t_0})$ be the pattern linking explanans and explanandum, and let $K(Y_t; X_{t_0})$ be the corresponding knowledge of the system at time $t$. Let let $Y_{t'}$ be the state of the explanandum at a different time $t' = t + \Delta t$.*

$$if \quad 0 < K(Y_t X_{t_0}) < 1 \quad then \quad K(Y_{t'}; X_{t_0}) < K(Y_t; X_{t_0}) \tag{52}$$

*Proof.* Proof follows from the assumption that the explanans $X_{t_0}$ remains unchanged, i.e. that at time $t'$ the knower has no additional information compared to what it had at time $t_0$. Because of this, the values of $H(Y_{t'})$ and $H(Y_{t'}|X_{t_0})$ can only be extrapolations from (i.e. functions of) the the explanandum measured at time t i.e. $H(Y_{t'}) = f(Y_t)$ and $H(Y_{t'}|X_{t_0}) = f(Y_t|X_{t_0})$. Under these conditions, the data processing inequality applies [36], which dictates that $I(X_{t_0}; Y_t) \geq I(X_{t_0}; Y_{t'})$. Note that $\Delta t$ can be positive as well as negative, in other words the inequality is true for $t' > t$ as well as $t' < t$, leading to generalize the theorem to distances forward as well as backwards in time.

The data processing inequality would allow for $I(X_{t_0}; Y_t) = I(X_{t_0}; Y_{t'})$, but this case is excluded under the conditions of imperfect knowledge imposed by the theorem. We see this by examining the derivative of K with respect to $t$ (allowed because $K$ is continuous and differentiable with respect to a continuous variable such as t). To simplify the notation, we will assume $t_0 = t = 0$ so that $\Delta t \equiv t$. The conditions for $K$ to be non-decreasing are:

$$\frac{dK(Y_t X)}{dt} \geq 0 \Leftrightarrow \quad \frac{dH(Y_t)}{dt}(1 - K(Y_t X)) \geq \frac{dH(Y_t|X)}{dt} \tag{53}$$

This condition is never satisfied unless one assumed that the entropy of $Y_t$ grows faster than that of its conditional counterpart. However, this is impossible because $H(Y|X) = H(Y) + H(X|Y) - H(X)$ and $H(X)$ is fixed by assumption, yielding $H'(Y|X) = H'(Y) + H'(X|Y)$. Since $0 \leq 1 - K(YX) \leq 1$, the condition of equality is possible only when either $K(Y_{t_0} X_{t_0}) = 0$, and there is no knowledge in the system to begin with, or when $K(Y_{t_0} X_{t_0}) = 1$, i.e. when knowledge is complete - a condition that is never satisfied in ordinary knowledge. □

Note that the theorem does not preclude the possibility that if the explandum is measured at a different $\delta t$ from $t_0$, the corresponding value of K might be higher. However, it precludes the possibility that the knower's *predictions* about the state of the explanandum could ever yield more knowledge than that obtained by the system at time t, *unless new information is available.*

**Inevitability of chaos** If knowledge is bound to decline over time, then every empirical system must have a temporal horizon beyond which knowledge of the system is zero. In other words, all systems possess a "chaos horizon".

**Lemma 0.11.** *For any system $Y_t X_{t_0}$ with $K(Y_t X_{t_0}) < 1$, given an arbitrary threshold $\epsilon \in [0, 1]$, there is a time $t^\dagger$ such that $K(Y_{t^\dagger}; X_{t_0}) \approx 0$.*

*Proof.* This conclusion follows from theorem 0.10. If for all empirical systems $K$ decreases over time, then for all empirical systems there must be a time $t^\dagger$ at which $H(Y_{t^\dagger}) - H(Y_{t^\dagger}|X_{t_0}) \leq \epsilon$. □

The lemma suggests that all empirical systems have what may be described as a chaos horizon, i.e. a point beyond which knowledge is impossible. A system is said to be chaotic when it is highly sensitive to initial conditions. Since accuracy of measurement of initial states is limited, future states of the system become rapidly unpredictable even when the system is seemingly simple and fully deterministic. Paradigmatic chaotic systems, such as the 3-body problem or the Lorenz equations that simulate the weather, share the characteristics of being strikingly simple and yet are extremely sensitive to initial conditions, which made their instability particularly notable [42, 43].

In standard chaos theory, the rapidity with which a system diverges from the predicted trajectory can be measured by an exponential function in the form:

$$\frac{d_N}{d_0} \approx e^{\lambda N} \tag{54}$$

in which $d_N/d_0$ is the relative offset after N steps, and $\lambda$ is the so called Liapunov exponent, a parameter that quantifies sensitivity of the system to initial conditions. Positive Liapunov exponents correspond to a chaotic system, negative values correspond to stable systems, i.e. systems that are resilient to perturbation.

The exact connection between the empirical concept of chaos in chaos theory and the knowledge horizon predicted by lemma 0.11 remains to be fully explored. However, we can use the concept of Liapunov exponent to model the loss of knowledge over time as an exponential function. Even if we ignore chaos theory, exponential functions are a common model of choice for random time-decay phenomena, because they possess useful

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v1 | CC-BY 4.0 Open Access | rec: 19 Apr 2016, publ: 19 Apr 2016

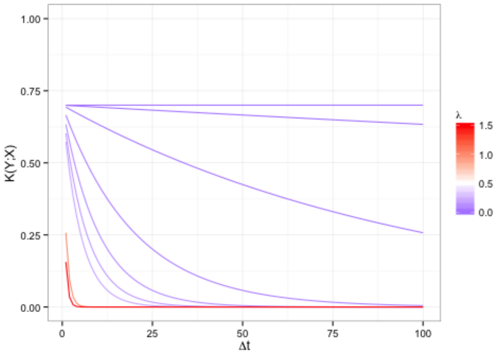

**Figure 6.** Rates of loss of knowledge over time, quantified by the $K$ function with values of the parameter *lambda*.

properties such as being memoryless. Therefore, given a system $Y_t X_0$ and an initial state of knowledge $K_0 = K(Y_t; X_{t_0})$, we can model the rate of knowledge loss as:

$$K(Y_{\Delta t}; X) \sim K_0 e^{-\lambda \Delta t} \tag{55}$$

in which $\Delta t = t' - t$ is the interval between the original time and $\lambda$ is the rate of information loss. Our $\lambda$ is analogous to a Liapunov exponent, but it is unclear how resilience connects to the concept of knowledge decline examined in this essay, Therefore, for our current purposes we will ignore the condition $\lambda < 0$, and assume that all knowledge systems have an associated value $\lambda \geq 0$, with equality corresponding to no loss of knowledge over time, which is the condition of logico-deductive systems (Fig 6).

Lemma 0.11 tells us that every empirical system with a defined value of $K(Y; X)$ will have a characteristic speed with which it approaches its chaos horizon. Since estimating when knowledge is completely lost will be, under most circumstances, easier than estimating conventional measures of decay (e.g. half life), we will use the point of chaos $t^{\dagger}$ to define a system-specific rate of knowledge loss:

$$\Lambda \equiv \frac{1}{t^{\dagger}} log \frac{K_0}{\epsilon} \tag{56}$$

This quantity is highly system-specific and depends not just on properties of the phenomena examined, but also on the accuracy with which these are measured.

## 0.9 Knowledge

We may distinguish three fundamental processes by which knowledge is acquired and/or used: 1) experience, in which information relating to individual events or objects is structured by encoded patterns; 2) experience cumulation, in which patterns are reinforced or weakened; 3) pattern creation, in which encoded patterns are tentatively recombined. The three processes are strictly intertwined, of course, and knowledge emerges dynamically from the interplay of the three.

**Experience** Experience requires the direct application of knowledge, and would not be possible without it. An event or object enters cognition only when it carries some level of uncertainty or, equivalently, non-zero information in its structure and/or frequency, i.e. $H(y) > 0$. The extent to which such uncertainty is reduced by encoded patterns quantifies the amount of *understanding* or *recognizing*. Using our notation:

$$K(y;x) \equiv \frac{H(y)-H(y|x)}{H(y)+H(x)} \tag{57}$$

in which $y, x$ are individual objects or events. The event is explained by the encoding of a relation $H(y|x)$ which is derived from the cumulation of experience (phase described below).

Note that for the $K$ function to have non-zero finite value, the term $H(x)$ can never be zero. This implies that, based on our definition of knowledge, *no object or experience can be explained without additional information.* In knowledge there is no “free lunch”. If this is true, however, and $H(x) > 0$, then it is also true that $K(y;x) < 1$ for any $y$. In other words, explanations can never be ”complete”, because some a priori information always needs to be input. There can be no ”theory of everything” in a literal sense, because such theory will never be able to explain itself.

**Knowledge ≡ Experience Cumulation and Aggregation** Knowledge builds upon experiences, by a process of cumulation in which all the terms involved are updated:

$$K(\overset{\circ}{\sum_n}(y;x)) \equiv \frac{H(\overset{\circ}{\sum}_n y) - H(\overset{\circ}{\sum}_n y|x)}{H(\overset{\circ}{\sum}_n y) + H(\overset{\circ}{\sum}_n x)} \equiv K(Y;X) \tag{cumulation}$$

Note that these individual events are “atomic” from the point of view of the specific cumulation process but may themselves consists of objects of any level of complexity. With the term “aggregation” will indicate a higher-order cumulation of encoded patterns:

$$K(\overset{\circ}{\sum_k}(\overset{\circ}{\sum_n}(y;x))) \equiv K(\overset{\circ}{\sum_k}(Y;X)) \equiv \frac{H(\overset{\circ}{\sum}_k Y) - H(\overset{\circ}{\sum}_k Y|X)}{H(\overset{\circ}{\sum}_k Y) + H(\overset{\circ}{\sum}_k X)} \tag{aggregation}$$

Postulating an aggregation phase is also useful to accommodate temporal and spatial discontinuities with which organisms might encode patterns. Individual organisms usually learn through a cumulation of experiences that might occur at different times in different places. There is also a sense in which populations of organisms encode patterns that are a weighted average (an aggregation) of the patterns encoded within each member of the population.

**Knowledge gained per experience** As $n$ experiences cumulate, knowledge changes at the $n+1$ step as $\Delta K \equiv K(\overset{\circ}{\sum}_{n+1}(y;x)) - K(\overset{\circ}{\sum}_n(y;x))$ which re-arranged, gives:

$$\Delta K \equiv \frac{(C-Q)(H(\overset{\circ}{\sum}_n y) - H(\overset{\circ}{\sum}_n y|x)) + \delta_{y|x} - \delta_y}{H(\overset{\circ}{\sum}_{n+1} y) + H(\overset{\circ}{\sum}_{n+1} x)} \tag{58}$$

with $C = 1 - w_{n+1}$, $Q = \frac{H(\overset{\circ}{\sum}_{n+1} y)+H(\overset{\circ}{\sum}_{n+1} x)}{H(\overset{\circ}{\sum}_n y)+H(\overset{\circ}{\sum}_n x)}$ and $\delta_{y|x}, \delta_y$ are shorthand for the information that is saved when the new events match old ones (see lemma 0.3). When $\Delta K > 0$, the pattern is reinforced and thus knowledge grows; when $\Delta K < 0$ the pattern is weakened and knowledge decreases.

If the new experience does not relate to unknown instances of either explanans or explanandum, then $Q < 1$, $d_y = 0$ and $d_{y|x} > 0$. As the number of experiences cumulated increases, then both C and Q should, all else being equal, converge to 1. Therefore, as intuition would suggest: 1) knowledge always increases when, all else being equal, a new experience is made about a fixed (a known) system; 2) vice versa, when the novel experience brings new uncertainty about explanandum or explanans, then knowledge about the phenomenon is likely to decrease; however, 3) knowledge becomes less likely to change in any direction as the number of cumulated experiences grows, though the rate of this change depends entirely on how new experiences are weighted compared to old ones.

**Creativity ≡ Knowledge Expansion** The ability to encode new patterns is necessarily generated by modification of pre-existing encoded patterns - nothing can be generated from nothing. In practice, this modification must therefore consist in the expansion of the pattern to new attributes of the explanandum, the explanans or both. Independent of the pattern-encoding substratum, the source of novelty will likely include a source of randomness on which the organism subsequently capitalizes. Just as random mutations generate novelty in genomes, it is likely that the random formation new synaptic potential connection is a source of creativity in brains. Whatever their generating mechanism, newly created patterns are by definition "tentative", because they have no pre-defined objective and simply create a potential that will be reinforced or weakened by cumulation and aggregation of experiences.

$$K(\overset{\circ}{\prod}(Y;X)) \equiv \frac{H(\overset{\circ}{\prod}Y) - H(\overset{\circ}{\prod}Y|X)}{H(\overset{\circ}{\prod}Y) + H(\overset{\circ}{\prod}X)} \qquad \text{(expansion)}$$

This process represents and "expansion" because it allows a knower to experience objects or events of increasing complexity (i.e.lower universal probability), opening up new uncertainty spaces to compress. This does not just include creating associations between new variables, but also increasing accuracy, making predictions, etc. Combinations of the $\otimes$ and $\oslash$ operations can be shown to underlie any process by which new potential knowledge is created. We avoid providing details in this section, although a scheme is provided in section 0.11.

## 0.10 Knowledge growth

When experiences are registered as individual events, all distinct and disconnected from one another, knowledge is zero and cannot grow:

$$\overset{\circ}{\sum}K(y;x) \equiv \frac{H(\nu_y) - H(\nu_y|x)}{H(\nu_y) + H(\nu_x)} = \frac{H(\nu) - H(\nu)}{2H(\nu)} = 0 \qquad \text{(disconnected experiences)}$$

When knowledge of individual systems is encoded, (i.e. $K(\overset{\circ}{\sum}y;x) > 0$) but these are cumulated as completely disaggregated, i.e. completely non-overlapping systems total $K$ equals:

$$\overset{\circ}{\sum}K(Y;X) \equiv \frac{\overset{\circ}{\sum}H(Y) - \overset{\circ}{\sum}H(Y|X)}{2H(\underline{w}) + \overset{\circ}{\sum}H(Y) + \overset{\circ}{\sum}H(X)} \qquad \text{(disconnected knowledge)}$$

The $H(\underline{w})$ term reflects the fact that if each system is encoded as non-overlapping, the total uncertainty equals the average uncertainty of each system plus the information necessary to distinguish each system. This term disappears at the numerator but not at the denominator, making disaggregated knowledge a decreasing function of the number of systems involved. We can show how cumulating and aggregating knowledge is an adaptive response to the growing information costs of unstructured information.

**Knowledge growth by cumulation** Cumulating knowledge is a viable strategy if $K(\overset{\circ}{\sum}Y;X) > \overset{\circ}{\sum}K(Y;X)$. We can re-arrange the condition and obtain, using the tilde to represent cumulation (i.e. $\tilde{Z} \equiv \overset{\circ}{\sum}Z$):

$$\overset{\circ}{\sum}D(Y|X||Y\tilde{|}X) - \overset{\circ}{\sum}D(Y||\tilde{Y}) < \overset{\circ}{\sum}K(Y;X)(2H(\underline{w}) - \overset{\circ}{\sum}D(Y||\tilde{Y}) - \overset{\circ}{\sum}D(X||\tilde{X})) \quad (59)$$

We know from lemmas 0.4, 0.5 that the right hand side is $\geq 0$, so the condition is satisfied whenever $\overset{\circ}{\sum}D(Y|X||Y\tilde{|}X) \leq \overset{\circ}{\sum}D(Y||\tilde{Y})$. The advantage of cumulating is null when $\overset{\circ}{\sum}K(Y;X) = 0$ and increases in proportion to the average $K$ and the number (entropy) of systems being aggregated.

**Knowledge growth by complexification** Similarly, expansion of knowledge is a beneficial when, $K(\overset{\circ}{\prod}Y;X) > \overset{\circ}{\sum}K(Y;X)$ which, knowing that $\overset{\circ}{\prod}K(Y;X) \approx \frac{\overset{\circ}{\sum}H(Y) - \overset{\circ}{\sum}H(Y|X)}{\overset{\circ}{\sum}H(Y) + \overset{\circ}{\sum}H(X)}$, can be simplified to:

$$2H(\underline{w}) > \frac{\overset{\circ}{\prod}K(Y;X) - K(\overset{\circ}{\prod}Y;X)}{K(\overset{\circ}{\prod}Y;X)} \quad (60)$$

Note that the right-hand side of the inequality is constrained whereas the left-hand side increases with the number of disaggregated systems. Therefore, as the number of systems in a disaggregated fashion cumulates, it becomes ever more convenient to merge them into a more complex (multidimensional) system, producing knowledge of higher complexity. A more restrictive condition would contemplate systems that do not contain any information of each other. In such case, $H(\underline{w}) = 0$ and equation 60 is satisfied if and only if $\overset{\circ}{\prod}K(Y;X) < K(\overset{\circ}{\prod}Y;X)$.

Therefore, the two essential properties at the basis of biological and cognitive evolution, i.e. the ability to combine sets of events and extract patterns and the tendency to extract patterns of ever higher complexity emerge naturally as strategies to maximize K (reduce redundancy). The higher the number of challenges faced by an organism and benefit accrued from each encoded pattern, the greater the advantage of integrating such patterns into a smaller number of systems of higher complexity. Once knowledge about a system of high complexity is encoded, simplification of the new system is likely to become beneficial, too.

**Simplification** Let $YX$ be a system, and let $T(Y)$ and $T(X)$ be their respective total information contents. As shown in section 0.3, we can always partition the total information in a Shannon entropy component and a ”residual complexity” component. The criterion of such partitioning is entirely arbitrary. A system is said to be simplified when the entropy components of explanandum or explanans are chosen such that:

$$\frac{H(Y) - H(Y|X) + D(Y||Y_U) - D(Y|X)||Y|X_U)}{H(Y) + D(Y||Y_U) + H(X) + D(X||X_U)} < \frac{H(Y) - H(Y|X)}{H(Y) + H(X)} \quad (61)$$

the simplification will be maximized when the partitioning is such that $D(Y||Y_U) = D(Y|X)$, i.e. when the explanans contains no relevant information about the residual complexity terms. Each term can thus be reduced to its essential properties, a process we call abstraction:

$$T(Y) \to T(Y \oslash Y^c) \equiv H(X) \quad \text{and} \quad T(X) \to T(X \oslash X^c) \equiv H(X) \tag{62}$$

The recursive interplay of cumulation and transformation and simplification creates structures of ever-increasing order and growing level of abstraction and complexity.

## 0.11 Science

Scientific knowledge, as argued in the introduction, is not fundamentally different from ordinary knowledge: it consists in the encoding of patterns that reduce uncertainty about phenomena. Science, however, is a particularly powerful pattern-encoding activity, which allowed human beings to move beyond their ordinary perceptual and cognitive capacities. This power comes to science by the adoption of rules and practices which we generically call the "scientific method". It would be incorrect to identify science with any specific set of rules, as some attempted to do in the past ( 2), because the scientific method is certainly in constant evolution. However, the *explication* of a methodology clearly represents a defining feature of any activity that aspires to be considered scientific - including a pseudoscience.

The term "methodology" is used in this essay in the broadest possible sense, to indicate all theories, models, procedures, instrumentation that contributed to a result. From a practical point of view, the methodology of a study is embodied (albeit imperfectly, as argued below) in the Introduction and Materials and Methods sections of the study's publication. From a theoretical point of view, a study's methodology should be conceived as the algorithm that, given the explanans as input, yields a certain amount of information (reduced uncertainty) about the explanandum.

If methodology is part of scientific explanation, then its natural position in the $K$ function is to be part of the explanans. More specifically, methodology is a component of the explanans which, at least in principle (i.e. in absence of bias, see section 0.15), is independent of both explanans and explanandum. Mathematically, this concept translates in the partitioning of the explanans in two components: $x \to xm$, and in positing that $H(xm) = H(x) + H(m)$, $H(y|m) = H(y)$, $H(y|mx) \leq H(y|x)$.

**Methodology as conditioning** The mathematization of scientific methodology proposed above suggests a rather new interpretation the concept of scientific methodology: it is *a conditioning* of the pattern, in other words a conditioning of the knowledge claimed by a study about a particular system $YX$. We can visualize this concept by imagining the set of all possibles methodologies applicable to a system $YX$.

Let $YX$ be a system of interest and let $m^*$ be the methodology adopted by a given study. Let $l^*$ be the minimum description length of $m^*$, expressed as a binary string. The content, and therefore the length, of $m$ will depend on the nature of the system and choices made by scientists. However, postulate 1 guarantees that this description is finite, and therefore that for every system $YX$ there is an $l^*$ of finite length. The universal probability of such string is $P_U(l^*) = 2^{-l^*}$, and therefore the methodology's Kolmogorov complexity, is $C(m^*) \approx log(2^{l^*}) = l^*$. Now, let $\aleph_{YX} : \{YXm_1, YXm_2...\}$ be the set of all possible studies about the system $YX$ that use methodology description lengths $l^*$. This set will have $2^{l^*}$ elements, each occuring with a probability $2^{-l^*}$. To pin-point a specific element of this set, we would need exactly $l^* \approx C(m^*)$ bits.

Therefore, the complexity of a study's methodology is also the minimum information required to identify that particular methodology amongst all its possible alternatives of

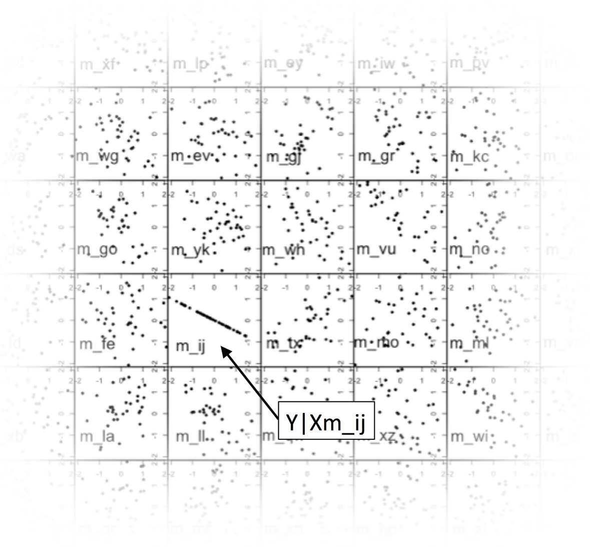

**Figure 7.** Pictorial representation of the set of all possible methodologies, $\aleph$, with fading borders to emphasize the very large size $L$ of this set. If the study with methodology $m_ij$ is one element of this set, the Kolmogorov complexity of the methodology's description, $C(m_ij)$ is equivalent to the information required to identify it, i.e. Shannon's information $H(L)$.

equivalent description length. It is therefore also the minimum information that identifies a particular study of $YX$ amongst all the possible studies that could have been conducted with the given methodology description length. The more "complex" a methodology, the longer its minimum description length, the larger the size of its associated $\aleph_{YX}$ set (Fig 7).

Specifying, as we will henceforth do, a study by its methodology $m$, is equivalent to representing study outcomes $YXm_i$ as instantiations of a random variable with uniform probability $2^{-C(m)}$ and conditioning that random variable it to the case $YXm_i = YXm$.

Whilst all the above holds in theory, in practice the methodology of a study is rarely accessible in complete form. A research publication, for example, will usually omit large amounts of information that are either assumed to be known by the targeted readers or are (incorrectly) deemed to be irrelevant. Most other information will not be omitted entirely but will only be referred to by acronym, name or by reference to other literature. In practice, therefore, if $m^*$ is the complete methodology and has complexity $C(m^*)$, its real-life equivalent will be a simpler object $m$, conditioned version of $m^*$, with description length $l < l^*$ and complexity $C(m) < C(m^*)$. This has important consequences for the concept of reproducibility (see section 0.13).

**Experience $\equiv$ Study** What in ordinary knowledge was defined as experience in science is generally referred to as a "study", which can be empirical or theoretical. Depending on the type of study, the explanandum $y$ may be a single case (event), a sample of events (i.e. an object whose probability is the joint probability of the events), a controlled experiment, or it might consist in a theoretical construct. For example, the uncertainty of the explanandum might correspond to the plausibility of a theory or hypothesis, in which case the explanans could be an event that is relevant to the theory's likelihood.

$$K(y;mx) = \frac{H(y) - H(y|mx)}{H(y) + H(x) + H(m)} \qquad \text{(individual study)}$$

In absence of bias (see section 0.15) $H(y)$ and $H(x)$ are independent of the methodology, and K will be maximized by making the latter as simple as possible. If $yx$ represent a sample of data, then a fundamental way to minimize the $m$ term consists in keeping it constant and independent of $yx$. Hence a tenet of the scientific method, namely that data should be collected exactly in the same way - deemed so important as to be considered by some the essence of science itself (e.g. [6])- emerges as a natural strategy to maximize $K$.

Knowledge of theoretical (i.e. logico-deductive) systems differs from empirical knowledge only in one respect. Since by definition logico-deductive systems have no measurement error, then knowledge about them is built on a series of "rigid" patters $H(Y|X) = 0$, in other words of *identities*. A typical hypothesis underlying a theoretical study, for example a mathematical conjecture, proposes that two theoretical entities that were previously believed to be disconnected are actually identical, because one can be derived from the other. If $y$ and $x$ are such entities, the expectation following the hypothesis is $H(y|x) = 0$ and the study consists in assessing whether a chain of identities $H(x'|x) = 0, H(x''|x') = 0...$ connects explanandum to explanans. In this case, the $m$ component of the explanans will represent the complexity of the (description of the) proof. Therefore, like all other forms of knowledge, logico-deductive knowledge is affected by the information costs (complexity) of the explanans. This explains why mathematical and theoretical researchers value very highly a results' simplicity and elegance, which the aesthetic translation of simplicity.

**Knowledge $\equiv$ Literature** The scientific equivalent of the cumulation and aggregation of experience is a literature. This is the knowledge transmitted through scientific publications, summarized in reviews, books and university courses, and generally the knowledge manifested in the expertise of scientists.

$$K(\overset{\circ}{\sum} y;mx) \equiv \frac{H(Y) - H(Y|MX)}{H(Y) + H(X) + H(M)} \qquad \text{(study cumulation)}$$

$$K(\overset{\circ}{\sum} Y;MX) \equiv \frac{H(\overset{\circ}{\sum} Y) - H(\overset{\circ}{\sum} Y|MX)}{H(\overset{\circ}{\sum} Y) + H(\overset{\circ}{\sum} X) + H(\overset{\circ}{\sum} M)} \qquad \text{(literature aggregation)}$$

The methodology terms $H\overset{\circ}{\sum} m$ and $H\overset{\circ}{\sum} M$ are by assumption controllable to some extent. In particular, their values will be minimized in proportion to how invariant methodologies are across studies. The system terms $Y$ and $X$ are instead external phenomena, which by assumption have independent properties.

**Definition: stable system**. Under ideal conditions, phenomena studied by a field maintain exactly the same characteristics from one study to the next. We can impose this condition on any system by requiring $H(\overset{\circ}{\sum} y) \approx H(y)$, $H(\overset{\circ}{\sum} Y) \approx H(Y)$ and $H(\overset{\circ}{\sum} x) \approx H(x)$, $H(\overset{\circ}{\sum} X) \approx H(X)$. A system with such property will be referred to as that of a “stable system”.

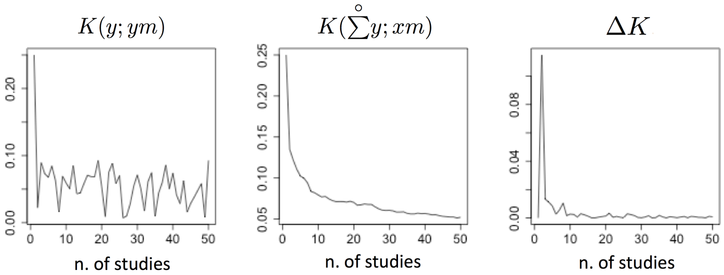

**Figure 8.** Simulation of cumulation of studies within a defined field. A) Values of K in simulated studies: the first study reports a strong result $K(y;ym) >> 0$, but subsequent studies report null or close-to-null results. B) Corresponding value of the cumulative knowledge $K(\overset{\circ}{\sum}y;xm)$. C) Corresponding information value of each study, as measured by $\Delta K$. Note the initial peak in $\Delta K$, which illustrates how negative results can be just as informative as positive result. The information value of either will however decline over time, particularly if results are consistent with each other.

**Knowledge gain per study** The calculation is the same as that for ordinary knowledge. Applied, to simplfy calculations, to a stable system, it yields:

$$\Delta K(y;xm) \equiv \frac{H(y) - H(\overset{\circ}{\sum}_{k+1}y|xm)}{H(y)+H(x)+H(\overset{\circ}{\sum}_{k+1}m)} - \frac{H(y) - H(\overset{\circ}{\sum}_{k}y|xm)}{H(y)+H(x)+H(\overset{\circ}{\sum}_{k}m)} = \frac{(1-Q)H(Y) + (Q-C)H(\overset{\circ}{\sum}_{k}y|xm) + d_{y|xm}}{H(Y)+H(X)+H(\overset{\circ}{\sum}_{k+1}m)} \tag{63}$$

with $C$ and $d$ defined as before and $Q = \frac{H(Y)+H(X)+H(\overset{\circ}{\sum}_{k+1}m)}{H(Y)+H(X)+H(\overset{\circ}{\sum}_{k}m)}$. Hence, how knowledge changes will be a function of the changes in cumulative complexity of methods, given by Q, as well as the cumulative effect size observed, given by $H(\overset{\circ}{\sum}_{k}y|xm)$. In particular, when methodology does not vary across studies, then $Q < 1$ and knowledge increases or decreases depending on whether the new effect aligns with previous ones, as measured by the sign of $d_{y|x}$. The magnitude of change, however, is inversely proportional (1-C), i.e. the relative weight of the new study. As studies cumulate, therefore, the information yielded by each additional study decreases.

Eventually, the cumulation/aggregation of new studies will cease to alter K, i.e. $|\Delta K(Y;XM)| \approx e$, with $e$ representing some arbitrary threshold. This condition marks the end of the cumulation/aggregation processes, at which point the hypothesis is deemed to be either verified or falsified (Fig 8).

**Verification vs. Falsification** Let $Y$ be an explanandum and let $\underline{X}_n$ be a joint distribution of conclusively verified explanantia such that $K(Y;\underline{X}_n) > 0$ and let $X_{n+1}$ be an additional candidate explanans that a literature has also conclusively investigated around the hypothesis $H(Y|\underline{X}_n \otimes X_{n+1}) < H(Y|\underline{X}_n)$. If the literature supports the hypothesis, then:

$$K(Y;M\underline{X}_n \otimes X_{n+1}) \geq v \tag{verification}$$

with $v >> 0$ being an arbitrary threshold that marks the substantive significance of the pattern. The knowledge gained in this case is given by

$\Delta K \equiv K(Y; M\underline{X}_n \otimes X_{n+1}) - K(Y; M\underline{X}_n)$ which, for a stable system gives $\Delta K \propto H(Y|M\underline{X}_n) - H(Y|M\underline{X}_n \otimes X_{n+1})$.

If the literature does not support the hypothesis:

$$K(Y; M\underline{X}_n \otimes X_{n+1}) < v \tag{falsification}$$

which is equivalent to the claim that $H(Y|M\underline{X}_n \otimes X_{n+1}) \approx H(Y|M\underline{X}_n)$. In this case, the only knowledge obtained comes from the elimination of the hypothesis itself. If we let $\Omega; \{X_1, X_2 ... X_m\}$ be a set of $m$ candidate explanantia and let $X_{n+1}$ be an element of $\Omega$ (let $m > n + 1$), then the information gained by conclusively falsifying hypothesis $X_{n+1}$, is $H(\Omega) - H(\Omega \oslash X_{n+1})$.

The result above points to an asymmetry between verification and falsification that verbal arguments might overlook. The uncertainty about explanantia is reduced also when the hypothesis is verified. To correctly analyse the case, therefore, we need to consider the total amount of knowledge that is potentially gained by testing the hypothesis. Omitting for simplicity the $M$ term:

$$\begin{aligned} K(Y;\Omega) \oplus K(\Omega;Y) &= \frac{w_1 H(Y) + w_2 H(\Omega) - w_1 H(Y|\Omega) - w_2 H(\Omega|Y)}{H(Y) + H(\Omega)} = \\ &= \frac{w_1 H(Y) + w_2 H(\Omega) - w_1 H(Y|\underline{X}_n) - w_2 H(\Omega \oslash \underline{X}_n)}{H(Y) + H(\Omega)} \end{aligned} \tag{64}$$

Following the conclusive test of a hypothesis $X_{n+1}$, this knowledge will change as:

$$\begin{aligned} &\Delta K \equiv K(Y; \underline{X}_n \otimes X_{n+1}) \oplus K(\Omega; \underline{X}_n \oslash X_{n+1}) - K(Y; \underline{X}_n) \oplus K(\Omega; \underline{X}_n \otimes X_{n+1}) = \\ &= \frac{w_1 (H(Y|\underline{X}_n) - H(Y|\underline{X}_n \otimes X_{n+1})) + w_2 (H(\Omega \oslash \underline{X}_n) - H(\Omega \oslash (\underline{X}_n \otimes X_{n+1})))}{H(Y) + H(\Omega)} \end{aligned} \tag{65}$$

Verifications and falsifications may be valued arbitrarily differently by altering the values of the weights $w_1, w_2$. However, unless one of these values is set to zero, a verification will always yield higher $\Delta K$ than a falsification (Fig 9). Under maximally informative conditions, all hypotheses are equally likely a priori, and the value of a falsification will be:

$$\Delta K_{falsif} \quad \propto \quad log(\frac{|\Omega \oslash \underline{X}_n|}{|\Omega \oslash \underline{X}_n| - 1}) \tag{66}$$

which is maximal when $|\Omega \oslash \underline{X}_n| = 2$, and declines rapidly when this quantity increases. Therefore, the value of a conclusive test of a single hypothesis is inversely proportional to the number of hypotheses that remain untested (Fig 9).

Equation 65 will be maximized, in particular, when $H(Y|\underline{X}_n) = H(Y), H(\Omega \oslash \underline{X}_n) = H(\Omega)$ and the conditional terms are zero, which means that the test ruled out all but one of the hypotheses, and that the explanandum is fully determined by the remaining hypothesis. In other words, there is a one-to-one correspondence between the instantiations of the explanandum and each of the hypotheses, i.e. $H(Y) = H(\Omega)$ and $H(Y|\Omega) = H(\Omega|Y) = 0$. This scenario corresponds to Popperian falsificationism as it is generally intended: a set of mutually exclusive outcomes $y$ each of which is univocally associated with an hypothesis $x$. As we have just shown, these are indeed conditions that maximize the knowledge gain. However, these are also conditions in which verification is redundant, because both explanans and explanandum carry the same information. Therefore, Popperian falsificationism

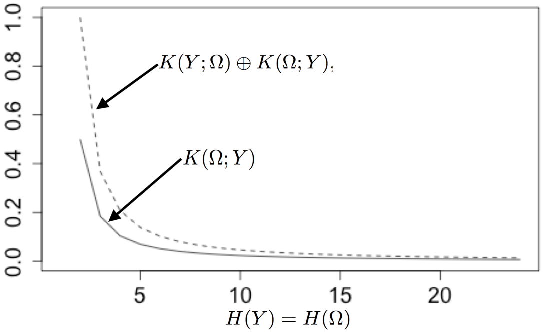

**Figure 9.** Knowledge gained by the conclusive falsification of one single hypothesis, as a function of the total number n of candidate hypotheses (candidate explanantia). Solid line represents the values of $K(\Omega;Y)$, i.e. the knowledge about the set of candidate hypotheses. Dotted line represents the values of $K(Y;\Omega) \oplus K(\Omega;Y)$, i.e. the combined knowledge of explanantia and explanandum, when $H(Y) = H(\Omega)$ and the falsification about an explanans brings a corresponding reduction of uncertainty about the explanandum. Knowledge is maximized when $H(Y) = H(\Omega) = 2$ and declines rapidly thereafter.

represents a limiting condition in which falsification and verification convey exactly the same information. In many, possibly most, fields of research, this condition might not be realizable, in which case verifications are more valuable than falsifications.

**Creativity $\equiv$ Hypothesis Generation** New hypotheses and new theories (which are just hypotheses about identities) are tentatively advanced by the same processes underlying creativity in general, i.e. by expansion and/or reduction of explanans and explanandum. Behind the infinite creativity of scientists, there seem to lie a few basic forms of expansion and reduction, which can be ranked by their increasing creativity and potential knowledge yield. We will use the symbol $\succ$ and $\prec$ in place of $>$ and $<$ to represent the tentative nature of these hypotheses:

1. increase/decrease accuracy of explanans and/or explanandum: e.g. $H(Y^{\alpha'_y}|X) \prec H(Y^{\alpha_y})$ with $Y^{\alpha'_y} = Y^{\alpha_y} \otimes Z$, with $H(Z) = log(n_z)$ and $\alpha'_y = \alpha_y / n_z$; and similarly for the explanandum. The same could apply to $X$. As shown in section 0.8, there is an optimal value of accuracy that is highly specific to each system. Once a pattern is discovered, it may be a natural research objective to identify the level of accuracy that maximizes it.

2. increase/decrease time: $H(Y_{t'}|X_{t_0}) \prec H(Y_t|X_{t_0})$ with $Y_{t'} = Y_t \otimes Y_{t'} \oslash Y_t$. Studies would, in this case, explore how far forward or backwards in time a certain explanans $X$ retains information about the explanandum.

3. expand the explanans: $H(Y|X \otimes X') \prec H(Y|X)$ with $X'$ a new candidate explanans. Explanations are combined, thereby increasing the overall complexity of the explanans but gaining proportionally more information about the explanandum.

4. reduce the explanandum: $H(Y'|X) \prec H(Y|X)$ with $Y' = Y \oslash Y^c$. This process might take the form, for example, of theoretical abstraction, (see section 0.9) or

empirical conditioning, in which the system is re-scaled to include only unexplained portions of the explanandum.

5. reduce the explanans: $H(Y|X') \prec H(Y|X)$ with $X' = X \oslash X^c$; and similarly for methodology. Theoretical explanantia and methodologies are abstracted and/or compressed, whereas empirical components are simplified, for example, by randomization, in which information between $X'$ and $Y$ is actively destroyed, or stabilization, in which a value is imposed, i.e. $X' = X'|do(X' = constant)$. In either case, the condition achieved is $H(Y|X') = H(Y)$.

6. expand the explanandum: $K(Y \otimes Z; X) \succ K(Y; X)$, in which new phenomena are subsumed under the same explanantions, which is a fundamental presupposition for knowledge growth.

The last two forms of scientific innovations lead to the greatest form of progress, in which an increasing range of phenomena are explained by a relatively decreasing set of explanatory principles. This is the phenomenon of consilience, rightfully indicated as the ultimate objective of scientific knowledge (and arguably also of ordinary knowledge, as well as of life itself). Since reduction is just the inverse of expansion, and since explanans, explanandum and methodology may be altered by separate and different processes of expansion, we can represent consilience as an independent expansion of the the three components:

$$K(\overset{\circ}{\prod}(Y; MX)) \equiv \frac{H(\overset{\circ}{\prod}_n Y) - H(\overset{\circ}{\prod}_n Y|\overset{\circ}{\prod}_k M \overset{\circ}{\prod}_m X)}{H(\overset{\circ}{\prod}_n Y) + H(\overset{\circ}{\prod}_m X) + H(\overset{\circ}{\prod}_k M)} \qquad \text{(consilience)}$$

with $n, m, k$ indicating the different dimensions of each component.

## 0.12 Scientific progress

If $K(Y; mX)_i$ is the amount of knowledge about a system at stage $i$, progress is achieved at stage $i + 1$ in proportion to $|\Delta K(Y; mX)| > 0$, with $\Delta K(Y; mX) \equiv K(Y; mX)_{i+1} - K(Y; mX)_i$. Scientific progress, however, takes different forms in the cumulation and expansion phases.

**Progress in verification/falsification** is achieved when a specific hypothesis is conclusively accepted or rejected. For a given system $YX$, this stage is achieved when new evidence ceases to be informative, i.e. when the cumulation/aggregation yields $\Delta K \approx 0$. Given a cumulation of $k$ studies, if $K(\overset{\circ}{\sum}_{k+1} y; mx) = K(\overset{\circ}{\sum}_{k+1} y; mx) \pm e$ with $e \geq 0$ an error term, then

$$(1 - Q)H(\overset{\circ}{\sum}_k y) + (Q - C)H(\overset{\circ}{\sum}_k y|xm) + d_{y|xm} \leq e \qquad (67)$$

with $C$ and $d$ defined as before and $Q = \frac{H(\overset{\circ}{\sum}_{k+1} y)+H(\overset{\circ}{\sum}_{k+1} x)+H(\overset{\circ}{\sum}_{k+1} m)}{H(\overset{\circ}{\sum}_k y)+H(\overset{\circ}{\sum}_k x)+H(\overset{\circ}{\sum}_k m)}$. Similar conditions would apply to the case of aggregation. Either case can be illustrated by taking the derivative of $K$ with respect to the number $k$ of cumulated/aggregated studies. Although technically only valid to the limit of an infinite number of studies, this is a practical way to analyze the speed of information gain per study. When new evidence ceases to be informative:

$$\begin{aligned}&\frac{dK(\overset{\circ}{\sum}_k y;xm)}{dk}=0 \iff \\ &\quad H'(\overset{\circ}{\sum}_k y)(1-K(\overset{\circ}{\sum}_k y;xm))-K(\overset{\circ}{\sum}_k y;xm)(H'(\overset{\circ}{\sum}_k x)+H'(\overset{\circ}{\sum}_k m))= \\ &\quad = H'(\overset{\circ}{\sum}_k y|xm)\end{aligned} \tag{68}$$

in which $H'$ indicates the first derivative of the entropy function with respect to the number of cumulated/aggregated studies, $k$. Of particular interest is the case in which the system is stable (explanans and explanandum do not change with cumulation) but the methodology varies across studies. The condition of stability simplifies equation 68 to:

$$H'(\overset{\circ}{\sum}_k Y|Xm) = -K(\overset{\circ}{\sum}_k Y;Xm)H'(\overset{\circ}{\sum}_k m) \leq -K(\overset{\circ}{\sum}_k Y;Xm)log(k) \tag{69}$$

The negative sign on the right side of the inequality proves that when every new study uses new methods and therefore adds complexity to the explanans, then no more knowledge is added even when the cumulation of studies is apparently reporting effects of increasing magnitude. Viceversa, we can reverse the inequality and see that $\frac{dK(\overset{\circ}{\sum}_k Y;Xm)}{dk} > 0$ leads to:

$$H'(\overset{\circ}{\sum}_k m) \geq -\frac{H'(\overset{\circ}{\sum}_k Y|Xm)}{K(\overset{\circ}{\sum}_k Y;Xm)} \tag{70}$$

which suggests that methodologies could increase in heterogeneity (complexity) in order to maintain an apparent knowledge increase in cumulation. Inequality 70 suggests that methodological heterogeneity is more likely to increase in a cumulating literature in which effects are small and decreasing.

**Speed of progress in verification/falsification** within a field is quantified by the rapidity which which the field reaches the stage $\Delta K \approx 0$. This speed can be empirically measured in years, man-years, total expenditure, the number of scientific studies required to reach stability, or perhaps ideally a combination of all these. The speed of cumulation is likely to vary enormously by field, depending both on characteristics of the phenomena studied (e.g. on the consistency of patterns across studies) as well as implicit or explicit choices made by researchers. These choices include, for example, the threshold below which $\Delta K$is considered negligible or the relative weight given to new studies compared to old (e.g. Fig 10).

**Progress in Consilience** represents the ultimate expression of scientific progress. Section 0.11 offered a schematic representation of how explanans and explanandum can be manipulated through expansions and reductions to generate new tentative hypotheses. When such hypotheses are confirmed (because stability is achieved in the verification/falsification phase) an innovation is established and progress is proportional to $\Delta K \equiv K(\overset{\circ}{\prod}_{k+1} Y;XM) - K(\overset{\circ}{\prod}_k Y;XM)$. We can analyse this process most effectively by taking a generic derivative of K, this time with respect to the number of systems (dimensions) that knowledge expands to. It is easy to show that the condition for progress is:

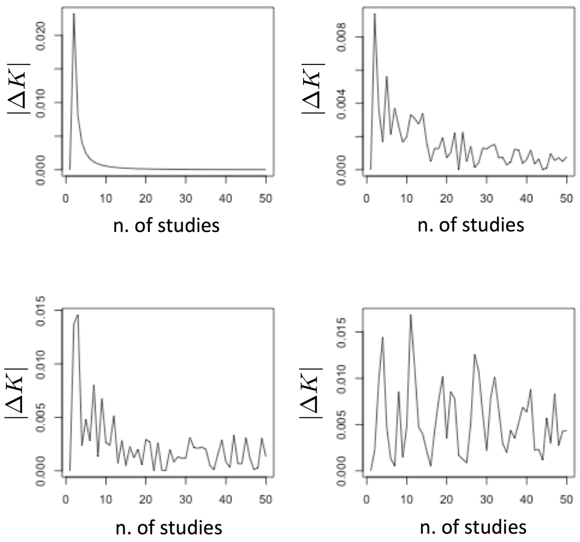

**Figure 10.** Simulation illustrating how a field's speed of verification/falsification might be affected by properties of its subject matter and by methodological choices. All four panels report values of $|\Delta K|$ following the cumulation of 50 studies under the following conditions: A) Stable system and identical methodologies and results across studies, all studies weighted equally (i.e. $w = 1/k$); B) Stable system, identical methodology but results subject to uniform noise, all studies weighted equally; C) Unstable system, varying methodologies and noise in results, all studies weighted equally; D) Unstable system, varying methodologies, noise in results and weights favouring later studies ($w = 1/\sqrt{k}$).

$$\frac{dK(\overset{\circ}{\prod}_n Y; XM)}{dn} > 0 \iff K(\overset{\circ}{\prod}_n Y; XM) < K(\overset{\circ}{\prod}_n Y'; X'M') \tag{71}$$

in which $K(\overset{\circ}{\prod}_n Y'; X'M') \equiv \frac{H'(\overset{\circ}{\prod}_n Y) - H'(\overset{\circ}{\prod}_n Y|MX)}{H'(\overset{\circ}{\prod}_n Y) + H'(\overset{\circ}{\prod}_n X) + H'(\overset{\circ}{\prod}_n M)}$ is the $K$ function with the entropies replaced by their first derivatives with respect to $n$. Assuming that explanans, explanandum and methodology are independent of one another, the effects that each has on progress are easy to tease apart. All else being equal, scientific progress is achieved in proportion to how much the explanans $X$ and the methodology $M$ are minimized (simplified) and how much the explanandum $Y$ is expanded, subject to the condition that the expansion of $Y$ must be larger than that of its conditional counter part $Y|X$.

## 0.13 Reproducibility

The conclusive verification of a hypothesis presupposes the independent replication of studies that support it. Following current terminology, a study whose results are confirmed is said to be reproducible. Failure to reproduce an original finding is very often interpreted as evidence that the original claim was false, and that the original methodology was biased or flawed. However, our approach to defining scientific methodology suggests that failed replications are virtually inevitable in all fields of science, and may have nothing to do with the validity of the claim being tested. Failed replications are the inevitable consequence of limited information about a study's methodology.

Let $K(y; xm) > 0$ be the claim made by a study that reported methodology $m$. Let $K(y'; x'm')$ be the result of a replication attempt. To simplify the analysis, we will assume that the system $YX$ is stable (and therefore $H(y) \equiv H(Y)$ and $H(x) \equiv H(X)$). Hence, a claim about successful reproducibility is translated as $K(Y; Xm) \approx K(Y; Xm')$, a condition which is solely dependent on whether $H(m) = H(m')$ and $H(Y|Xm) = H(Y|Xm')$. If $m = m'$, the two methodologies are indeed identical, then the condition is certainly true. However, such identity could be assumed to occur only if all the information required to produce the pattern is available and fully matched. This condition can be pictured as having the code to retrieve the specific study out of the set of all possible studies $\aleph_{YX}$ with methodology of complexity $C(m) = l$ (see Fig 7).

We have argued in section 0.11 that in practice the description of a study's methodology is never complete. It omits details that the authors of a study assume that their colleagues know, and it also omits details that, unbeknownst to the authors of the study, are crucial to produce the result. Therefore let $m^*$ be the ideal description of the study $m$'s methodology, i.e. the information that is necessary to condition system $YX$ exactly as study $m$ did, and let $l^* > l$ be the ideal description length. The incompletely reported methodology $m$ is missing $L = l^* - l$ bits, which is equivalent to saying that methodology $m$ represents not one study, but a subset of $\aleph_Y X$ that counts $2^L$ possible other methodologies. All these methodologies contain the description $m$, but vary at random with respect to the remaining $L$ bits.

Therefore, even if technically identical to $m$, a replication methodology $m'$ is effectively a random draw from the set of all $2^L$ possible variations of $m$. The expected value of $K(Y; Xm')$, therefore, is not $K(Y; Xm)$ but $K(\overset{\circ}{\sum}_{2^L} Y; Xm)$, and the condition of reproducibility is given by

$$K(Y; Xm) \approx K(Y; Xm') \iff H(Y|Xm) \approx \overset{\circ}{\sum}_{2^L} H(Y|Xm) + D(Y|Xm||Y|Xm_U) \tag{72}$$

With $0 \leq D(Y|Xm||Y|Xm_U) \leq L$. Hence, even under the ideal conditions of a stable system, any study whose result was not perfectly overlapping with the original claim would fail to meet the condition of successful replication, unless it reported a *stronger* pattern than the original study. In a condition of perfect randomness and no bias, this will only occur 50% of the time.

Naturally, the specific reproducibility success rate will be determined by the criteria of success, i.e. by the similarity threshold above which results are considered to be reproduced. This analysis suggests that such criteria need to be tailored to the characteristics of the system being analysed, in ways that future work should explore in detail.

Albeit preliminary and limited in many ways, this analysis suggests that no study should be expected to be perfectly reproducible, because no study is likely to report complete information about its methodology. Some of the missing information might correspond to what is defined as "bias" (analysed in section 0.15), but most information is likely to be omitted knowingly. When information about a methodology is omitted because it is assumed to be known, then the expertise of the replicators will substantially affect the likelihood to reproduce a finding. When however information is omitted unknowingly, because it is (incorrectly) assumed to be irrelevant, then the reproducibility study will truly be a random draw of the kind described above. In either case, the probability to reproduce a study is likely to be directly proportional to the amount of missing information and to the robustness of study's results to methodological variation. Studies with complex methodologies and complex systems, i.e. systems that are sensitive to multiple variables, are at greater risk from reproducibility failure.

## 0.14 Soft science

The various criteria proposed to distinguish stereotypically "hard" sciences like physics from stereotypically "soft" ones like sociology cluster along two relevant dimensions (see [27]):

- Complexity: from the physical to the social sciences subject matters go from being simple and general to being complex and particular. This increase in complexity corresponds, intuitively, to an increase in systems' number of relevant variables and the intricacy of their interactions.
- Consensus: from the physical to the social sciences, there is a decline in the ability of scientists to reach agreement over the interpretation and the relevance of findings, over the correct methodologies to use, even on the relevant research questions to ask, and therefore ultimately on the validity of any particular theory.

Both concepts have a straightforward mathematical interpretation, which points to the same underlying characteristic: having a relatively complex explanans and therefore a low $K$. A system with many interacting variables is a system for which $H(X)$ and/or $H(Y|mX)$ are high. A system (a field, in this case) for which consensus is low is one in which $H(\overset{\circ}{\sum} m) \equiv H(M)$ is high. Therefore, at the core of the differences between a "soft" and a "hard" field is a difference in K, i.e. a difference in the quantity of knowledge that can be attained with a given explanandum.

We can define and quantify the concept more formally as follows. Let $A$ and $B$ be two fields with K values $K(Y_A; X_A M_A)$ and $K(Y_B; X_B M_B)$. Field A is softer than B if when $K(Y_A; X_A M_A) = K(Y_B; X_B M_B)$ then $H(X_A M_A) > H(X_B M_B)$ and, vice versa, when $H(X_A M_A) = H(X_B M_B)$ then $K(Y_A; XM_A) < K(Y_B; XM_B)$. We can express both conditions effectively by partitioning $K(Y; XM)$ and defining $A$ as softer than $B$ if

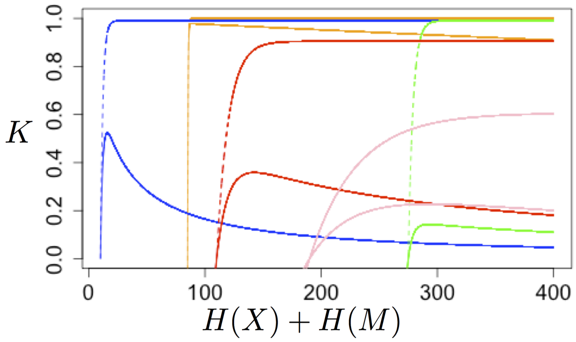

**Figure 11.** Simulated examples of knowledge curves with different values of $H(Y), \Lambda, \gamma, H(x_0)$. Dotted lines represent values of $k(Y;XM)$ whilst solid lines correspond to $k(Y;XM) \times h$. See text for further details.

$$k(Y_A;M_AX_A)h_A < k(Y_B;M_BX_B)h_B \tag{73}$$

in which

$$k(Y;XM) \equiv \frac{H(Y) - H(Y|MX)}{H(Y)} \quad \text{and} \quad h \equiv \frac{1}{1 + \frac{H(M)+H(X)}{H(Y)}} \tag{74}$$

are, respectively, the uncorrected $K$ - in essence a measure of effect size - and a “hardness” factor expressing the information cost of such $k$, i.e. the complexity of the explanans relative to the explanandum.

The relationship defined by condition 73 is never an absolute or a static condition. On the contrary, softness thus defined is a relative concept and a dynamic one, which can be increased or decreased depending on choices which are largely under control of scientists. These choices are about the system to be studied as well as all other parameters than can alter over time the system and its potential to yield knowledge. Therefore, the $K$ of a specific field is determined by an interplay between intrinsic properties of the system and methodological choices made in studying such system, which include choosing the explanantia and explanandum themselves, as well as the time at which they are measured, the accuracy of measurement, the “size” (entropy) of the explanans, etc.

If we identify a field with a specific explanandum, we can represent its dynamics as a curve that connects successive values of $H(X) + H(M)$ with their corresponding $K(Y;XM)$ (Fig 11). This curve is highly field-specific and can be shown to represent an upper limit to all real-life curves (see S1-Text). As figure 11 illustrates, the relative softness of a field is a highly contingent process, which appropriate manipulations of the explanans or the explanandum can change even under the most ideal conditions. Real-life curves are sub-optimal realizations of the curves in figure 11, which means that they grow less rapidly and reach lower maxima. Therefore, a real (non-ideal) research field will typically be a “softer” version of its ideal self.

**Example:** We wish to compare the hardness of two fields that use similar methodologies, namely multiple regression. Starting with the simplest case, let $X_A$ and $X_B$ be two binary dummy variables with the same distribution, such that $H(X_A) = H(X_B)$, and let $H(Y_A) = H(Y_B)$. Assuming that these effects are well-established within each field, the harder field is the one whose models explains the largest amount of variance with the greatest accuracy. If instead we had

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v1 | CC-BY 4.0 Open Access | rec: 19 Apr 2016, publ: 19 Apr 2016

$H(Y_A) = H(Y_B)$ and each fields regression models explained exactly the same amount variance using only one variable, but with $X_A$ having lower information (e.g. $X_A$ is a binary category whilst $X_B$ is continuous and measured to the third decimal point), then field $A$ would he harder than $B$. Under the same conditions, if the models included multiple independent variables, the softer field would likely be the one with the greatest number of high-entropy independent variables. However, a more accurate estimation could attempt to reconstruct the field's knowledge curve, by adding variables progressively to the model, in decreasing order of effect size, and then fitting an exponential (knowledge gain) curve to each pair $H(X), k(Y;X)$ obtained.

## 0.15 Bias

Bias as commonly discussed in the meta-research literature entails that particular methodological choices made in one or more phases of a study (e.g. its conception, design, data collection, analysis, interpretation, publication) conferred an "unfair advantage" to a specific outcome. A common distinction made in the literature is between choices that are conscious and perhaps intentionally misleading and choices that are unconscious and unintentionally flawed. From an information theoretic point of view, however, it is more relevant to draw a distinction between methodological choices made *after* the results are known (post-hoc) and choices made *before* (ante-hoc).

**Post-hoc methodological choices** make use of the same information that they are supposed to produce, to yield a secondary result. Whether such information is processed consciously or unconsciously is irrelevant, because the mechanism is the same. Although a bias can go in any direction, it is usually realistic to assume that the choices go in the direction of producing a positive and significant result, i.e. a higher value of $K$. We will make this assumption henceforth. A broad range of behaviours discussed in the literature fall into this category, including data falsification, data-related "questionable research practices", p-hacking, HARKing etc. We can schematically represent the information flow in post-hoc methodologies as a causal chain, going from ante-hoc methodology $m_{t_0}$, to data, to a post-hoc expansion of the methodology:

$$m_{t_0} \to x_{t_1} y_{t_1} \to m_{t_0} \otimes b_{t_2} \tag{75}$$

in which $t_2 > t_1 \geq t_0$. To qualify as a genuine post-hoc bias, $b_{t_2}$ should be demonstrably caused by the event of researchers coming to known the data $y, x$. Using the equivalence noted for causal $K$, we can define an experimentally measurable parameter $\mu$ that quantifies how much the alleged bias is truly post-hoc:

$$\mu \equiv K(b_{t_2}; Y_{t_1} X_{t_0} | do(Y_{t_1} X_{t_0})) = k(b_{t_2}; Y_{t_1} X_{t_0}) \equiv \frac{H(b_{t_2}) - H(b_{t_2} | Y_{t_1} X_{t_0})}{H(b_{t_2})} \tag{76}$$

The notation was shifted from events to random variables to emphasize that we are treating the post-hoc methodology as a standard methodology, i.e. one systematically adopted in a field. We want to estimate how much knowledge would be produced *by a field whose methodology consists in* $m \otimes b$. We are taking, that is, a completely agnostic stance with respect to the validity of the methods. Our only objective is to estimate the amount of knowledge that such methods actually produce.

Knowing that $H(Y|mXb) = H(YmX) + H(b|YmX) - H(mX) - H(b|mX)$, and assuming that choices for method $b$ are made only after explanans and explanandum are known, i.e. $H(b|mX) = H(b)$, we have that $H(b|mX) - H(b|YmX) = \mu H(b_{t_2})$, and therefore:

$$K(Y_{t_1};m_{t_0}X_{t_1}\otimes b_{t_2}) = \\ = \frac{H(Y_{t_1})-H(Y_{t_1}|m_{t_0}X_{t_1})+\mu H(b_{t_2})}{H(Y_{t_1})+H(X_{t_1})+H(m_{t_0})+H(b_{t_2})} = K(Y_{t_1};m_{t_0}X_{t_1})\times\frac{C_1}{C_2}+\frac{\mu H(b_{t_2})}{C_2} \quad (77)$$

where $C_1 = H(Y_{t_1})+H(X_{t_1})+H(m_{t_0})$, and $C_2 = H(Y_{t_1})+H(X_{t_1})+H(m_{t_0})+H(b_{t_2})$. By re-arranging the first and last term of the equation we can retrospectively calculate the actual $K$ of the study, i.e. the $K$ obtained once the cost of post-hoc methodological choices is taken into account:

$$K(Y_{t_1};m_{t_0}X_{t_1}) = \frac{H(Y_{t_1})-H(Y_{t_1}|m_{t_0}X_{t_1}b_{t_2})-\mu H(b_{t_2})}{H(Y_{t_1})+H(X_{t_1})+H(m_{t_0})} \quad \text{(K with post-hoc bias)}$$

where $C=\frac{C_2}{C_1}$. Note that if it were experimentally determined that $\mu = 0$, the choices made at time $t_2$ are independent of the data, the condition would revert to those of a standard study. In all cases, the knowledge obtained is defined to be:

$$K(Y_{t_1};m_{t_0}X_{t_1}) \propto H(Y_{t_1})-H(Y_{t_1}|m_{t_0}X_{t_1}b_{t_2})-\mu H(b_{t_2}) \quad (78)$$

and therefore, unlike the unbiased $K$ encountered so far, can assume negative values. Equation 78 quantifies the actual costs of post-hoc methodological choices, which are costs in terms of information. In addition to increasing the total information cost of methodology, post-hoc methodological choices distrupt the flow of information, introducing secondary uncertainty that adds to that of the explanandum itself and thus reduces the actual knowledge obtained.

**Example**: Let $YX$ be a system and let $b$ be a post-hoc methodology in which data are removed from the sample depending on whether the values of $Y|X$ fall below a certain threshold. Methodology $b$ therefore consists of a binary choice, whose entropy will depend on how frequently data falls below the threshold: $H(b) = H(P(y|x > \tau))$. Let's assume that the methodology claims to reduce the uncertainty about the explanandum by one bit, i.e. $H(Y) - H(Y|X) = 1$. An experiment is set up to determine to what extent the methodology is post-hoc, and let's assume that the experiment establishes that the methods are fully post-hoc, i.e. $\mu = 1$. Then, the level of bias is determined by how much data is discarded using the methodology. If $P(y|x > \tau) = 0.5$ (the biased scientists discards on average half of the data), knowledge will be zero and the study can be said to be completely biased. If $P(y|x > \tau) < 0.5$ and less than half data has to be discarded, then $H(b) < 1$ and some knowledge is produced despite the post-hoc bias. If $P(y|x > \tau) > 0.5$, or if the possible choices made by the scientists are multiple, such that $H(b) > H(Y) - H(Y|Xb)$, then the study yields negative knowledge.

**Fabrication** Albeit a trivial case, for completeness we will analyse the case of data fabrication, in which methods and/or data are entirely made up. When this is the case, all the values in $K(Y;X)$ are generated, which is to say caused, by the knower itself. We have:

$$K(Y|do(Y=y);MX|do(MX=mx)) = \\ = \frac{H(Y|do(Y=y))-H(Y|do(YMX=ymx))}{H(Y|do(Y=y))+H(X|do(X=x))+H(M|do(M=m)))} = \frac{0-0}{0+0+0} = DNE \quad (79)$$

Unsurprisingly, data fabrication is just a non-starter for knowledge.

**Ante-hoc methodological choices** are just choices in the study design that favour a particular result. As in the case of post-hoc choices, it makes no difference whether ante-hoc choices are conscious or unconscious. The result is the same and it entails a lack of independence between methodology and system. If $YmX$ is a system and $K(Y;mX)$ is the knowledge claimed about the system, ante-hoc bias occurs when a "hidden" explanans $z$ produces part of the effect (or null effect). This $z$ component is of course "hiding" in the methodology $m$. Therefore, we can describe an ante-hoc biased methodology as $m : \{m^c, Z\}$, i.e. $K(Y;mX) \equiv K(Y;m^cXz)$ with $m^c \equiv m \oslash z$.

Just as we did with post-hoc choices, we can isolate the effect of this ante-hoc choice by noticing that
$H(Y) - H(Y|m^cXz) = H(Y) - H(Y|m^cX) - H(H(Y|m^cX) - H(Y|m^cXZ)$, and therefore:

$$K(Y;mX) = K(Y;m^cX)R_1 - K(Ym^cX;z)R_2 \tag{80}$$

in which $R_1 = \frac{H(Y)+H(X)+H(m^c)+H(z)}{H(Y)+H(X)+H(m^c)}$ and $R_2 = \frac{H(Y)+H(X)+H(m^c)+H(z)}{H(Y|m^cX)+H(z)}$. The ante-hoc biased study will yield actual knowledge about the system $YX$ (i.e. ignoring the effect of z) subject to the condition

$$K(Y;mX) > 0 \iff K(Ym^cX;z) < K(Y;m^cX) \times R \tag{81}$$

in which $R \equiv \frac{R_1}{R_2} \equiv \frac{H(Y|m^cX)+H(z)}{H(Y)+H(X)+H(m^c)}$ is the ratio between the uncertainty spaces of the two $K$ functions. Therefore, for a given subcomponent $z$ in $m$ that has a non-zero effect, the magnitude of the bias - the lost information - is proportional to the relative magnitude of the uncertainty spaces of the terms involved. Similarly to post-hoc biases, ante-hoc biases can lead to negative values of $K$.

Even when not quantifiable directly, the validity of methodologies can be assessed by reference to relevant literature. In particular, we can estimate the information costs of including studies with "deviant" methodologies in a literature. To estimate these costs, we introduce the two new quantitites: Oddity and Discrepancy.

**Oddity relative to aggregate** Let $m_i$ be a generic object or event, and let $M \equiv \overset{\circ}{\sum}_N m$ be an aggregate (cumulation/aggregation) of which $m_i$ is an element. We define as the the "oddity of $m_i$ with respect to M" , or more simply the "Oddity" of $m_i$, the difference between the total information of $m_i$ and that of the aggregate $M$.

$$\begin{aligned} O(m_i||M) \equiv T(m_i) - T(M) = \\ = log\frac{1}{P(m_i)} - H(M) + D(m_i||P_U(m_i)) - D(M||U) \end{aligned} \tag{82}$$

The Oddity thus defined is analogous to a Kullback-Leibler distance, and expresses the dissimilarity of a specific event or object from a set to which it pertains. This quantity is proportional to the number of bits required to pin-point (describe) that particular element of the set. As we have done throughout the essay, we will use the entropy function to represent all types of information.

We can use the oddity to evaluate the contribution made by a study. Let $\overset{\circ}{\sum}_n ymx$ be a cumulation of n studies about system $YX$, which for simplicity we assume to be a stable system. Let's assume that one study in the aggregate, which we will call study "B", uses methodology $b$ and reports a suspiciously strong result compared to all other studies. Ignoring as usual the complexity terms, the Oddity of $b$ relative to the cumulated methodologies $\overset{\circ}{\sum} m \equiv M$ is $O(m_i||M) = H(b) - H(M)$. Study $B$ reports a genuinely stronger pattern than the other studies in the literature if:

$$
\begin{aligned}
& K(YX;b) > K(YX;M) \Leftrightarrow \\
& \qquad \frac{H(Y|X) - H(Y|Xb)}{H(Y) + H(Y|X) + H(b)} > \frac{H(Y|X) - H(Y|XM)}{H(Y) + H(Y|X) + H(M)}
\end{aligned}
\tag{83}
$$

Knowing that $H(b) = H(M) + O(b||M)$, and after re-arrangement, we get:

$$H(Y|XM) - H(Y|Xb) > O(b||M)K(Y;XM) \tag{84}$$

This inequality states that, for $b$ to yield a higher than average K, the information (uncertainty of $Y$) saved by using methodology $M = b$ must be as large or larger than the oddity of its methods corrected for the overall K produced by the aggregate.

This result quantifies intuitive principles guiding the assessment of bias and flaws within a literature (e.g. [44]). First, since methods are by assumption producing positive results $K(Y;XM) > 0$, and the conditions for $b$ to yield higher than average K become restrictive when methods $M$ tield higher $K$. Moreover, the oddity $O(b||M)$ is inversely proportional to $H(M)$, and therefore inversely proportional to the size of the literature and its methodological heterogeneity (and/or average methodological complexity), and directly proportional to the relative rarity (and/or relative complexity) of the methodology. A study using unusual or unusually complex methodologies in an otherwise homogeneous literature is unlikely to actually add to the overall knowledge. Vice versa, in a small literature and/or a literature characterized by high methodological heterogeneity there is no ground to mark out a study as biased or flawed.

However, even if the literature about a specific system $YX$ is too small or too heterogeneous to mark out a methodology as exceedingly "odd", a similar, and arguably definitive, judgement can be made about the compatibility of the methodology with the *rest* of the literature.

**Discrepancy of objects or events** Let $m, b$ be two objects or events. $m, b$ are said to be discrepant if their universal probability is lower when conditioned upon each other than when considered independently of each other, i.e. $P_U(b|m) < P_U(b)$ and $P_U(m|b) < P_U(m)$. Equivalently, this implies that $T(m \otimes b) > T(m) + T(b)$ and therefore that, for the two entities to become part of one system through a process of expansion, additional information needs to be added, i.e. $T(m \otimes b) = T(m) + T(b) + D(m;b)$. This latter term, which quantifies the additional information, is the discrepancy

$$
\begin{aligned}
& D(m;b) \equiv T(m \otimes b) - T(m) - T(b) \equiv \\
& \quad \equiv H(m \otimes b) - H(m) - H(b) + D(m \otimes b||P_U(m \otimes b) - D(m||P_U(m) - D(b||P_U(b)
\end{aligned}
\tag{85}
$$

Ignoring as usual the complexity terms to slim calculations, we note that $D(m;b) = H(mb) - H(m) - H(b) = -I(m;b)$ and therefore Discrepancy, like the Oddity, is analogous to a Kullback-Leibler distance, i.e. a measure of information distance: $D(m;b) \equiv D((b) \otimes (m)||(b \otimes m)) \equiv D(b||b|m) = D(m||m|b)$. Objects or events are compatible in proportion to how negative is the value of $D(mb)$ and are incompatible in proportion to how positive it is.

Let $YX$ be a stable system, let $m$ be a methodology applied to the system, and let $b$ be a new methodology. Expanding $m$ to incorporate $b$ will increase knowledge subject to

$$K(Y;Xm{\otimes}b) > K(Y;Xm) \iff H(Y|Xm){-}H(Y|Xmb) > (H(b){+}D(m;b))K(Y;Xm) \quad (86)$$

The inequality is always satisfied when $D(m;b) = min(D(m;b)) = -H(b)$ as well as when $K(Y;Xm) = 0$, conditions in which, respectively, $b$ is fully compatible (in fact, identical) with $m$ and $m$ yields zero knowledge to begin with. If the methods are discrepant, however, additional information is needed to combine them, entailing a cost $D(m;b)$. When $D(m;b) > \frac{H(Y|Xm)-H(Y|Xmb)}{K(Y;Xm)} - H(b)$ the methodology $b$ brings negative knowledge to the system.

Acceptance or rejection of a methodology is determined by how much knowledge would be gained or lost *in total* if the literature were to be expanded to include the new system and method. Let $Z, W$ be an aggregate of systems representing the totality of phenomena currently explained by the aggregated methodologies $M$, and let $b$ be a method by which knowledge $K(Y;Xb)$ of system $YX$ is claimed. The method will be deemed unacceptable if it satisfies the conditions for knowledge growth by conslience, i.e. $K(Z;WM \otimes Y;Xb) > K(Z;WM) \otimes K(Y;Xb)$, which, after re-arrangements yields $I(Z|WM;Y|Xb) - I(Z;Y) > D(M;b)(K(Z;WM) \otimes K(Y;Xb))$ where $I()$ is the mutual information. Since scientific knowledge requires a methodology, we can posit that $I(Z;Y) = 0$ and express the general condition for knowledge growth in the presence of methodological discrepancies as:

$$H(Z|WM) + H(Y|Xb) - H(Z|MW \otimes Y|bX) > D(M;b)K(Z;WM) \otimes K(Y;Xb) \quad (87)$$

The value of $D(M;b)$ is inversely proportional to $H(M)$ and directly $H(b|m)$. Therefore, the condition 87 is less likely to be satisfied when, all else being equal, the methodology involved is more information-costly, which is to say is is more complex.

## 0.16 Pseudoscience

Etymologically, the term "pseudoscience" indicates an activity that pretends to be scientific but is not. Section 0.11 defined science as a knowledge-producing activity that has an explicit methodology. Therefore, a pseudoscience should be any activity that explicitly declares to possess a methodology $m$ that yields $K(Y;Xm) >> 0$, whilst in actually producing no knowledge, i.e. $K(Y;Xm) \leq 0$.

The phenomenon that allows discrepancy of $K$ is bias. Therefore, our thesis is that pseudo-sciences are just activities that manifest extreme forms of bias. A functional connection between bias and pseudoscience, is apparent in all typical examples of pseudoscience. Astrology, Freudian psychoanalysis, homeopathy, Intelligent Design and others (see [4]) appear to be very different activities, but share at least two characteristics: 1-they appear to produce relevant amounts of knowledge, but only of explanatory kind - they provide, in other words, high "understanding", but only of individual events; 2-their methodologies are at odds with those of established sciences. Both these conditions, it will be shown below, match the conditions we identified for bias, to levels that are extreme enough to make $K$ equal or lower than zero.

**Extreme post-hoc bias or methodological complexity** Given a specific object or event, a pseudoscience will offer a seeming valid understanding, such that $K(y;xm) >> 0$. However, knowledge presupposes the subsistence of a pattern such that $K(\overset{\circ}{\sum} y;xm) > 0$. If the methodology of a pseudoscience is kept constant across cumulated studies, then $K(\overset{\circ}{\sum} y;xm) = 0$, because by assumption the pattern does not

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v1 | CC-BY 4.0 Open Access | rec: 19 Apr 2016, publ: 19 Apr 2016

actually subsist. If, on the other hand, a repertoire of multiple methodologies (or equivalently of multiple explanations) is accessible to the pseudoscientist, such that $H(\overset{\circ}{\sum} m) = H(M)$, then $H(\overset{\circ}{\sum} y|xm) = \overset{\circ}{\sum} H(Y|Xm) + H(M)$. Leading to the condition:

$$K(Y;XM) \leq 0 \iff H(M) \geq H(Y) - \overset{\circ}{\sum} H(Y|Xm) \tag{88}$$

This is a generalization of Popperian falsificationism. The paradigmatic unfalsifiable pseudoscience *sensu* Popper corresponds to one in which explanations are seemingly perfect, i.e. $H(Y|XM) \approx 0$ and one explanation is available for every possible outcome $H(M) = H(Y)$ (the condition already encountered in section 0.11). However, note that equation 88 allows for a broader variety of scenarios. First, it can accommodate post-hoc methodological bias as a source of unfalsifiability. Second, it allows explanations to be imperfect (with $\overset{\circ}{\sum} H(Y|Xm) > 0$) and it allows overlap between the understanding of different explanations (since $H(\overset{\circ}{\sum} y|xm) = \overset{\circ}{\sum} H(Y|Xm) + \overset{\circ}{\sum} D(Y|Xm||\overset{\circ}{\sum} Y|Xm)$. Third, it allows for scenarios in which the uncertainty costs of the (pseudo-)methodology are higher than the uncertainty of the explanandum, i.e. $H(M) > H(Y) - \overset{\circ}{\sum} H(Y|Xm)$. Fourth, it embodies the equivalence between $H(\overset{\circ}{\sum} m)$, $H(M)$, $H(\overset{\circ}{\sum} M)$, which unifies scenarios in which a pseudoscience has multiple alternative methodologies or only one but sufficiently complex as to cover any possible event in the explanandum.

**Extreme ante-hoc bias** The definition proposed here might identify as pseudoscience fields that traditional falsificationist (or verificationist) approaches might exclude. In particular, our definition includes cases of extreme ante-hoc bias, in which the pattern that is believed to underlie the knowledge claim is in fact a pseudo-pattern, produced by the methods. It is easy to quantify this condition from equation 89. A pseudoscience yields $K(Y;mX) \leq 0$, and therefore has:

$$K(Ym^cX;z) \geq K(Y;m^cX) \times R \tag{89}$$

with all terms defined as before. This condition would correspond to a research field that looks entirely legitimate and that makes predictions that are successfully tested but yields negative knowledge nonetheless.

**Extreme methodological discrepancy** Consilience is the ultimate form of scientific progress, and discrepancy ( 87) is the ultimate criterion by which methodologies can be judged. When discrepancy is extreme, a methodology $b$ is fully incompatible with accepted methodology $m$, implying $P(b \otimes m) = 0$ and therefore $D(b;m) = \infty$ and $K(Y;Xm \otimes b) = 0$ for any system $YX$.

Let $ZW$ be a second system for which non-zero knowledge is obtained using methodology $m$, i.e. $K(Z;Wm) > 0$. To simplify the analysis, we will assume that $Y$ and $Z$ are completely independent phenomena, as are their respective explananda $X$ and $W$ such that $H(Y \otimes Z) = H(Y) + H(Z), H(X \otimes W) = H(X) + H(W)$. We will also assume that there are no cross-effects, $H(Y|W) = H(Y), H(Z|X) = H(Z)$ independent of the methods used. Due to the incompatibility of $m$ and $b$, the knower should choose method $b$ over $m$ if:

$$K(Y;Xb \otimes Z;Wb) > K(Y;Xm \otimes Z;Wm) \tag{90}$$

Which, under the assumptions made above, after a few re-arrangements, yields:

$$\frac{H(Y)\times k(Y;Xb)+H(Z)\times k(Z;Wb)}{H(Y)\times k(Y;Xm)+H(Z)\times k(Z;Wm)} > \frac{H(b)+C}{H(m)+C} \tag{91}$$

with $C = H(Y) + H(Z) + H(X) + H(W)$. The knower should reject methodology $m$ in favour of $b$ if the latter yielded a relatively high knowledge $k$ over a broader range of phenomena $H(Y)$ at a relatively small cost in added complexity $H(b)$. Phenomena that are typically defined as pseudosciences do not meet any of these conditions: their methods are overtly complex and only (appear to) explain a narrow range of phenomena.

The three forms of bias are not mutually exclusive, of course, and in a pseudoscience are likely to co-exist and reinforce each other. In particular, the presence of post- and ante-hoc biases can make the left-hand side of inequality 91 be equal or lower than zero, which makes the condition impossible to satisfy. An activity that was genuinely aimed at increasing knowledge would reject the flawed methodology or at least attempt to increase the value of $K(Y;Xb)$ and reduce that of $D(b;M)$. One that is pseudoscientific, however, does not. By doing so, a pseudoscience hampers its own progress, since:

$$\lim_{n\to\infty} K(\overset{\circ}{\sum}_n Y;Xb) \le 0 \quad and \quad \lim_{n\to\infty} K(\overset{\circ}{\prod}_n Y;Xb) \le 0 \tag{92}$$

## 0.17 Hierarchy of the Sciences (and Pseudosciences)

The great polymath and philosopher Auguste Comte (1798-1857) first proposed that the diversity of the sciences could best be understood as a progression. From astronomy to sociology, disciplines seemed to form an ordered hierarchy, along which subject matters became more complex, less general and less amenable to mathematization. This order also seemed to reflect sciences' historical recency, decreasing speed of progress, increasing susceptibility to biases and increasing relevance to human affairs. Sociology was considered by Comte to be the most important and yet the least developed of all sciences.

Comte's vision was certainly simplistic and incorrect in many details, but not fundamentally wrong. A quantitatively complete and generalized account of the diversity of the sciences is obtained by merging all functions and quantities developed in this essay in a single quantity which we indicate with the capital Greek letter $\Xi$:

$$\Xi = \frac{ykh}{\Lambda} \tag{93}$$

in which $y = T(y) = n \times (H(Y) + D(Y||Y_U))$ is the total information of the explanandum (section 0.3)), $k = \frac{T(Y)-T(Y|MX)}{T(Y)}$ is the *bias-corrected* $k$ (sections 0.4, 0.15), $h = \frac{1}{1+\frac{T(MX)}{T(Y)}}$ is the "hardness" factor (section 0.14)), and $\Lambda = \frac{1}{t^{\dagger}} log \frac{K(Y;MX)}{\epsilon}$ is the knowledge loss rate (section 0.8). Since $y \in (0, +\infty)$, $k \in (-1, 1)$, $h \in (0, 1)$ and $\Lambda \in [0, +\infty)$, the quantity $\Xi$ ranges between $(-\infty, +\infty)$.

$\Xi$ quantifies pseudosciences and sciences along a gradient of softness measured on a universal scale (Figure 12). We could also write $\Xi = T(y) \times K(Y_t; MX)$, but this would obscure the fact that $\Xi$ depends on four relatively independent quantities, which for all means and purposes can be analysed separately. $\Xi$ can be imagined as a volume in four dimensions, which corresponds the amount of total knowledge yelded by a field about an explanandum.

In reality, the dimensions of $\Xi$ are five, because $y = n \times E[C(y)]$, in which the $n$ term quantifies the frequency of encounter of the explanandum. This term introduces a subjective element of value-judgement in the analysis of knowledge. Different

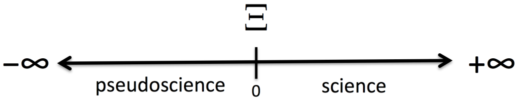

**Figure 12.** Visual representation of Ξ and how it gives mathematical solution to the demarcation problem.

individuals are likely to encounter phenomenon $y$ with different frequencies in their lives, and will therefore value $K(Y; mX)$ differently. For example, since most people have to handle other human beings more frequently than atoms, most people might ascribe a larger $n$ to psychology than to quantum physics - although physicists will give the latter a relatively higher value. Alternatively, the $n$ term might be taken to represent the relative abundance of $y$ in the universe. In this case, since there are more atoms than there are people, knowledge about the former would be valued much more. Philosophers and meta-researchers, when studying knowledge itself, might assume $n = 1$.

# Discussion

## 0.18 Summary of findings

From two simple postulates - information is finite and knowledge is information compression - we have derived a consistent mathematical theory that unifies and quantifies fundamental epistemological concepts including simplicity, accuracy, chaos, science, soft science, bias and pseudoscience.

We started by defining a $K$ function (section 0.3), which quantifies knowledge as a standardized, accuracy-dependent and time-dependent version of Shannon's mutual information function. The $K$ function quantifies how much uncertainty about an explanandum is reduced by an explanans. Explanans and explanandum may consists in sources of information of any kind: variables, single events or objects. The information that in a random variable is measured by Shannon's entropy, in an object is quantified by Kolmogorov complexity. The equivalence between information, uncertainty and complexity was made explicit by introducing the notion of "Total Information" (section 0.3).

The $K$ function lends itself to a Bayesian as well as a frequentist interpretation (section 0.4), but differs from typical statistical quantities in at least two ways. First, being based on Shannon's entropy, it computes quantities that are logarithmic and accuracy-dependent (Fig 3). Second, it incorporates the information costs entailed in the explanans, which typical statistical measures of effect size ignore. The information costs of the explanans can only be excluded in particular circumstances, i.e. when the explanans is assumed to be perfectly knowable, a condition quantified by a variant of the $K$ function defined as "uncorrected K" and symbolized with a lower-case "k".

Explanation and causation are expressions of knowledge that find a direct quantification in the $K$ function (section 0.5). Theoretical models are also shown to be explanantia like any other: theories are just devices which structure a phenomenon and thus reduce uncertainty about it. Like all other explanantia, theories must contain non-zero, finite amounts of information.

Section 0.6 introduced two fundamental operations through which information is processed and knowledge evolves. The $\otimes$ operation expands information by adding dimensions, thus multiplying the attributes that are objects of knowledge. The $\oplus$ operation cumulates information over individual dimensions, updating and summarizing knowledge about an attribute. A number of useful mathematical properties and theorems are derived in section 0.7.

The validity of $K$ as a measure of knowledge was supported in section 0.8, by showing that K embodies three properties that knowledge is expected to possess. The first property is Occam's razor, which turns out to be implicit in our mathematization of knowledge. The second property is $K$'s dependence upon accuracy, which is defined by the quantization of explanans and explanandum. We defined measurement error as the defining property of empirical knowledge, and it was shown that $K$ possesses a single optimal level of accuracy. The third property is $K$'s decline over time, measured as a distance between explanans and explanandum. This decline only occurs for empirical systems, and implies that knowledge of all empirical systems encounters a "chaos horizon", beyond which $K \approx 0$. Logico-deductive systems are different from empirical systems solely in not having measurement error and therefore no chaos horizon.

Ordinary knowledge was described in section 0.9 by distinguishing three essential processes. The first one is experience, in which knowledge (encoded patters) is applied to a single event or object, yielding "understanding" which is quantified by $K$. The second is knowledge proper, which results from the cumulation of experiences about a system, allowing the reinforcing or weakening of patterns. Cumulation can be recursive, and knowledge can therefore be aggregated in ever more complex structures. The third component is creativity, in which encoded patterns are tentatively expanded to new domains, generating the potential for new knowledge. We quantified the knowledge acquired per experience and the conditions that favour the cumulation and expansion of knowledge. These conditions might correspond to evolutionary pressures that lead to increasing cognitive complexity as well as abstract and symbolic thinking.

Science (section 0.11) operates exactly as ordinary knowledge, but is made more potent by a distinctive characteristic: it makes explicit a "methodology", i.e. an algorithm that underlies a knowledge claim (a pattern). Thus, a distinctive "methodology" component $m$ flanks the explanans $x$ in the $K$ function for science. Scientific methodology effectively operates like a statistical conditioning factor: it narrows down the universe of possible studies to that corresponding to a specific knowledge claim. The amount of information needed to describe a methodology reflects the specificity of the conditioning, i.e. extent to which the universe is narrowed down (Fig 7).

In all other respects, scientific knowledge is in one-to-one correspondence with ordinary knowledge. Scientific studies corresponds to experiences; a literature is a cumulation of experiences yielding knowledge; hypotheses are expression of creativity and innovation. Scientific progress occurs at two levels. Within scientific fields progress occurs when hypotheses are conclusively verified or falsified. Across fields, scientific progress is determined by the pace of expansions and abstractions, which moves towards increasing consilience, i.e. maximum information compression with minimum methodology.

Contrary to common verbal falsificationist arguments, we found that verification yields more information than falsification. Karl Popper's account of falsification represents a special case of symmetry between explanans and explanandum, unlikely to be realized in most real sciences 0.11. Similarly, we found that research reproducibility is a highly idealized concept (section 0.13). Virtually no empirical study can be 100% reproducible, because no methodology is likely to be completely described. The success rate of reproducibility studies will depend on multiple factors, including a methodology's complexity and amount of missing information.

Sciences that are typically considered to be "soft" were straightforwardly identified as those which encode relatively weak patterns ( 0.14). The weakness is due to high complexity (uncertainty) of explanans and/or methodology, relative to explanandum. Low consensus, believed to be characteristic of soft sciences, is manifest in the cumulation of multiple non-overlapping methodologies, or equivalently a single highly complex methodology. Therefore, scientific softness is directly quantified by the K function, which can be partitioned into an ”uncorrected k” component, which measures effect sizes, and a ”hardness” factor h, which measures the information costs caused by the complexity of explanations and methodologies. Intrinsic properties of subject matter are the prime determinants of scientific softness, but choices under control of scientists can modulate $h$ and $k$ and thus maximize knowledge in all fields.

The concept of bias found a new interpretation in light of this theory (section 0.15). Irrespective of whether it is conscious or unconscious, intentional or unintentional, benevolent or malevolent, what empirical scientists call ”biases” are methodological choices that, either ante-hoc or post-hoc, subtract information from the knowledge claim, making $K$ lower than it appears. Post-hoc and ante-hoc biases can be quantified experimentally. The ultimate validity of a methodology is determined in relation to a literature, and to this end we proposed measures of methodological Oddity and Discrepancy.

Methodologies that are typically defined as ”pseudoscientific” turn out to be just extreme manifestations of bias. In particular, these are methodologies that subtract more information than they produce, yielding non-positive K. Classic demarcation criteria, in particular Popper’s falsificationism and Lakatosh’s degeneracy, are covered by this definition, which might cover a broader class of possible pseudosciences, the status of which can be quantified experimentally.

Finally, we proposed a new, quantitative and more general version of the hierarchy of the sciences. The status of an activity as hard science, soft science, or pseudoscience is measured on a universal quantity Ξ, which subsumes all key parameters and concepts proposed in this essay (Fig 12).

## 0.19 Predictions and conclusions

This essay proposed a theory, not a model. The function $K(Y;X)$ was not derived to emulate knowledge, but to quantify what knowledge actually *is*. The correspondence between $K$ and knowledge was supported by finding that properties typically ascribed to knowledge are intrinsic properties of the function. However, the validity of the claim that $K$ is knowledge ultimately rests on postulating that knowledge, in any of its forms, consists in pattern encoding. To the best of the author’s knowledge, there is no evidence that contradicts the postulate. Any counter-example or counter-argument would significantly undermine the generality of the theory. The generality of the theory also depends on the generality of the first postulate. If information cannot be assumed to be finite, even in the context of measurement, then the theory would need to be modified to accommodate infinities, which might create insurmountable contradictions.

Although likely to fall on the ”soft” side of the Ξ spectrum, this theory makes unique and testable predictions. In particular, it makes the overarching prediction that the uncertainty space of a field (see 0.3) is inversely proportional to the field’s speed of progress and directly proportional to the field’s exposure to bias. A connection between complexity of subject matter and progress or bias has been suggested in the past to explain broad differences between scientific domains. However, in addition to offering accurate quantitative versions of these pre-existing preditions, the theory makes the new prediction that these relations should hold within any discipline, at the level of individual fields. Even within a highly structured discipline such as mathematics the theory predicts the existence of slow-progressing, ”softer” fields. Conversely, it predicts the presence of relatively hard and fast-progressing fields in the social sciences.

Novel predictions are also made with respect to reproducibility, the probability of which appears difficult to explain [45]. The theory predicts that no field can exhibit 100% reproducibility, and that the reproducibility rate should be inversely proportional to the cumulative uncertainty space. Post-hoc and publication bias might further undermine a study's reproducibility, but reproducibility failures should occur independent of the study's publication status. Moreover, all else being equal, the reproducibility of a study should be proportional to the Kolmogorov complexity of the description of the study's methodology and, controlling for this factor, to the level of expertise of the the replicating scientists and the length of the description of the methodology available to them.

A further overarching prediction is that a field's uncertainty space should exert the same effects irrespective of the relative size of its components (explanans, explanandum and methodology). For example, a field that tests simple explanations of complex phenomena should progress at a similar speed to a field that tests complex explanations of simple phenomena. Methodological complexity, moreover, should have similar effects when dispersed in multiple alternative methodologies or a single one of equivalent complexity. The equivalence between information, uncertainty and complexity is perhaps the most striking prediction of the theory. Unfortunately, it is also the most difficult to test, because Kolmogorov complexity is not computable in principle and hard to estimate even in practice. Current compression algorithms approximate measures of complexity, but their accuracy is inversely proportional to the size of the object to compress. The relative simplicity of mathematical objects suggests that mathematics, physics and other math-intensive fields might be the most suitable domains to test the theory.

If correct, the theory offers consistent and unified explanations for a variety of phenomena noticed about the sciences. For example, it explains why the social sciences lack methodological consensus and why they do not make progress like the physical sciences, a problem that had entire books dedicated to it [46]. It also explains why and how pseudosciences are formed and maintained, and it connects the emergence of pseudosciences to well-established scientific theories, including measurement theory and complexity science. Bias, soft methodology, pseudoscience are proposed to be measurable concepts, that can be tracked in the literature, quantified experimentally and intervened upon to foster progress in all areas of knowledge.

The quantitative approach proposed might also inform policies and interventions to improve research, by offering insights into the specific heuristics of each field. For example, we found that, whilst negative results are informative when testing a defined hypothesis, the conclusive falsification of a hypothesis is generally not as valuable as a verification (see 0.11). This suggests that when theoretical consensus is low the publication of exploratory negative results could be of very little value. Future, more complete analyses should examine how the benefits of study registration, negative results repositories and other strategies to counter bias vary by field and should be tailored accordingly. Reproducibility was also shown to be field-dependent, in ways that can be quantified and predicted (section 0.13).

In addition to explaining phenomena about science, this theory might help to understand general cognitive processes. For example, it explains why public debates on highly complex issues of great political and social importance - e.g. climate change, or the effects of GMO - tend to polarize around two extremely simplistic positions (e.g. is climate change man-made or not? Are GMOs good or bad for health?). Extreme simplification is a strategy to maximize K when information is limited. Other cognitive phenomena, including art, humour intelligence and genius could be quantified and analysed.

In conclusion, the theory offers a scientific approach to epistemology that avoids naïve reductionism and positivism. It allows knowledge to make progress in all domains and all subject matters - physical, biological, behavioural and cultural - but also allows it to be shaped by contingent, psychological and sociological choices. It recognizes and rationalizes epistemological pluralism, but reconciles it with a unitary view of the scientific enterprise. The unity of the sciences reflects the unity of knowledge, a phenomenon that, albeit erratic, can be measured and explained by two simple postulates.

## Supporting Information

**Knowledge progress curve** Let a field be defined by an explanandum $Y$, and let $\Omega : \{X_1, X_2 ... X_n\}$ be a set of candidate explanantia. Our aim is to build a curve that expresses $K(Y;X)$ as an optimized function of $H(\overset{\circ}{\prod} X)$.

To do so, we can imagine a process in which the field (a collection of studies all addressing a specific system $YX$ verifies/falsifies all candidate explanantia $X \in \Omega$ individually, and identifies the one, say $X_i$, that maximizes $K$, i.e. minimizes the value $H(Y|X)$. The knower then proceeds to test all pairs $X_i, X_j$ to again identify the $X_j$ that maximizes K, and so on. Iterations of this process would produce a series of pairs $\{(H(X_1); H(Y|X_1)), (H(X_2); H(Y|X_2)), ...(H(X_n); H(Y|X_n))\}$ which could be plotted. If fitted by a curve, these points would yield a curve that is non-increasing, convex on the average, and that reaches a minimal value $H(Y|X^*)$ corresponding to the optimal explanation for the set $\Omega$. Having exhausted a set of candidate explanantia, we can imagine the field expanding the original $\Omega$ to a new set of explanantia $\Omega'$ and trying out all combinations of the optimal explanation with each new candidate explanantia. Indeed, we could imagine the field re-starting the process, in order to find an optimal new explanans $\underline{X}_{n'} \in \Omega \otimes \Omega'$ which by definition would reach a new minimal value of $H(Y|X^{*'}) < H(Y|X^*)$.

The process of described above could in principle be applied to any field (any explanandum $Y$), and would yield a curve with similar characteristics: it would be rapidly a declining curve that approaches asymptotically the value $H(Y|X) = 0$. If the field had access to all possible candidate explanantia and built a maximally optimal curve, such curve would correspond to an expontential function. Any explanandum $Y$ and a set of candidate explanantia $\Omega$ can therefore be characterized by a unique function in the form $H(Y|X) \sim H(Y) * e^{(-\gamma(-H(X_0)+H(X))}$, with a negative intercept term $-H(X_0)$ that allows the curve the start at any value of the explanans.

Plugging that function back into the $K$ function, we get a characteristic curve for each combination of $Y$ and $\Omega$ in the form:

$$K(Y;X) \sim \frac{H(Y)(1 - e^{-\gamma(H(X)-H(X_0))})}{(H(Y)+H(X))e^{\lambda t}} \tag{94}$$

This curve is idealized as it fits a simple exponential curve that reaches asymptotically its maximal value. The model could be made more realistic by including a suboptimal asymptote to which the value $H(Y|X)$ tends and could accommodate irregularities in the curves by using higher order polynomials, but these details are removed to simplify notation at no cost for the analysis. Independent of the complexity of the polynomial, the growth of K will be determined by the parameter $\gamma$, which represents the average effect size and will be referred to as the knowledge gain rate. All knowledge gain curves tend asymptotically to the maximum value of K that can be achieved given a specific explanandum $Y$. If the explanandum, and therefore $H(Y)$, is constant, then $K$ is a geometrically decreasing function of $H(X)$. It follows that every knowledge gain curve will, if $H(Y)$ is kept constant, initially increase, reach a maximized value, and subsequently decrease approaching asymptotically 0 (Fig 11).

## References

1. Ioannidis JPA, Fanelli D, Dunne DD, Goodman SN. Meta-research: Evaluation and Improvement of Research Methods and Practices. PLoS Biol. 2015 10;13(10):1–7.
2. Laudan L. The Demise of the Demarcation Problem. In: Grünbaum A, Cohen RS, Laudan L, editors. Physics, Philosophy and Psychoanalysis: Essays in Honour of A. Grünbaum. Boston Studies in the Philosophy of Science. Springer; 1983. .
3. Dupre JA. The Disorder of Things. Metaphysical foundations of the disunity of science. Harvard University Press; 1993.
4. Pigliucci M. The Demarcation Problem: A (Belated) Response to Laudan. In: Pigliucci M Massimo e Boudry, editor. Philosophy of Pseudoscience: Reconsidering the Demarcation Problem. University of Chicago Press; 2013. .
5. Comte A. Cours de philosophie positive. vol. 6 vols. Paris: Rouen first, then Bachelier; 1830-1842.
6. Pearson K. The grammar of science. 2nd ed. London: A. and C. Black; 1900.
7. Poincaré H, Maitland F. Science and method. London: T. Nelson and sons; 1914.
8. Wittgenstein L. Tractatus logico-philosophicus. New York: Harcourt, Brace & company, inc.; 1922.
9. Popper KR, Popper KR. The Logic of Scientific Discovery. Harper Torchbooks. HarperCollins Canada, Limited; 1959.
10. Lakatos I. Falsification and the Methodology of Research program. In: Criticism and the Growth of Knowledge. Cambridge: Cambridge University Press.; 1970. p. 91–97.
11. Merton RK. The Normative Structure of Science. In: he Sociology of Science: Theoretical and Empirical Investigations. Chicago: University of Chicago Press; 1942(1973). .
12. Kuhn TS. The structure of scientific revolutions. 2nd ed. Chicago: The University of chicago Press; 1970.
13. Fuller S. Dissent Over Descent: Intelligent Design's Challenge to Darwinism. Icon; 2008.
14. Whewell W. The philosophy of the inductive sciences: founded upon their history. London: J.W. Parker; 1840.
15. Windelband W. History and Natural Science. Theory & Psychology. 1894 (1998);8(1):5–22.
16. Russell B. Our knowledge of the external world as a field for scientific method in philosophy. Chicago: The Open Court Publishing Co.; 1914.
17. Conant JB. Science and common sense. New Haven: Yale University Press; 1951.
18. Storer NW. Hard sciences and soft - Some sociological observations. Bulletin of the Medical Library Association. 1967;55(1):75–&.

19. Bunge M. The maturation of science. In: Lakatos I, Musgrave A, editors. Problems in the Philosphy of Science. vol. 3. Amsterdam: Norh-Holland Publishing Company; 1967. .

20. de Solla Price DJ. 1. In: Citation measures of hard science, soft science, technology, and nonscience. Lexington, MA: Heath Lexington Books, D.C. Heath and Company; 1970. p. 3–22.

21. Zuckerman HA, Merton RK. Age, aging, and age structure in science. In: Storer N, editor. The Sociology of Science, by R. K. Merton. Chicago: University of Chicago Press; 1973. p. 497–559.

22. Cole S. The hierarchy of the sciences? American Journal of Sociology. 1983;89(1):111–139.

23. Humphreys P. A conjecture concerning the ranking of the sciences. Topoi-an International Review of Philosophy. 1990;9(2):157–160.

24. Braxton JM HL. Variation among academic disciplines: Analytical frameworks and research. In: Higher education: handbook of theory and research. New York: Agathon Press.; 1996. .

25. Simonton DK. Scientific status of disciplines, individuals, and ideas: Empirical analyses of the potential impact of theory. Review of General Psychology. 2006;10(2):98–112.

26. Fanelli D. 'Positive' Results Increase Down the Hierarchy of the Sciences. PLoS ONE. 2010 04;5(4):1–10.

27. Fanelli D, Gl‰nzel W. Bibliometric Evidence for a Hierarchy of the Sciences. PLoS ONE. 2013 06;8(6):1–11.

28. Lloyd S. Computational Capacity of the Universe. Phys Rev Lett. 2002 May;88:237901.

29. Hand DJ. Measurement Theory and Practice: The World Through Quantification. Wiley; 2004.

30. March E. The economical nature of physical inquiry. In: Popular Scientific Lectures by Ernst Mach [1895]. The Open Court Publishing Co.; 1882. .

31. Brillouin L. Science and Information Theory: Second Edition. Dover Publications; 1962.

32. Popper KR. The Logic of Scientific Discovery. Classics Series. Routledge; 2002.

33. Nola R, Sankey H. Theories of Scientific Method: an Introduction. Taylor & Francis; 2007.

34. McAllister J. Algorithmic randomness in empirical data. Studies in History and Philosophy of Science. 2003;34:633?646.

35. Michalowicz JV, Nichols JM, Bucholtz F. Handbook of Differential Entropy. Taylor & Francis; 2013.

36. Cover TM, Thomas JA. Elements of Information Theory. Wiley; 2012.

37. Li M, Vitányi P. An Introduction to Kolmogorov Complexity and Its Applications. Texts in Computer Science. Springer New York; 2009.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v1 | CC-BY 4.0 Open Access | rec: 19 Apr 2016, publ: 19 Apr 2016

38. Losee J. Theories of Causality: From Antiquity to the Present. Transaction Publishers; 2012.

39. Pearl J. Causality. Cambridge University Press; 2009.

40. Grimes DA, Schulz KF. Bias and causal associations in observational research. The Lancet. 2016/04/13;359(9302):248–252.

41. Vandenbroucke JP, Broadbent A, Pearce N. Causality and causal inference in epidemiology: the need for a pluralistic approach. International Journal of Epidemiology. 2016;.

42. Bertuglia CS, Vaio F. Nonlinearity, Chaos, and Complexity:The Dynamics of Natural and Social Systems: The Dynamics of Natural and Social Systems. OUP Oxford; 2005.

43. Kautz R. Chaos: The Science of Predictable Random Motion. OUP Oxford; 2011.

44. Ioannidis JP. Why most published research findings are false. PLoS medicine. 2005;2(8):e124.

45. Collaboration OS. Estimating the reproducibility of psychological science. Science. 2015;349(6251).

46. Cole S. Why sociology doesn't make progress like the natural sciences. In: What's wrong with sociology? Transaction Publishers; 2001. .

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v1 | CC-BY 4.0 Open Access | rec: 19 Apr 2016, publ: 19 Apr 2016