A theory and methodology to quantify knowledge

Daniele Fanelli[1]

1 Department of Methodology, London School of Economics and Political Science, London, United Kingdom

* email@danielefanelli.com

1 Abstract

This article proposes quantitative answers to meta-scientific questions including "how much knowledge is attained by a research field?", "how rapidly is a field making progress?", "what is the expected reproducibility of a result?", "how much knowledge is lost from scientific bias and misconduct?" "what do we mean by soft science?", "what demarcates a pseudoscience?".

Knowledge is suggested to be a system-specific property measured by K, a quantity determined by how much the information contained in an *explanandum* is compressed by an *explanans*, which is composed of an information "input" and a "theory/methodology" conditioning factor. The three arguments of the K function are quantifiable using methods of classic and algorithmic information theory. This approach is justified on three grounds: 1) K results from postulating that information is finite and knowledge is information compression; 2) K is compatible and convertible to ordinary measures of effect size and algorithmic complexity; 3) K is physically interpretable as a measure of entropic efficiency. Moreover, the K function has useful properties that support its potential as a measure of knowledge.

Examples from a variety of fields are given to illustrate the possible uses of K. These examples include quantifying: the knowledge value of proving Fermat's last theorem; the accuracy of measurements of the electron's mass; the half life of eclipse predictions; the usefulness of evolutionary models of reproductive skew; the significance of gender differences in personality; the sources of irreproducibility in psychology; the impact of scientific misconduct and QRP; the knowledge value of astrology. Furthermore, a cumulative K may complement ordinary meta-analysis and may give rise to a universal classification of sciences and pseudosciences.

Simple and memorable mathematical formulae summarize the theory's key results and implications. In addition to practical uses in meta-research, these formulae may have conceptual applications in philosophy and in research policy, and may guide scientists to make progress on all frontiers of knowledge.

2 Introduction

A science of science is flourishing in all disciplines and promises to boost discovery on all research fronts [1]. Commonly branded "meta-science" or "meta-research", this growing literature of empirical studies, experiments, interventions, and theoretical models explicitly aims to take a "bird's eye view" of science and a decidedly cross-disciplinary approach to studying the scientific method, which is dissected and experimented upon as any other topic of academic inquiry. To bloom into a fruitful research field, however

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

meta-research needs a cross-disciplinary, quantitative, and operationalizable theory of scientific knowledge that can tell apart "good" from "bad" science.

This article proposes such a meta-scientific theory. By analytical methods and practical examples, it suggests that a system-specific quantity named "K" can help answer meta-scientific questions including "how much knowledge is attained by a research field?", "how rapidly is a field making progress?", "what is the expected reproducibility of a result?", "how much knowledge is lost from scientific bias and misconduct?" "what do we mean by soft science?", "what demarcates a pseudoscience?".

The growing success and importance of meta-research make the need for a meta-theory ever more salient and pressing. Growing resources are invested, for example, in ensuring reproducibility [1], but there is little agreement on how reproducibility ought to be predicted, measured and understood (see [2,3]). Graduate students are trained in courses to avoid scientific misconduct and questionable research practices, and yet the definition, prevalence and impact of questionable behaviours across science are far from clear [4]. Increasing efforts are devoted to measure and counter well-known problems such as publication bias, even though inconclusive empirical evidence [5] and past failures of similar initiatives (e.g. the withering and closure of most journals of negative results [6]) suggest that these problems are incompletely understood.

Theoretical models currently used in meta-research are derived from very specific fields, which may not be exportable to other contexts. The most prominent example is offered by the famous claim that most published research findings are false [7]. This land-mark analysis has deservedly inspired meta-studies in all disciplines. However, its predictions are based on an extrapolation of statistical techniques used in genetic epidemiology that have several limiting assumptions. These assumptions include: that all findings are generated by stable underlying phenomena, independently of one another, with no information on their individual plausibility or posterior odds, and with low prior odds of any one effect being true. These assumptions are unlikely to be fully met even within genetic studies [8], and the extent to which they apply to any given research field remains to be determined.

Similar limiting assumptions are manifest in meta-research methodologies. Reproducibility and bias, for example, are measured using meta-analytical techniques that treat sources of variation between studies as either fixed or random [9,10]. This assumption may be valid when aggregating results of randomized control trials [11], but may be inadequate when comparing results of fields that use varying and evolving methods (e.g. ecology [12]) and that study complex systems subject to non-random variation (e.g. reaction norms [13]).

Statistical models can be used to explore the effects of different theoretical assumptions (e.g. [14–17]) as well as other conditions that are believed to conduce to bias and irreproducibility (e.g. [18,19]). However, the knowledge produced by such models will not amount to a theory. A genuine "theory of meta-science" ought to offer a general framework that, from maximally simple and universal assumptions, explains, predicts and classifies knowledge phenomena across all research fields.

This article builds a theoretical and methodological framework for meta-research using notions of classic and algorithmic information theory. The key innovation introduced is a function of three variables that, it will be argued, quantifies the essential phenomenology of knowledge, scientific or otherwise. This approach builds upon a long history of advances made in combining epistemology and information theory. The concept that scientific knowledge consists in pattern encoding can be traced at least to the polymath and father of positive philosophy August Comte (1798-1857) [20], and the connection between knowledge and information compression *ante litteram* to the writings of Ernst Mach (1838-1916) and his concept of "economy of thought" [21]. Claude Shannon's theory of communication gave academics a mathematical language to

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

quantify information [22], whose applications to physical science were soon examined by Léon Brillouin (1889-1969) [23]. The independent works of Solomonoff, Kolmogorov and Chaitin gave rise to algorithmic information theory, which dispenses of the notion of probability in favour of that of complexity and compressibility of strings (see [24]). The notion of learning as information compression of data was formalized in Rhissanen's Minimum Description Length principle [25], which has found fruitful and expanding applications in statistical inference and machine learning [26, 27]. From a philosophical perspective, the relation between knowledge and information was explored by Fred Dretske [28], and a computational philosophy of science was elaborated by Paul Thagard [29]. To the best of the author's knowledge, however, the formulae and ideas presented in this article were never proposed before (see Discussion for further details).

The rest of this article is organized as follows. In section 3, the core mathematical approach is presented. This verges on a single equation, the K function, whose validity is justified in section 3.2 based on theoretical, statistical and physical arguments. Section 3.3 explains and discusses properties of the K function. These properties further support the role of K as a universal quantifier of knowledge, laying out the ground for developing a methodology. The methodology is illustrated in section 4, which shows how the theory may help to answer typical meta-research questions. These questions include: how to quantify theoretical and empirical knowledge (sections 4.1 and 4.2, respectively), how to quantify scientific progress within or across fields (section 4.3), how to forecast reproducibility (section 4.4), how to estimate the knowledge value of null and negative results (section 4.5), how to compare the knowledge costs of bias, misconduct and QRP (section 4.6), and how to define a "soft" science (section 4.8) and a pseudoscience (section 4.7). These results are summarized in a set of memorable formulae and discussed in section 5. These sections cross-reference each other but can be read in any order with little loss of comprehensibility.

3 Analysis

3.1 The quantity of knowledge

The core claim of this article is that knowledge is a system-specific property measured by a quantity that we will indicate with the letter K and is given by the function

$$K(Y^{n_Y}; X^{n_X}, \tau) \equiv \frac{n_y H(Y) - n_y H(Y|X, \tau)}{n_y H(Y) + n_x H(X) - \log p(\tau)} \tag{1}$$

in which $H(\cdot)$ is Shannon's entropy function, and $p(\cdot)$ is a uniform probability distribution whose characteristics are explained further below. The three terms in equation 1 are defined as follows:

- **Y** constitutes the *explanandum*, latin for "what is to the explained. Examples of explananda include: response variables in regression analysis, physical properties to be measured, experimental outcomes, unknown answers to questions.
- **X** and τ together constitute the *explanans*, latin for "what does the explaining". In particular
  - **X** will be referred to as the "input", and it will represent information acquired externally. Examples of inputs include: results of any measurement, explanatory variables in regression analysis, physical constants, arbitrary methodological decisions and all other factors that are not "rigidly" encoded in the theory or methodology.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

– τ will be referred to as the "theory" or "methodology". A typical τ is likely to contain both a description of the structure relating Y and X, as well as a specification of all other conditions that allow the relationship between X and Y to be expressed. Examples of τ include: an algorithm to reproduce Y, a description of a physical law relating Y to X, a description of the methodology of a study (i.e. description of the population characteristics, the interventions, how measurements were made, etc.).

Specific examples of all of these terms will be offered repeatedly throughout the essay. Mathematically, all three terms ultimately consist of sequences, produced by random variables, and therefore characterized by a specific quantity of information. In the cases most typically discussed in this essay, explanandum and input will be assumed to be sequences of lengths n_y and n_x, respectively, resulting from a series of independent identically distributed random variables, Y and X, with discrete alphabets $\mathcal{Y}, \mathcal{X}$, probability distributions p_Y, p_X and therefore Shannon entropy $H(Y) = -\sum p_Y(y) \log p_Y(y)$ and $H(X)$.

The object representing the theory or methodology τ will be typically more complex than Y and X, as it will consist in a sequence of random variables that have distinctive alphabets (are non-identical) and are all uniformly distributed. This sequence of random variables represents the sequence of choices that define a theory and/or methodology. Indicating with T a uniform RV with uniform probability distribution p_T, resulting from a sequence of l RVs T_i of probability distributions p_{T_i}, we have

$$-\log p(\tau) \equiv \log \frac{1}{p_T(\tau)} = \log \frac{1}{Pr\{T_1 = \tau_1, T_2 = \tau_2, ...T_l = \tau_l\}} = \sum_{i \leq l} \log \frac{1}{p_{T_i}(\tau_i)} \quad (2)$$

As will be described shortly, the alphabet of each individual RV composing τ may have any size larger than 2, which represents a binary choice. For example, let τ correspond to the description of three components of a study's method: $\tau =$ ("randomized", "human subject", "female"). In the simplest possible condition, this sequence represents a draw from three independent choices: 1 ="randomized vs. not", 2 ="human vs not", 3 ="female vs not". Representing each choice as a binary RV T_i, the probability of τ is $Pr\{T_1 = \tau_1\} \times Pr\{T_2 = \tau_2\} \times Pr\{T_3 = \tau_3\} = 0.5^3 = 0.125$ and its information content is 3 bits.

Equivalent and useful formulations of equation 1 are

$$K(Y^{n_Y}; X^{n_X}, \tau) \equiv \frac{H(Y) - H(Y|X, \tau)}{H(Y) + \frac{n_X}{n_Y} H(X) - \frac{1}{n_Y} \log p(\tau)} \quad (3)$$

and

$$K(Y^{n_Y}; X^{n_X}, \tau) \equiv k \times h \quad (4)$$

in which

$$k = \frac{H(Y) - H(Y|X, \tau)}{H(Y)} \quad (5)$$

will be referred to as the "effect" component, because it embodies what is often quantified by ordinary measures of effect size (see section 3.2.2), and

$$h = \frac{1}{1 + \frac{n_x H(X) - \log p(\tau)}{n_Y H(Y)}} \quad (6)$$

will be referred to as the "hardness" component, because it quantifies the informational costs of a methodology, which is connected to the concept of "soft science", as will be explained in section 4.8.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

3.2 Why K is a measure of knowledge

Why do we claim that equation 1 quantifies the essence of knowledge? This section will offer three different arguments: a theoretical argument, based on postulating what knowledge consists in; a statistical argument, which illustrates how the K function includes the quantities that are typically computed in ordinary measures of effect size; and a physical argument, which explains how the K function, unlike ordinary measures of effect size or information compression, has a direct physical interpretation in terms of negentropic efficiency.

3.2.1 1) Theoretical argument: K as a measure of pattern encoding

Equation 1 is the mathematical translation of two postulates concerning the nature of the phenomenon we call knowledge:

1. *Information is finite.* The universe is presented to us (or, equivalently, is *necessarily* perceived by us) as a set of discrete, distinguishable states. In absence of distinctions, there is no knowledge to be had. Although in the abstract such states could be infinite (countably infinite, that is), physical limitations ensure that they never are or can be.

2. *Knowledge is information compression.* Knowledge is manifested as an encoding of patterns that connect states, thereby permitting the anticipation of states not yet presented, based on states that are presented. All forms of biological adaptations consist in such pattern encoding, and human cognition and science are merely highly derived manifestations of this process.

Physical, biological and philosophical arguments in support of these two postulates are offered in S1 text.

The most general quantifier of patterns between states is given by Shannon's mutual information function

$$I(Y;X) = H(Y) + H(X) - H(Y,X) = H(Y) - H(Y|X) \tag{7}$$

in which $H(\cdot) = -\sum p(\cdot)\log p(\cdot)$ is Shannon's entropy function. The Mutual Information function is the most general quantifier of patterns, because it is free from any assumption concerning the shape and distribution of the random variables involved (see figure 1). In order to turn equation 7 into an operationalizable quantity of knowledge, we formalize the following properties:

1. the pattern between Y and X is explicitly expressed by a conditioning. We therefore posit the existence of a third random variable, T with alphabet $\mathcal{T} = \{\tau_1, \tau_2 ....\tau_z\}$, such that $H(Y,X|T) = H(Y|T) + H(Y|X,T)$ and $H(Y,X|T) = H(Y) + H(X)$ if $\mathcal{T} = \emptyset$. Unlike Y and X, T is assumed to be uniformly distributed, and therefore the size of its alphabet is $z = |\mathcal{T}| = 2^n$, where n the minimum number of bits required to describe each τ in the set. The uniform distribution of T also implies that $H(T) = -logPr\{T = \tau\} = n$.

2. the quantification of such pattern is standardized, to allow comparisons across systems.

The two requirement above lead us to formulate knowledge as resulting from the contextual, system-specific connection of thee quantities, defined by the following equation

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

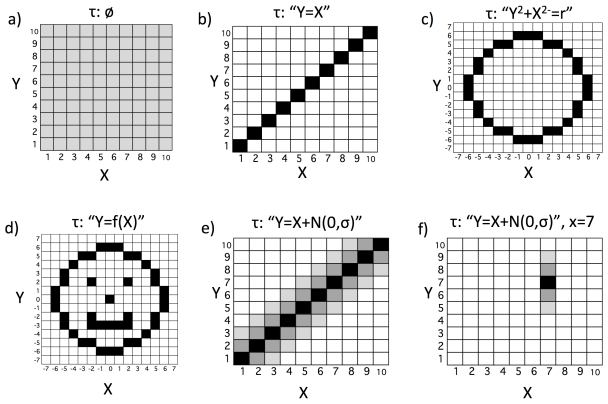


Figure 1. Pictorial representation of the variety of patterns that the K function could quantify. The description of the patterns τ are purely illustrative and are not necessarily literal descriptions of what the pattern encodings would look in practice. A square's level of gray is intended to convey the relative frequency of cells.

.

$$\frac{I(Y;X|T)}{H(Y)+H(X)+H(T)} \equiv \frac{H(Y)-H(Y|X,T)}{H(Y)+H(X)+H(T)} \tag{8}$$

in which, to simplify the notation, we always implicitly assume that $H(Y|T) \equiv H(Y)$ and $H(X|T) \equiv H(X)$.

Note how, at this stage, the value computed by equation 8 is potentially very low, because $H(Y|X,T) = \sum_{i \in \mathcal{T}} P(i)H(Y|X,T=i)$ is the average value of the conditional entropy for every possible theory of description length $-\log p(\tau)$. The more complex the average $\tau \in \mathcal{T}$ is, the larger the number of possible theories of equivalent description length, and thus the smaller the proportion of theories τ_i that yield $H(Y|X,T=\tau_i) < H(Y)$ (because most realizable theories are likely to be nonsensical).

Knowledge is realized because, from *all* possible theories, only a specific theory (or possibly a subset of theories) is selected (Figure 2). This selection is not merely a mathematical fiction, but the result of natural selection or analogous neurological or computational processes. At any rate, regardless of the details of the processes involved, the selection "fixes" the random variable T in equation 8 on a particular realization $\tau \in \mathcal{T}$, with two consequences. On the one hand, the entropy of T goes to zero (because there is no longer any uncertainty about T), but on the other hand, the selection itself entails a non-zero amount of information.

Since T has a uniform distribution, the information necessary to identify this realization of T is simply $-logP(T=\tau) = log2^{l(\tau)} = l(\tau)$, which is the shortest description length of τ. This quantity is an informational cost that needs to be computed in the standardized equation 8. Therefore, we get

$$K(Y;X,\tau) = \frac{H(Y)-H(Y|X,T=\tau)}{H(Y)+H(X)+H(T|T=\tau)+l(\tau)} \equiv \frac{H(Y)-H(Y|X,\tau)}{H(Y)+H(X)-\log p(\tau)} \tag{9}$$

Equation 1 is arrived at by generalizing 9 to the case in which the knowledge encoded by τ is applied to multiple *independent* realizations of explanandum and/or explanans.

3.2.2 2) Statistical argument: K as a universal effect size

Despite having been derived theoretically, and being potentially applicable to phenomena of any kind, i.e. not merely statistical ones, equation 1 bears structural

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

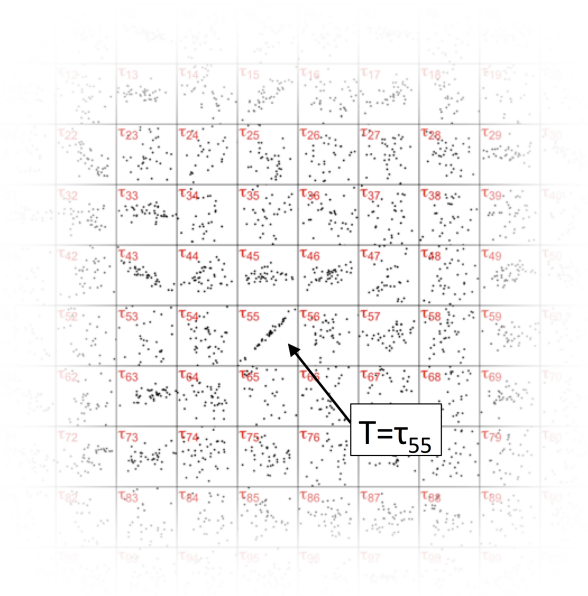


Figure 2. Pictorial representation of a set $\mathcal{T} = \{\tau_1, \tau_2, ...\tau_z\}$ of theories of a given description length that condition the relation between two variables. This set constitutes the alphabet of the uniformly distributed random variable T, from which a specific outcome, in this case τ_{55}, is selected. For further discussion, see text.

similarities with ordinary measures of statistical effect size. Such similarities ultimately follow from the argument made above that the K function captures the essence of any knowledge process - indeed, they offer additional support to the theoretical argument.

To illustrate such similarities, it is useful to point out that the value of the K function can be approximated from the quantization of any continuous probability distribution. For information to be finite as required by the K function, the entropy of a normally distributed quantized random variable X^{Δ} can be approximated by $H(X^{\Delta}) = \log\sqrt{2\pi e}\sigma$, in which σ is the standard deviation rescaled to a lowest decimal (for example, from $\sigma = 0.123$ to $sigma = 123$, further details in S2 text).

There is a clear structural similarity between the "effect" component of equation 4 and the coefficient of determination R^2. Since the entropy of a random variable is monotonically increasing function of the variable's dispersion (e.g. its variance) this measure is directly related to K. For example, if y and $y|x$ are normally distributed with variance σ_y,

$$R^2 \equiv \frac{TSS - SSE}{TSS} \equiv \frac{n \times (\sigma_y^2 - \sigma_{y|x}^2)}{n \times \sigma_y^2} = f\left(\frac{\log\sigma_y - \log\sigma_{y|x}}{\log\sigma_y + C}\right) = f(K(Y; X, \tau)) \quad (10)$$

in which TSS is the Total Sum of Squares, SSE is the Sum of Squares of the Residuals, and n is sample size. The adjusted coefficient of determination R^2_{adj} is also directly related to K since

$$R^2_{adj} = \frac{\frac{TSS}{n-1} - \frac{SSE}{n-k}}{\frac{TSS}{n-1}} = g\left(\frac{\log\sigma_y^2 - \log(\sigma_{y|x}^2 \times A)}{\log\sigma_y^2}\right) = f(K(Y; X, \tau)) \quad (11)$$

with $A = \frac{n-1}{n-k}$.

From this relation follows that multiple ordinary measures of statistical effects size used in meta-analysis are also functions of K. In particular, for any two random

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

variables, $R^2 = r^2$, with r their Spearman's correlation coefficient. And since most popular measures of effect size used in meta-analysis, including Cohen's d and Odds Ratios, are approximately convertible to and from r [9], they are also convertible to K.

The direct connection between K and measures of effect size like Cohen's d implies that K is also related to the t and the F distribution, which are constructed as ratios between the amount of what is explained and what remains to be explained, and are therefore constructed similarly to an "odds" transformation of K

$$OK(Y;X,\tau) \equiv \frac{K(Y;X,\tau)}{1-K(Y;X,\tau)} = \frac{n_Y(H(Y)-H(Y|X,\tau))}{n_Y H(Y|X,\tau)+n_X H(X)-\log p(\tau)} \tag{12}$$

Other more general tests, such as the Chi-squared test, can be shown to be an approximation of the Kullback-Leibler distance between the probability distributions of observed and expected frequencies [30] and therefore a measure of the mutual information between two random variables, i.e. the same measure on which the K function is built. We can also show how in practice the structure of the Chi-squared test is analogous to that of K (see S2 text).

Figure 3 illustrates how these are not merely structural analogies, because K can be approximately or exactly converted to ordinary measures of effect size. However, note that the K function computes properties that are critical to knowledge and that are ignored by ordinary measures of effect size. Such properties include the number of repetitions (which, depending on analyses, may correspond to the sample size or to the intended total number of uses of a τ); accuracy of scale (section 3.3.6); distance in time and space and methods (sections 3.3.5); and Ockham's razor (section 3.3.1). The latter property also makes K conceptually analogous to measures of minimum description length, discussed below.

Minimum Description Length Principle The Minimum Description Length principle is a formalization of the principle of inductive inference and of Ockham's razor that has many potential applications in statistical inference, particularly with regards to the problem of model selection [26]. In its most basic formulation, the MDL states that the best model to explain a data is the one that minimizes the quantity

$$L(H) + L(D|H) \tag{13}$$

in which $L(H)$ is the description length of the hypothesis (i.e. a candidate model for the data) and $L(D|H)$ is the description length of the data given the model. The K equation has equivalent properties to equation 13, with $L(H) \equiv -\log p(\tau)$ and $L(D|H) \equiv n_y H(Y|X,\tau)$. Therefore, the values of that minimize equation 13 maximize the K function.

The MDL principle may be considered as the extension of earlier information-based measures model fit, such as the popular Akaike Information Criterion (AIC), which can be shown to give equivalent results to K when sample sizes are assumed to be infinite (a condition that K theory, however, excludes, see S4 text). Analogous reasoning applies to other measures closely related to MDL, such as Bayesian Information Criterion BIC, which differs from AIC in weighting the number of model parameters by the logarithm of the sample size.

***

The reader may question why, if K is equivalent to existing statistical measures of effect size, we couldn't just use the latter to quantify knowledge. There are at least three reasons. First, because K is a truly universal measure of effect size. The quantity measured by K is completely free from any distributional assumptions about the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

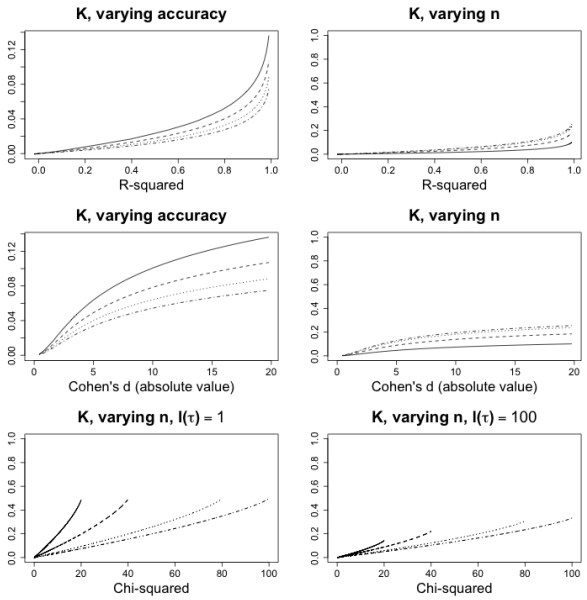


Figure 3. Relation between K and common measures of effect size, with varying conditions of accuracy (i.e. of resolution, see section 3.3.6), number of repetitions n (i.e. the n_y in equation 1) and size of τ. The relation with R-squared and Cohen's d was derived assuming a normal distribution of the explanandum. Increasing accuracy thus corresponded to calculating entropies with a standard deviation measured with one additional significant digit, at each step, from solid line to dotted line. The values of n for R^2 and Cohen's d were, from dotted to solid line, $\{1, 2, 10, 100\}$ respectively. The relation with Chi-squared was derived from the probability distribution of a 2×2 contingency table. From solid to dotted line, the value of n was $20, 40, 80, 100$, and the description length of τ was varied as indicated in the panels' titles. The code used to generate these and all other figures is available in S11 text

.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

subject matter being assessed. It can be applied not only to quantitative data produced by any probability distribution (e.g. Figure 1), but also to any other explanandum that has a finite description length (although this potential application will not be examined in detail in this essay). In essence, K can be applied to anything that is quantifiable in terms of information, which means any phenomenon that is the object of cognition - any phenomenon amenable to being "known". Second, as mentioned above, K is not only free from assumptions but is also more complete than any individual measure of effect size, and therefore is a more complete representation of knowledge phenomena. Third, because, unlike any of the statistical and algorithmic approaches discussed above, K has a straightforward physical interpretation, which is presented in the next section.

3.2.3 3) Physical argument: K as a measure of negentropic efficiency

The physical meaning of equation 1 is derived from the physical meaning of information, which was revealed by the solution to the famous paradox known as Maxwell's Demon. In the most general formulation of this *Gedankenexperiment*, the Demon is an organism or a machine that is able to manipulate molecules of a gas, for example by operating a trap door, and thereby is able to separate molecules that move at high speed from the slower ones. By doing so, the Demon can lower the entropy of a system, seemingly without dissipation. This created a theoretical paradox as it would contradict the second law of thermodynamics, according to which no process can have as its only result the transfer of heat from a cooler to a warmer body.

In one variant of this paradox, called the "pressure Demon", a cylinder is immersed in a heat bath and has a single "gas" molecule moving randomly inside it. The demon inserts a partition right in the middle of the cylinder, thereby trapping the molecule in one half of the cylinder's volume. It then operates a measurement to assess in which half of the cylinder the molecule is, and pushes down, with a reversible process, a piston in the half that is empty. The demon could then remove the partition, allowing the gas molecule to push the piston up, and thus extract work from the system, apparently with no dissipation.

Objections to the paradox that involve the energetic costs of operating the machine or of measuring the position of the particle [23] were proven to be invalid, at least from a theoretical point of view [24, 31]. The conclusive solution to the paradox was given in 1982 by Charles Bennett, who showed that dissipation in the process occurred as a byproduct of the Demon's need to process information [32]. In order to know which piston to lower, the Demon must memorize the position of the molecule, storing one bit of information, and it must eventually re-set its memory to prepare it for the next measurement. The recording of information can occur with no dissipation, but the *erasure* of it is an *irreversible* process that will produce heat that is at least equivalent to the work extracted from the system, i.e $kT \ln 2$ joules. where k is Boltzmann's constant. This solution to the paradox proved that information is a measurable physical quantity.

Figure 4) illustrates how the K function relates to Maxwell's pressure Demon. The explanandum $H(Y)$ quantifies the uncertainty about the molecule's position. The input $H(X)$ is the information aquired by measurement, external input of information, obtained by a measurement. The input corresponds to the colloquial notion of "information" as something that is acquired, and may be perhaps more correctly defined as *negentropy* [23].

The theory τ contains a description of the information-processing structure that allows the pressure Demon to operate. The size of τ may be hard to quantify in absolute sense, but can be described and measured in relative terms. A minimal description will include at least an encoding of the identity relation between the state of X and that of Y, i.e. "$X = Y$" as distinguished from its opposite "$X \neq Y$". This theory requires at least a binary alphabet, and one bit of memory storage. A more comprehensive

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

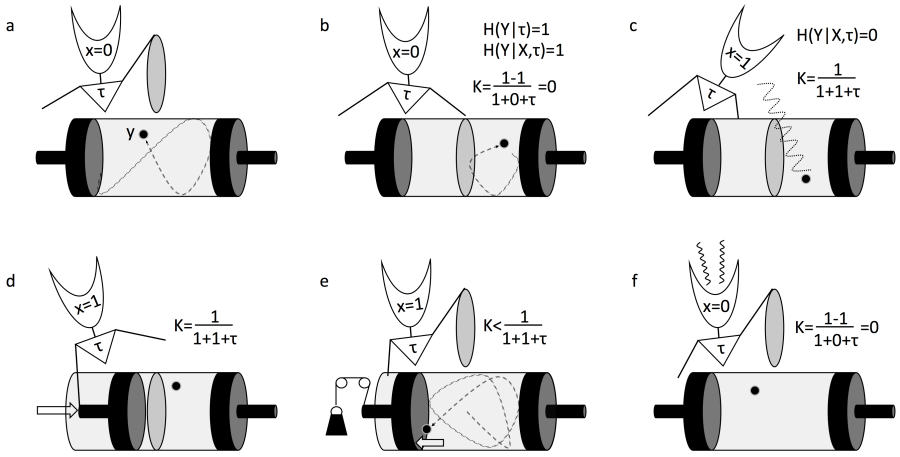


Figure 4. Illustration of Maxwell's "Pressure Demon" paradox, and how it relates to K. See text for further discussion.

.

description will include a description of the algorithm that enables the negentropy in X to be exploited - something like "if X=left, press down right piston, else, press left piston". An even broader physical description of the pressure Demon ought to encode instructions to set up the entire system, i.e. the heat bath, the partition etc. In other words, a complete τ includes the genetic code to reproduce pressure Demons.

The K function quantifies the efficiency with which the Demon converts information into work. At the start of the cycle, the Demon's K is zero. After measuring the particle's position, the demon has stored one bit of information, and has knowledge $K > 0$. By setting the piston and removing the partition, the demon puts his knowledge to use, and extracts $k \ln 2$ of work from it. Once the piston is fully pushed out, the Demon no longer knows where the molecule is ($K = 0$) and yet still has one bit stored in memory, a trace of its last experience. The Demon has now two possible options. First, as in Bennett's solution to the process, it can simply erase that bit, re-setting X to an initial state, getting $H(X) = 0$ and releasing $k \ln 2$ in the environment. At each cycle, the negentropy is renewed, whereas the fixed τ component remains unaltered. Since the position of the molecule at each cycle is independent of previous positions, the explanandum grows by one bit, whereas the theory component remains unaltered. For n cycles, the total K is:

$$K = \frac{nH(Y)}{nH(Y) + nH(X) - \log p(\tau)} = \frac{1}{1 + 1 - \frac{\log p(\tau)}{n}} \tag{14}$$

Which, to the limit of infinite cycles is

$$\lim_{n \to \infty} K = 1/2 \tag{15}$$

The value of $K = 1/2$ constitutes the absolute limit for knowledge that requires a direct measurement and/or a complete and direct description of the explanandum.

Alternatively, the Demon could keep the value of X in memory, and make space for new information to be gathered in the next cycle (see [24]). As Bennett also pointed out, in practice it could not do so forever. In any physical implementation of the experiment, the Demon would eventually run out of memory space and would be forced to erase some of it, releasing the entropy locked in it. If, *ad absurdum* the Demon stored

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

an infinite amount of information, then at each cycle the input would grow by one bit yielding:

$$K = \frac{1}{1 + n - \frac{\log p(\tau)}{n}} \tag{16}$$

To the limit of infinite cycles, $\lim_{n\to\infty} K = 1/(1+n) = 0$, again independent of τ. This is a further argument to illustrate how information is necessarily finite, as we postulated (section 3.2.1).

More realistically, we can imagine that the number of physical bits available to the Demon is finite. As cycles progress, the Demon could try to allocate as many resources as possible to the memory $n_X X$, for example by reducing the space occupied by τ. This is why knowledge entails compression and pattern encoding (see also section 3.3.1).

Elaborations on the pressure Demon experiment shed further light on the meaning of K and its implications for knowledge. First, let's imagine that the movement of the gas molecule is not actually random, but that, acted upon by some external force, the molecule periodically and regularly finds itself alternatively on the right and left side of the cylinder, and expands from there. If the Demon kept a sufficiently long record of past measurements, say a number z of bits, it might be able to discover the pattern. Its τ could then store a new, slightly expanded algorithm, like "if last position was left, new position is right, else, new position is left". With this new theory, and one bit of input to determine the initial position of the molecule, the Demon could extract unlimited amounts of energy from the heat bath. In this case,

$$K = \frac{1}{1 + \frac{1}{n} - \frac{\log p(\tau)}{n}} \tag{17}$$

which to the limit of infinite cycles tends to $\lim_{n\to\infty} K = 1$.

Intermediate cases are also easy to imagine, in which the behaviour of the molecule is predictable only for a limited number of cycles, say m. In such case, K would increase as the number of necessary measurements n_x is reduced to n_x/m. At any rate, this example illustrated how the Demon's ability to implement knowledge (in order to extract work, create order, etc.) is determined by the presence of regularities in the explanandum as well as the efficiency with which the Demon can identify and encode patterns, i.e. by minimizing the explanans. The Demon is selected to be as "intelligent" and "informed" as possible.

As a second case, let's imagine instead that the gas molecule moves at random and that its position is measurable only to limited accuracy. A single measurement yields the position of the molecule with an error η. However, each additional measurement reduces η by a fraction a. The Demon, in this case, could benefit from increasing the number of measurements. Indicating with m the number of measurements and with τ_m the corresponding theory:

$$K = \frac{1 - \eta \times a^{-m}}{1 + m - \frac{\log p(\tau_m)}{n}} \tag{18}$$

which to the limit of infinite cycles tends to

$$\lim_{n\to\infty} K = \frac{1 - \eta \times a^{-m}}{1 + m} < \frac{1}{2} \tag{19}$$

The absolute amount of work extracted at each cycle will be $k \ln 2(1 - \eta \times a^{-m})$. Therefore, K expresses the efficiency with which work can be extracted from a system, given a certain error rate a and number of measurements m (see S4 text for an illustration).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

3.3 Properties of knowledge

This section will illustrate how K possesses properties that would be expected by a measure of knowledge. In addition to offering support for the three arguments given above, these properties underlie some of the results presented in section 4.

3.3.1 Ockham's razor is relative

As discussed in section 3.2.2, the K function encompasses the MDL principle, and therefore computes a quantification of Ockham's razor. However, the K formulation of Ockham's razor highlights a property that not all MDL formulations encompass: that Ockham's razor is relative to the size of the explanandum. For a given Y and X and two alternative theories τ and τ' that have the same effect $H(Y|X,\tau) = H(Y|X,\tau')$ and that can be applied to a number of repetitions n_Y and n'_Y, respectively, we have that:

$$\frac{-\log p(\tau')}{n'_Y} < \frac{-\log p(\tau)}{n_Y} \Rightarrow K(Y^{n'_Y}; X, \tau') > K(Y^{n_Y}; X, \tau) \tag{20}$$

and similarly for the case in which $\tau = \tau'$ whilst $n_X H(X) \neq n'_X H(X')$:

$$\frac{n'_X H(X')}{n'_Y} < \frac{n_X H(X)}{n_Y} \Rightarrow K(Y^{n'_Y}; X^{n'_{X'}}, \tau) > K(Y^{n_Y}; X^{n_X}, \tau). \tag{21}$$

Therefore, the knowledge relevance of an explanans' simplicity, i.e. Ockham's razor, is modulated by the number of times that the explanans can be applied to the explanandum.

3.3.2 Prediction is more costly than explanation, but preferable to it

The K function can be used to quantify either explanatory or predictive efficiency. When the terms of the K function are entropies, i.e. expectation values of uncertainties, then K quantifies the expected (average) explanatory or predictive efficiency of an explanans with regards to an explanandum. Conversely, if the explanandum is an event that has already occurred and that carries information $-\log P(Y = y)$, K quantifies the value of an explanation, by minimizing the surprisal of explanatory conditions $-\log P(X = x)$ and/or the complexity of the theory linking such conditions to the event, $-\log P(T = \tau)$, in order to maximize K. Conversely, if a theory and/or an input are pre-determined, these give rise to a predicted probability distribution that can be compared to observations. To any extent that observations do not match predictions, the observed and predicted distribution will have a non-zero informational divergence. The latter is the extra amount of information that would be needed to "adjust" the predictions, post-hoc or ante-hoc, to make them match the observations. It follows that, indicating with the tilde sign the predictive theory, we can calculate an "adjusted" K as

$$K_{adj} = K_{obs} - D(Y|X,\tau||Y|X,\tilde{\tau})\frac{h}{H(Y)} \tag{22}$$

in which $K_{obs} = K(Y; X, \tau)$ is the K observed, and $D(\cdot)$ is the Kullback-Leibler divergence between the observed and the predicted distribution (proof in S5 text). Since $D(Y|X,\tau||Y|X,\tilde{\tau}) \geq 0$, $K_{adj} \leq K_{obs}$, with equality corresponding to perfect fit between observations and predictions. An analogous formula could be derived for the case in which the explanandum is a sequence, in which case the distance would be calculated following methods suggested in section 4.3.3.

Now, note that the observed K *is* the explanatory K, and therefore is always greater or equal to the predictive K for individual observations. When evidence cumulates, then the explanans of an explanatory K is likely to expand, reducing the cumulative K

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

(see section 4.3). Replacing a "flexible" explanation with a fixed one avoids these latter cumulative costs, allowing a fixed explanans to be applied to a larger number of cases n_y, with no cumulative increase in its complexity.

Therefore, predictive knowledge is simply a more generalized, ideally unchanging form of explanatory knowledge. As intuition would suggests, prediction can never yield more knowledge than a post-hoc explanation for a given event (e.g., an experimental outcome). However, it becomes cumulatively more valuable, to the extent that it allows to explain, with no changes, a larger number of events, backwards or forward in time.

3.3.3 Causation entails correlation, and is preferable to it

Properties of the K function also suggests why the knowledge we gain from uncovering a cause-effect relation is often, but not always, more valuable than that derived from a mere correlation. Definitions of causality have a long history of subtle philsophical controversies [33]. No defintion of causality, however, can dispense with relying on counterfactuals and/or with assuming that manipulating present causes can change future effects [34]. The difference between a mere correlation and a causal relation can be formalized as the difference between two types of conditional probabilities, $P(Y = y|X = x)$ and $P(Y = y|do(X = x))$, where "$do(X = x)$" is a shorthand for "$X|do(X = x)$" and the "do" function indicates the manipulation of a variable. In general, the presence of a correlation entails $P(Y = y) \leq P(Y = y|X = x)$ and $P(Y = y) = P(Y = y|do(X = x))$ whereas a causation $P(Y = y) \leq P(Y = y|X = x) \leq P(Y = y|do(X = x))$.

If knowledge is correlational, then $K(Y; X = x, \tau) > 0$ and $K(Y; do(X = x), \tau) = 0$, otherwise $K(Y; X = x, \tau) > 0$ and $K(Y; do(X = x), \tau) > 0$. Hence, all else being equal, the knowledge attainable via causation is larger under a broader set of conditions. Moreover, note that in the correlational case knowledge is only attained once an external input of information is obtained, which has an informational cost $n_Y H(X) > 0$. In the causal case, conversely, the input has no informational cost, i.e. $H(X|do(X = x)) = 0$, because we have no uncertainty about its state. However the explanans is expanded by an additional $\tau_{do(X=x)}$, which is the description length of the methodology that imposes a value to X. Therefore, the value of causal knowledge is defined as:

$$K(Y; \tau, \tau_{do(X=x)}) = \frac{n_Y H(Y) - n_Y H(Y|X, \tau)}{n_Y H(Y) - \log p(\tau) - \log p(\tau_{do(X=x)})} \equiv \frac{H(Y) - H(Y|X, \tau)}{H(Y) + \frac{-\log p(\tau_{do(X=x)}) - \log p(\tau)}{n_Y}} \quad (23)$$

It follows that there is always an $n^* \in \mathbb{N}$ such that $K(Y^{n^*}; \tau, \tau_{do(X=x)}) > K(Y^{n^*}; X^{n^*}, \tau)$. Specifically, assuming τ to be constant, causal knowledge is superior to correlational knowledge when $n^* > -\log p(\tau_{do(X=x)})/H(X)$.

3.3.4 Knowledge growth requires lossy information compression

Both theoretical and physical arguments suggest that K is maximized when τ is minimized. A simple calculation shows that such minimization must eventually consist in the encoding of abstract patterns that offer an incomplete account of the explanandum, otherwise knowledge cannot grow indefinitely.

Let τ be a theory that is not encoding a relation between X and Y, but merely lists the co-occurrence of each element $x \in \mathcal{X}$ with one or more elements $y \in \mathcal{Y}$. Clearly, such τ would always yield $H(Y|X, \tau) = 0$, but its costs increase exponentially with the size of the two alphabets. Indicating the sizes of the two alphabets as $s_X = |\mathcal{X}|$ and

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

$s_Y = |\mathcal{Y}|$, the size of τ would be proportional to $log\frac{s_Y!}{|s_Y - s_X|!}$. As the size of of the alphabet $\mathcal{Y}$ grows, knowledge declines because

$$\lim_{s_Y \to +\infty} K(Y;X,\tau) = \lim_{s_Y \to +\infty} \frac{n_Y H(Y)}{n_Y H(Y) + n_X H(X) + \frac{log(s_y!)}{s_y}} = 0 \tag{24}$$

independent of the probability distribution of Y and X. Therefore, as the explanandum is expanded (i.e. its total information and/or complexity grows), knowledge rapidly decreases, unless τ is something other than a listing of y, x pairs. In other words, knowledge cannot grow unless τ consists in a relatively short description of some pattern that captures a redundancy in the system. The knowledge cost of a finite level of error or missing information $H(Y|X,\tau) > 0$ will soon be preferable to an exceedingly complex τ.

3.3.5 Decline with distance in time, space and/or explanans

Both predictions and explanations about empirical phenomena tend to become less accurate as the distance in time or space of the explanandum increases. The informational equivalent of our physical intuition of distance is quantified by the information divergence measures. It can be shown that such divergence leads, under most conditions, to a decline of K that can be described by a simple exponential function of the form

$$K(Y';X',\tau') = K(Y;X,\tau) \times A^{-\boldsymbol{\lambda}\cdot\boldsymbol{d}} \tag{25}$$

in which Y', X', τ' are a system having an overall distance (i.e. informational divergence) $\boldsymbol{d}$ from Y, X, τ, and $\boldsymbol{\lambda}\cdot\boldsymbol{d} = d_Y\lambda_Y + d_{\tau 1}\lambda_{\tau 1} + d_{\tau 2}\lambda_{\tau 2} + ... + d_{\tau l}\lambda_{\tau 1}$ defines the decline rate depending on the divergence between explanandum and explanans and between explanantia (proof in S6 text).

3.3.6 Knowledge has an optimal resolution

Accuracy of measurement is a special case of the general informational concept of resolution, quantifiable as the number of bits that are available to describe explanandum and explanans. It can be shown both analytically and empirically that any system Y, X, τ is characterized by an optimal level of resolution (the full argument is offered in S7 text).

We may start by noticing how, even if empirical data is assumed to be measurable to infinite accuracy (against one of the postulates in section 3.2.1), the resulting K value will be inversely proportional to measurement accuracy, unless special conditions are met. When K is measured on a continuous, normal and quantized random variable Y^{Δ} (see section 3.2.2), to the limit of infinite accuracy only one of two values is possible:

$$\lim_{n \to \infty} K(Y^{\Delta};X,\tau) = \begin{cases} \lim_{n\to\infty} \frac{h(Y)+n-h(Y|X,\tau)-n}{h(Y)+n+X+\tau} = 0 \\ \lim_{n\to\infty} \frac{h(Y)+n}{h(Y)+n+X+\tau} = 1 \end{cases} \tag{26}$$

The upper value occurs if and when $h(Y|X,\tau) > 0$, i.e. by assumption there is a non-zero residual uncertainty that needs to be measured. This is the typical case of empirical knowledge. The lower value presupposes *a priori* that $h(Y|X,\tau) = 0$, i.e. the explanandum is perfectly known via the explanans. This case is only represented by logico-deductive knowledge.

We now relax the assumption of infinitely measurable accuracy, which is commonly made in mathematical models, but it is unrealistic. Models such as the normal distribution are approximation, and the actual information that may be extracted from

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

a variable is related to its finite quantization. Therefore, we can define empirical systems, in opposition to logico-deductive systems, as systems that have a finite maximal resolution. Let $X^{\alpha} \equiv n_X X^{\alpha}$ be an explanans or explanandum quantized to resolution α, and let $\alpha' = \alpha/n$ be an increased quantization. The maximal resolution of X^{α} is a quantity $e > 0, e \in \mathbb{Q}$ such that:

$$H(X^{\alpha'}) - H(X^{\alpha}) = log(n), \forall \alpha \leq e \tag{27}$$

For all explananda and explanantia characterized by such maximal resolution, the corresponding K is maximized by an optimal, non-zero and non-infinite resolution $\alpha*_y$ and $\alpha*_x$ such that:

$$K(Y^{\alpha*_Y}; X^{\alpha*_X}, \tau) > K(Y^{\alpha_Y}; X^{\alpha_X}, \tau) \forall \alpha_Y \neq \alpha*_Y, \alpha_X \neq \alpha*_X \tag{28}$$

In general, moreover,

$$K(Y^{\frac{\alpha_Y}{2}}; X^{\alpha_X}, \tau) > K(Y^{\alpha_Y}; X^{\alpha_X}, \tau) \iff$$
$$H(Y^{\frac{\alpha_Y}{2}} | X^{\alpha_X}, \tau) - H(Y^{\alpha_Y} | X^{\alpha_X}, \tau) < 1 - K(Y^{\alpha_Y}; X^{\alpha_X}, \tau) \tag{29}$$

which entails that K may increase with higher resolution of Y only up to a maximal level, and subsequently decline back to zero. For the increase of the resolution of X, the condition is $H(Y^{\alpha_Y}|X^{\alpha_X}, \tau) - H(Y^{\alpha_Y}|X^{\frac{\alpha_x X}{2}}, \tau) > K(Y^{\alpha_Y}; X^{\alpha_X}, \tau)$ (see S7 text).

The dependence of K on resolution reflects its status as a measure of entropic efficiency (section 3.2.3) and entails that, to compare systems for which the explanandum is measured to different levels of accuracy, the K value needs to be re-scaled. Such re-scaling can be attained rather simply, by multiplying the value of K by the entropy of the corresponding explanandum. The resulting product $H(Y) \times K(Y; X, \tau)$ quantifies in absolute terms how many bits are extracted from the explanandum by the explanans.

A system's optimal resolution is determined by the shape of the relation between explanandum and explanans. Two simulations in Figure 5 illustrate how both K and $H(Y)K$ may vary depending on resolution (the complete figures are in S7 text and S7 text.).

4 Results

This section will illustrate, with practical examples, how the tools developed so far can be used to answer meta-scientific questions. Each of the questions is briefly introduced by a problem statement, followed by the answer, which comprises of a mathematical equation and an explanation, and one or more examples. Most of the examples are offered as suggestions of potential applications of the theory, and the specific results obtained should not be considered conclusive.

4.1 How much knowledge is contained in a theoretical system?

Problem: Unlike empirical knowledge, that is amenable to errors that can be verified against experiences, knowledge derived from logical and deductive processes conveys absolute certainty. It might therefore seem impossible to compare the knowledge yield of two different theories, for example two mathematical theorems. The problem is made even deeper by the fact that any logico-deductive system is effectively a tautology, i.e. a system that derives its own internal truths from a starting set of principles taken to be true. How can we quantify the knowledge contained in a theorem?

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

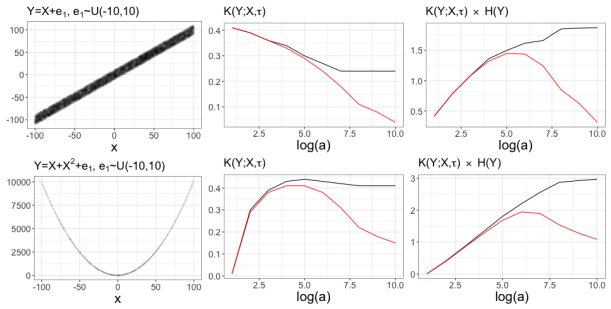


Figure 5. Illustrative example of how, as the resolution with which explanandum and explanans is changed, K varies dependong in the shape of the pattern. The figures and all the calculations were derived from a simulated dataset, in which the pattern linking explanandum to explanans was assumed to have a noise with uniform distribution. Black line: entropies and K values calculated by maximum likelihood method (i.e. bin counting). Red line: entropies and K values calculated using the "shrink" method described in [35] (the R code used to generate the figures is provided in S11 text). Note how the value of K and its re-scaled version $H(Y)K$ are maximized at a single optimal resolution. A complete figure, showing the different levels of resolution, is available in S7 text and S7 text.

.

Answer: The value of theoretical knowledge is quantified as

$$K = h \tag{30}$$

in which K corresponds to equation 1 and h to equation 6.

Explanation: Logico-deductive knowledge, like all other forms of knowledge, ultimately consists in the encoding of patterns. Mathematical knowledge, for example, is produced by revealing previously unnoticed logical connections between a statement with uncertainty $H(Y)$ and another statement, which may or may not have uncertainty $H(X)$ (depending on whether X has been proven, postulated or conjectured), via a set of passages described in a proof τ. The latter consists in the derivation of identities, creating an error-free chain of connections such that $P(Y|X,\tau) = 1$.

When the proof of the theorem is correct, the effect component k in equation 4, is always equal to one, yielding equation 30. However, when the chain of connections τ is replaced with a τ' at a distance $d_\tau > 0$ from it, k is likely to be zero, because even minor modifications of τ (for example, changing a passage in the proof of a theorem) break the chain of identities and invalidate the conclusion. This is equivalent to assuming that $\lambda_\tau \approx \infty$. Therefore, the reproducibility of mathematical knowledge, as it is embodied in a theorem, is either perfect or null:

$$K_r = K \quad if \quad d_\tau = 0, \quad 0 \quad otherwise \tag{31}$$

Alternative proofs τ', however, might also occur, and their K value will be inversely proportional to their length, since the shorter proof yields a higher h.

Once a theorem is proven, its application will usually not require invoking the entire proof τ. In K, we can formalize this fact by letting τ be replaced by a single symbol encoding the nature of the relationship itself. The entropy of τ will in this case be minimized to that of a small set of symbols, e.g. $\{=, \neq, >, < ...\}$. In such case, the value of the knowledge obtained will be mostly determined by n_Y, which is the number of times that the theorem is going to be invoked and used.

Hence, as will be proven in the example below, we reach the general conclusion that *the value of a theory is inversely related to its length and directly related to the frequency*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

of its use.

4.1.1 Example: The proof of Fermat's last theorem

Fermat's last theorem (henceforth, FLT) states that there is no solution to the equation $a^n + b^n = c^n$ when all terms are positive integers and $n > 2$. The french mathematician Pierre de Fermat (1607-1665) claimed to have proven such statement, but his proof was never found. In 1995, Andrew Wiles published a proof of FLT, winning a challenge that had engaged mathematicians for three centuries [36]. How valuable was Wiles' contribution?

We can describe the explanandum of FLT's as a binary question: "does $a^n + b^n = c^n$ have a solution"? In absence of any proof τ, the answer can only be obtained by calculating the result for any given set of integers $[a, b, c, n]$. Let n_y be the total plausible number of times that this result could be calculated. Of course, we cannot estimate this number exactly, but we are assured that this number is an integer (because a calculation is either made or not), and that it is finite (because the number of individuals, human or otherwise, who have, will, or might do calculations is finite). Therefore, the explanandum is $n_y H(Y)$. For simplicity, we might assume that in absence of any proof, individuals making the calculations are genuinely agnostic about the result, such that $H(Y) = 1$.

Indicating with τ the maximally succinct (i.e. maximally compressed) description of this proof, the knowledge yielded by it is:

$$K(Y^{n_Y}; \tau) = \frac{n_Y H(Y)}{n_Y H(Y) - \log p(\tau)} \equiv \frac{1}{1 + \frac{-\log p(\tau)}{n_y}} \tag{32}$$

Here we assume that any input is contained in the proof τ. The information size of the latter is certainly calculable in principle, since, in its most complete form, it will consists in an algorithm that derives the result from a small set of axioms and operations.

Wiles' proof of FLT is over 100 pages long and is based on highly advanced mathematical concepts that were unknown in Fermat's times. This suggest that, assuming Fermat had actually discovered a correct proof of the theorem, then his proof was considerably simpler and shorter than Wiles'. Mathematicians are now engaged in the challenge of discovering such a simple proof.

How would a new, simpler proof compare to the one given by Wiles? Indicating this simpler proof with τ' and since n_Y is constant and independent of the proof, the proportional gain in knowledge is

$$K(Y; \tau') - K(Y; \tau) = \frac{1}{1 - \log p(\tau')} - \frac{1}{1 - \log p(\tau)} = \frac{-\log p(\tau) - (-\log p(\tau'))}{(1 - \log p(\tau'))(1 - \log p(\tau))} \approx \frac{\log p(\tau') - \log p(\tau)}{\log p(\tau') \times \log p(\tau)} \tag{33}$$

equation 33 reflects the gain in knowledge obtained by devising a simpler, shorter proof of a previously proved theorem.

Given two theorems addressing different questions, instead, the difference in knowledge yield will depend on the lengths of the respective proofs as well as the number of computations that each theorem allows to be spared. The general formula is, indicating with Y' and τ' an explanandum and explanans different from Y and τ:

$$K(Y'; \tau') - K(Y; \tau) = \frac{n' \log p(\tau') - n \log p(\tau)}{(n' - \log p(\tau'))(n - \log p(\tau))} \tag{34}$$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

4.2 How much knowledge is contained in an empirical system?

Problem: Science is at once a unitary phenomenon and highly diversified and complex one. It is unitary in its fundamental objectives and in general aspects of its procedures, but it takes a myriad different forms when it is realized in individual research fields, whose diversity of theories, methodologies, practices, sociologies and histories mirrors that of the phenomena being investigated. How can we compare the knowledge obtained about widely different subject matters?

Answer: The knowledge produced by a study, a research field, and generally a methodology is quantified as

$$K = k \times h \tag{35}$$

in which K is given by equation 1, k by equation 5 and h by equation 6.

Explanation: Knowledge entails a reduction of uncertainty, attained by the processing of stored information. Equation 35 quantifies the efficiency with which uncertainty is reduced. This is a scale-free, system-specific property. The system is uniquely defined by a combination of explanandum, explanans and theory, the information content of which is subject to physical constraints. Such physical constraints ensure that, amongst other properties, every system Y, X, τ has an optimal resolution, non-zero and non-infinite, and therefore a unique identifiable value K (section 3.3.6). As discussed in section 3.3.6, this quantity can also be re-scaled to $K \times H(Y)$, which gives the total net number of bits that are extracted from the explanandum by the explanans. Since $k \leq 1$, theoretical knowledge is always as large or larger than empirical knowledge. Equation 35 applies to descriptive knowledge as well as correlational or causal knowledge, as examples below illustrate.

4.2.1 Example 1: The mass of the electron

Centuries of progressively accurate measurements have yielded the current value for the mass of the electron of $m_e = 9.10938291 \pm 40 \times 10^{-31}$ Kg, with the error term representing the standard deviation of normally distributed errors [37]. Since this is a fixed number of (currently) nine significant digits, the explanandum is quantified by the amount of storage required to encode it, i.e. a string of information content $-\log P(Y = y) = 9 \times \log(10)$, and the residual uncertainty is quantified by the entropy of the normal distribution of errors with $\sigma = 40$. These measurements are obtained by complex methodologies that are in principle quantifiable as a string of inputs and algorithms, $-\log p(x) - \log p(\tau)$. However, the case of physical constants is similar to that of a mathematical theorem, in that the explanans becomes negligible to the extent that the value obtained can be used in a very large number of subsequent applications. Therefore, we estimate our current knowledge of the mass of the electron to be

$$K(m_e) = \frac{9 \log 10 - \log \sqrt{2\pi e} 40}{9 \log 10} \frac{1}{1 + \frac{-\log p(x) - \log p(\tau)}{n_Y 9 \log 10}} \approx 0.754 \tag{36}$$

with the last approximation due to the case that the value can be stored and used for a very large n_Y times, yielding $h \approx 1$. More accurate calculations would require estimating the h component, too. In particular, to compare $K(m_e)$ to the K value of another constant, the relatively frequency of use would need to be taken into account. The corresponding re-scaled value is $K(m_e) \times 9 \log 10 = 22.53$.

Since our estimation of $K(m_e)$ is mainly derived from the number of digits used to encode the value of the electron, we could estimate it using a lower number of digits and ignore the current measurement error. It would be tempting, in particular, to quantify knowledge for the mass measured to six significant digits only (which is likely to cover

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

at least 99% of possible values). By doing so, we would obtain $K(m_e) \approx 1$, showing that at that level of accuracy we have virtually perfect knowledge of the mass of the electron. However, if we were to re-scale this value we would have $K(m_e) \times 6 \times \log 10 = 19.93$, a lower number of bits actually spared. By lowering the resolution, our knowledge increased in relative but not in absolute terms.

It should be emphasized that we are measuring here the knowledge value of the mass of the electron in the narrowest possible sense, i.e. by restricting the system to the mass itself. However, the knowledge we derive by measuring (describing) phenomena such as a physical constant acquires value also in a broader context, in its role as *as an input* required to know other phenomena, as the next example illustrates.

4.2.2 Example 2: Predicting an eclipse

The total solar eclipse that occurred in North America on August 21st 2017 (henceforth, E_{2017}) was predicted with a spatial accuracy of 1-3 km, at least in publicaly accessible calculations [38]. This error is mainly due to irregularities in the Moon's surface and, to a lesser extent, to irregularities of the shape of the Earth. Both sources of error can be reduced further with additional information and calculations (and thus a longer τ), but we will limit our analysis to this estimate and therefore assume an average prediction error of 4Km^2.

What is the value of the explanans for this knowledge? The theory component of the explanans consists in calculations based on the JPL DE405 solar system ephemeris, obtained via numerical integration of 33 equations of motion, derived from a total of 21 computations [39]. In the words of the authors, these equations are deemed to be "correct and complete to the level of accuracy of the observational data" [39], which means that this τ can be used for an indefinite number n_Y of computations, suggesting that we can assume $n_\tau/n_Y \approx 0$ and therefore that the weight of τ is negligible.

The input is in this case not a random variable but a sequence, with information content $H(X) = -\log p(x)$. It includes 98 values of initial conditions, physical constants and parameters, measured to up to 20 significant digits, plus 21 auxiliary constants used to correct previous data and the radii of 297 asteroids [39]. Assuming for simplicity that on average these inputs take five digits, we estimate the total information of the input to be $(98 + 21 + 297) \times 5 \times \log 10 \approx 6910$ bits. The accuracy of predictions is primarily determined by the accuracy of these estimations, which moreover are in many cases subject to revision. Therefore, in this case $n_X/n_Y > 0$, and the value of $H(X)$ is less appropriately neglected. Nonetheless, we will again assume for simplicity that $n_Y >> n_X$ and thus $h \approx 1$.

Therefore, since the surface of the Earth is circa $510,072,000\ \text{Km}^2$ we estimate our astronomical knowledge as:

$$K(E_{2017}; X, \tau) \approx \frac{\log(510072000) - \log(4)}{\log(510072000)} = 0.9309 \tag{37}$$

and a re-scaled value of $K(E_{2017}; X, \tau) \times \log(510072000) = 26.9261$.

It may be surprising that the value obtained for predicting eclipses is larger than that obtained for physical constants (section 4.2.1). However, our analysis is not complete. Firstly, note that the assumption of a negligible explanans for the latter is a coarser approximation compared to the former, since physical constant are required to predict eclipses, and not vice-versa. Secondly, and most importantly, our knowledge about eclipses is susceptible to declining with distance between explanans and explanandum. This is in stark contrast to the case of physical constants, which are, by definition, fixed in time and space, such that $\lambda_y \approx 0$.

What is λ_y in the case of eclipses? We will not examine here the possible effects of distance in methods, and we will only estimate the knowledge loss rate over time. We

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

can do so by taking the most distant prediction made for a total solar eclipse to occur in April 26 3000AD [40]. The estimated error is circa 7.8° of longitude, which at the predicted latitude of peak eclipse (21.1N, 18.4W) corresponds to an error of circa 815 Km in either direction. Therefore, the estimated K for predicting an eclipse 982 years from now is:

$$K(E_{3000}; X, \tau) \approx \frac{\log(510072000) - 2\log(815)}{\log(510072000)} = 0.3314 \tag{38}$$

solving $K(E_{3000}; X, \tau) = K(E_{2017}; X, \tau) \times 2^{-\lambda \times 982}$ yields a knowledge loss rate of

$$\lambda_t = 0.0015 \tag{39}$$

per year. Which corresponds to a knowledge half life of $\lambda^{-1} \approx 667$ years. Therefore our knowledge about the position of eclipses, based on the JPL DE405 methodology, is halved for every 667 years of time-distance to predictions.

4.3 How much progress is a research field making?

Problem: Knowledge is a dynamic quantity. Research fields are known to be constatly evolving, splitting and merging [41]. Theories and methodologies are modified, enlarged or simplified, and may be extended to new explananda and explanantia or narrowed down to reduced ones. When do these processes represent progress?

Answer: Progress occurs if an only if the following condition is met:

$$n_X \Delta H(X) - \Delta \log p(\tau) < n_Y H(Y) \frac{k' - k}{kh} \tag{40}$$

in which $\Delta H(X) = H(X') - H(X)$ and $-\Delta \log p(\tau) = -\log p(\tau') - \log p(\tau)$ are expansions or reductions of explanantia, and $\tau' \equiv \tau'_{y|x,x',\tau}$, $k = \frac{H(Y) - H(Y|X,\tau_{y|x})}{H(Y)}$, $k' = \frac{H(Y) - H(Y|X,X',\tau_{Y|X},\tau'_{Y|X,X',\tau})}{H(Y)}$, $h = \frac{n_Y H(Y)}{n_Y H(Y) + n_X H(X) - log p(\tau_{y|x})}$ (see S8 text).

Explanation: Knowledge occurs when prigressively larger explananda are accounted for by relatively smaller explanantia. This is the essence of the process of consilience, which has been recognized for a long time as the fundamental goal of the scientific enterprise [42]. Consilience drives progress at all levels of scientific knowledge. At the research frontier, new research fields are being created by identifying new explananda and/or new combinations of explanandum and explanans for which K is low. Their K grows by a process of "micro-consilience", whereas a "macro-consilience" occurs when knowledge-containing systems are extended and unified across fields, disciplines and entire domains. Equation 40 quantifies the conditions for consilience to occur.

The inequality 40 is satisfied under several conditions. First, when the explanantia X' and/or τ' produce a sufficiently large improvement in the effect, from k to k'. Second, equation 40 is satisfied when $k' \leq k$, if $\Delta H(X) - \Delta \log p(\tau)$ is sufficiently negative, i.e. the input, theory or methodology are being reduced or simplified. Finally, if $\Delta H(X) - \Delta \log p(\tau) \approx 0$, condition 40 is satisfied provided that $k' > k$, which would occur by expansion of the explanandum. In all cases, the conditions for consilience are modulated by the extent of application of the theories themselves, quantified by the n_x and n_y indices.

4.3.1 Example 1: Evolutionary models of reproductive skew

Reproductive skew theory is an ambitious attempt to explain reproductive inequalities within animal societies according to simple principles derived from kin selection theory (see [43] and references within). In its earliest formulation, reproductive skew was

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

predicted to be determined by a "transactional" dynamic between dominant and subordinate individuals, according to the condition

$$p_{min} = \frac{x_s - r(k - x_d)}{k(1 - r)} \tag{41}$$

in which p_{min} is the minimum proportion of reproduction required by the subordinate to stay, x_s and x_d are the number of offspring that the subordinate and dominant, respectively, would produce if breeding independently, r is the genetic relatedness between subordinate and dominant, and k is the productivity of the group. The theory was later expanded to include an alternative "compromise" model approach, in which skew was determined by direct intra-group conflict. Subsequent elaboration of this theory have extended its range of possible conditions and assumptions, leading to a proliferation of models whose overall explanatory value has been increasingly questioned [43].

We can use equation 40 to examine the conditions under which introducing a new parameter or a new model would constitute net progress within reproductive skew theory, using data from a comprehensive review of empirical tests [43]. In particular, we will focus on one of the earliest and most stringent predictions of transactional models, which concerns the correlation between skew and dominant-subordinate genetic relatedness. Contradicting earlier reported success [44], empirical tests in populations of 21 different species unambiguously supported transactional models only in one case (data taken from Table 2.2 in [43]).

Since this analysis is intended as a mere illustration, we will make several simplifying assumptions. First, we will assume that all parameters in the model are measurable to two significant digits, and that their prior expected distributions are uniform (in other words, any group from any species may exhibit a skew and relatedness ranging from 0.00 to 0.99, and individual and group productivities ranging from 0 to 99). Therefore, we assume that each of these parameters has an information content equal to $2 \log 10 = 6.64$. Second, we will assume that the data reported by [43] are a valid estimate of the average success rate of reproductive skew theory in any non-tested species. Third, we will assume that all of the parameters relevant to the theory are measured with no error. For example, we assume that for any organism in which a "success" for the theory is reported, reproductive skew is explained or predicted exactly. Fourth, we will assume that the extent of applications of skew theory, i.e. n_y, is sufficiently large to make the τ component (which contains a description of equation 41 as well as any other condition necessary to make reproductive skew predictions work) negligible. These assumptions make our analysis extremely conservative, leading to an over-estimation of K values.

Indicating with Y, X_s, X_d, X_r, X_k the information values of p_{min}, x_s, x_d, r, k in equation 41, we obtain the value corresponding to the K of transactional models

$$k = \frac{2 \log 10 - \frac{20}{21} 2 \log 10}{2 \log 10} = \frac{1}{21} = 0.048 \tag{42}$$

and

$$h = \frac{y}{y + x_s + x_d + x_r + x_k - \log p(\tau)} = \frac{1}{5 - \frac{\log p(\tau)}{n_y 2 \log 10}} \approx 0.2 \tag{43}$$

Plugging these values in equation 40 and re-arranging, we derive the minimal amount of increase in explanatory power that would justify adding a new parameter or model X':

$$k' > k \left(1 + \frac{n_X h H(X')}{n_Y H(Y)}\right) = 0.048 \left(1 + 0.2 \frac{H(X)}{6.64}\right) \tag{44}$$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

This suggests, for example, that if X' is a new parameter measured to two significant digits, with $H(X) = 2 \log 10$, adding it to equation 41, would represent theoretical progress if $k' > 1.2k$, in other words if it increased the explanatory power of the theory by 20%. If instead X' represented the choice between transactional theory and a new model then, assuming conservatively that $H(X) = 1$, we have $k' > 1.03k$, suggesting that any improvement above 3% would justify it.

Did the introduction of a single "compromise" model represent a valuable extension of transactional theory? The informational cost of expanding transactional theory consists not only in the equations τ' that need to be added to the theory, but also in the additional binary choice X' that will need to be made between the two models for each new species to which the theory is applied. The latter condition entails $n_x H(X) = n_y$. According to Nonacs et al [43], compromise models were successfully tested in 2 out of the 21 species examined. Therefore, the $k = 3/21 = 0.14$ attained by adding a compromise model amply compensated for the corresponding increased complexity of reproductive skew theory.

The analysis above refers to results for tests of reproductive skew theory across groups within populations. When comparing the average skew of populations, conversely, transactional models were compatible with virtually all of the species tested, especially with regards to the association of relatedness with reproductive skew. In this case, assuming that $k \approx 1$, i.e. that transactional models are compatible with every species encountered, then progress within the theory could only be achieved by simplifying equation 41. This could be obtained by removing or recoding the parameters with the lowest predictive power, or by deriving the theory in question from more general theories. The latter is what the authors of the review did, by suggesting that the cross-population success of the theory is explainable more economically in terms of kin selection theory, from which these models are derived [43].

These results, are merely preliminary and likely to over-estimate the benefits of expanding skew theory. In addition to the conservative assumptions made above, we have assumed that only one transactional model and one compromise model exist, whereas in reality several variants of these models have been produced, which entails that the choice X' is larger than binary. Moreover, we have assumed that the choice between transactional and compromise models is made *a priori*, for example based on some binary property of organisms. If the choice is made *after* the variables are known then the costs of this choice have to be accounted for, with potentially disastrous consequences (see section 4.6).

4.3.2 Example 2: Gender differences in personality factors

In 2005, psychologist Janet Hyde proposed a "gender similarity hypothesis", according to which men and women are more similar than different on most (but not all) psychological variables [45]. According to her review of the literature, human males and females exhibit average differences that, for most measured personality factors, are of small magnitude (i.e. Cohen's d ≤ 0.35). Assuming that these traits are normally distributed within each gender, this finding implies that the empirical distributions of male and female personality factors overlap by more than 85% in most cases.

The gender similarity hypothesis was challenged by Del Giudice et al [46], on the basis that, even assuming that the distributions of individual personality factors do overlap substantially, the joint distribution of these factors might not. For example, if Mahalanobis distance D, which is the multivariate equivalent of Cohen's d, was applied to 15 psychological factors measured on a large sample of adult males and females, the resulting effect was large (D=1.49), suggesting an overlap of 30% or less [46] (see Figure 6a).

The multivariate approach proposed by Del Giudice was criticized by Hyde primarily

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

for being "uninterpretable" [47], because it is based on a distance in 15-dimensional space, calculated from the discriminant function. This suggests that such a measure is intended to maximze the difference between groups. Indeed, Malanobis' D will always be larger than the largest uni-dimensional Cohen's d included in its calculation (see Figure 6a).

The K function offers an alternative approach to examine the gender differences vs. similarities controversy, using simple and intuitive calculations. With K, we can quantify directly the amount of knowledge that we gain, on average, about an individual's personality by knowing their gender. Since most people self-identify as male and female in roughly similar proportions, knowing the gender of an individual corresponds to an input of one bit. In the most informative scenario, males and females would be entirely separated groups along any given dimension, and knowing gender would return exactly one bit along any dimension. Therefore, we can test to what extent the gender factor is informative by setting up a one-bit uncertainty in the explanandum: we divide the population in two groups, above and below the median for each dimension.

Since the explanandum may consist more than one dimension, we can calculate the ensuing knowledge under two scenarios: we may assume to have no knowledge in (or interest in discounting) any possible correlations between personality factors, or assume that that structure is known and that we wish to discount it from K. The former scenario, which will call "multi-dimensional K", is psychologically realistic and intuitively interpretable, and is calculated as

$$K_{md} \equiv \frac{\sum_{i=1}^{z} H(Y_i) - \sum_{i=1}^{z} H(Y_i|X, \tau_{Y_i|X})}{\sum_{i=1}^{z} H(Y_i) + H(X) - \sum_{i=1}^{z} \log p(\tau_{Y_i|X}) \frac{1}{n_Y}} \tag{45}$$

in which z is the number of dimensions considered and $\tau_{Y_i|X}$ is the theory linking gender to each dimension i.

The latter scenario, which we will call "multi-variate K", quantifies the "net" relevance of gender to explaining overall personality differences, once all other structures in the data are accounted for, whether or not they are knowable by individuals. Since people are unlikely to carry accurate information on the cross-correlations between personality dimensions, this measure is psychologically less interpretable. Moreover, its value will strictly depend on which variables are included or not in the measure. However, this is also an objectively more accurate estimate of the absolute relevance of the gender variable to a data set, once any other structure in the data is controlled for. Its value is given by:

$$K_{mv} \equiv \frac{H(Y_z, Y_{z-1}...Y_1|\tau_{Y_z|Y_{z-1}...Y_1}) - H(Y_z, Y_{z-1}...Y_1|X, \tau_{Y_z,Y_{z-1}...Y_1|X})}{H(Y_z, Y_{z-1}...Y_1|\tau_{Y_z|Y_{z-1}...Y_1}) + H(X) - \log p(\tau_{Y_z,Y_{z-1}...Y_1|X}) \frac{1}{n_Y}} \tag{46}$$

in which $H(Y_z, Y_{z-1}...Y_1|\tau_{Y_z|Y_{z-1}...Y_1})$ is the joint entropy of the $z-$dimensional explanandum, the structure of which is determined by a theory $\tau_{Y_z|Y_{z-1}...Y_1}$ that is computed prior and independently of the explanans of interest.

Note that K_{md} is just a special case of K_{mv}, and that, whereas the maximum value attainable by the unidimensional K is $1/2$, that of K_{md} (and therefore K_{mv}) is $15/16 = 0.938$. This value illustrates how, as the explanandum is expanded to new dimensions, K_{md} could approach indefinitely the value of 1, value that would entail that input about gender yields complete information about personality. Whether it does so, and therefore the extent to which applying the concept of gender to multiple dimensions represent progress, is determined by conditions in 40.

To illustrate the potential applications of these measures, the values of K, average K, as well as K_{md} and K_{mv} were calculated from a data set (N=10^6) simulated using data on the variance and covariance of personality factors estimated by [46, 48]. All unidimensional personality measures were split in lower and upper 50% percentile,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

yielding one bit of potentially knowable information. In K_{md}, these were then recombined, yielding a 15-bit total explanandum. In K_{mv}, the internal variance-covariance structure reduced the entropy of the explanandum to 13.8 bits.

Figure 6b reports results of this analysis. As expected, the unidimensional K values are closely correlated with their corresponding Cohen's d values (Figure 6 a and b, black bars). However, the multidimensional K values offer a rather different picture from that of Malanobis D. In particular, both K_{md} and K_{mv} are considerably smaller than the largest unidimensional effect measured, and are in the range of the second-largest effect. They are, therefore, somewhat intermediate in magnitude, although larger than a simple average (given by the orange bar in Figure 6b).

Therefore, we conclude that the overall knowledge conferred by gender about the 15 personality factors together is comparable to some of the larger, but not the largest, values obtained on individual factors. This is a more directly interpretable comparison of effects, which stems from the unique properties of K.

We can also calculate the absolute number of bits that are gained about an individual's personality by knowing a person's gender. For the unidimensional variables, where we assumed $H(Y) = 1$, this is equivalent to the K values shown. For the multidimensional and multivariate Ks, however, we have to multiply by 15 and 13.84, respectively, obtaining 0.26 and 0.19, respectively (Figure 6b). These values are considerably larger that the unidimensional ones, and suggest that, at least among the 15 dimensions considered, receiving the one-bit input about an individual's gender allows to save, respectively, at least 1/4 and 1/5 of a bit in predicting their personality.

These results are intended as mere illustrations of the potential utility of the methods proposed. Such potential was under-exploited in this particular case, because the original data was not available, and therefore our analyses were based on a re-simulation of the data, based on estimated variances and co-variances. Therefore, our analysis inherited the assumptions of normality and linear covariance that are necessary but limiting components of traditional multivariate analyses, and were a source of criticism for data on gender differences too [47].

Unlike ordinary multivariate analyses, a K analysis requires no distributional assumptions. If it were conducted on a real data set about gender, the analysis might reveal non-linear structures in personality factors, and/or identify the optimal level of resolution at which each dimension of personality ought to be measured (see section 3.3.6). This would yield a more accurate answer concerning how much knowledge about people's personality is gained by knowing their gender.

4.3.3 Example 3: Does cumulative evidence support a hypothesis?

The current tool of choice to assess whether the aggregate evidence of multiple studies supports an empirical hypothesis is meta-analysis, in which effect sizes of primary studies are standardized and pooled in a weighted summary [9]. The K function may offer a complementary tool in the form of a cumulative K_{cum}. This is conceptually analogous to the K_{md} described above but, instead of assuming that the various composing explananda lie on orthogonal dimensions and the explanans is fixed, it assumes that both explanandum and explanans lie on a single dimension, and their entropy results from a mixture of different sources.

It can be shown that, for a set of RV $Y_1, Y_2 ... Y_m$ with probability distributions $p_{Y_1}(\cdot), p_{Y_2}(\cdot) ... p_{Y_m}(\cdot)$, the entropy of their mixed distribution $\sum w_i p_{Y_i}$ is given by:

$$H(\sum_{i \leq m} w_i p_{Y_i}) = \sum_{i \leq m} w_i H(Y_i) + \sum_{i \leq m} w_i D(p_{Y_i} || \sum_{i \leq m} w_i p_{Y_i}) \equiv \overline{H(Y)} + \overline{d_Y} \qquad (47)$$

where the right-hand terms are a notation introduced for convenience, and

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

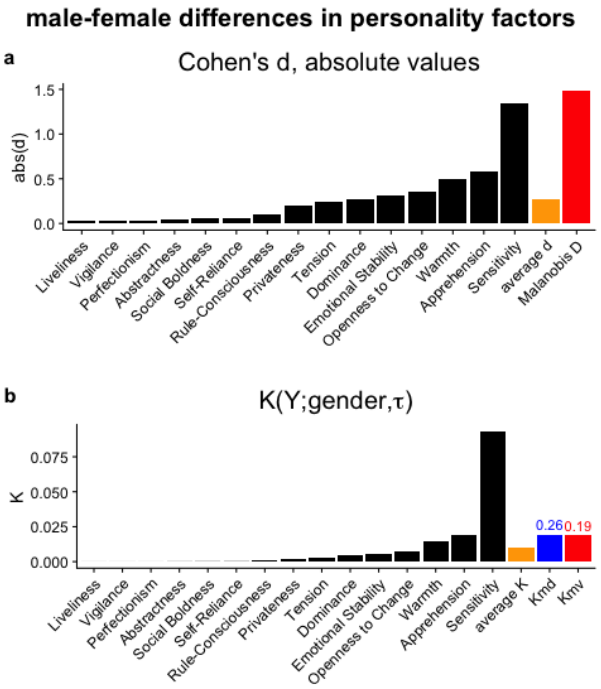


Figure 6. Uni- and multivariate analyses of gender differences in personality factors. a) Cohen's d values and Malanobis D calculated in [46]. b) K values calculated on a data set of one million individuals, reproduced using the covariance matrices for males and females estimated in [46]. Orange bar: average of uni-dimensional K values. Blue bar: multidimensional values (K_{md}), calculated assuming that all factors are orthogonal, calculated as in equation 45. Red bar: multivariate K (Kmv) calculated on the joint entropy of the 15 personality factors combined (equation 46). The numbers above the blue and red bars represent the re-scaled values $K_{md}H(Y)$ and $K_{mv}H(Y)$, respectively. For further details, see text.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

$D(p_{Y_i}||\sum_{i\leq m} w_i p_{Y_i})$ represents the Kullback-Leibler divergence between each RV and the mixed distribution.

For sequences, and particularly for those representing the theory τ, we can define a mixing process as follows. For each composing uniform RV T_i, the mixing with one or more distributions T_j results in a uniform RV whose alphabet is the union set of the mixed alphabets $\mathcal{T} = \{\mathcal{T}_i \cup \mathcal{T}_j\}$. It can then be shown that, if for example $\tau_i = (\tau_{i,1}, \tau_{i,2}...\tau_{i,l})$ and $\tau_j = (\tau_{j,1}, \tau_{j,2}...\tau_{j,m})$ are two sequences of length l and m with $l > m$, their mixture will yield the quantity

$$\overline{\tau} + \overline{d_\tau} \equiv l + \sum_{u\leq l} \log \frac{|\mathcal{T}_u|}{2} \tag{48}$$

in which $|\mathcal{T}_u|$ is the size of the alphabet resulting from the mixture. For the mixing of s theories $\{\tau_1, \tau_2...\tau_s\}$, $\bar{\tau}$ will be equal to the longest *binary* description length of the set, l^* say, and

$$0 \leq \overline{d_\tau} \leq l^* \log \frac{s+2}{2} \tag{49}$$

with the latter condition occurring if all the elements of all the s sequences are different from each other.

For example, if the methodology $\tau_i = (\text{"}randomized\text{"}, \text{"}human\text{"}, \text{"}female\text{"})$ were mixed with $\tau_j = (\text{"}randomized\text{"}, \text{"}human\text{"}, \text{"}male + female\text{"})$, the resulting mixture would have composing RV $T_1 = \{\text{"}randomized\text{"}, \text{"}not\text{"}\}$, $T_2 = \{\text{"}humansubject\text{"}, \text{"}not\text{"}\}$, $T_3 = \{\text{"}female\text{"}, \text{"}male + female\text{"}, \text{"}not\text{"}\}$, and its information content would equal $-log(2) - log(2) - log(3) = 3.58$ or equivalently $\overline{\tau} + \overline{d_\tau} = 3 + log(3/2) = 3 + 0.58$.

Therefore, the value of the cumulative K is given by

$$K_{cum} \equiv \frac{n_Y \left(\overline{H(Y)} - \overline{H(Y|X,\tau)} + \overline{d_Y} - \overline{d_{Y|X,\tau}}\right)}{n_Y \overline{H(Y)} + n_X \overline{H(X)} + \overline{\tau} + n_Y \overline{d_Y} + n_X \overline{d_X} + \overline{d_\tau}} \tag{50}$$

in which the $\bar{d}$ terms represent the average divergences from the mixed expananda or explanatia. Equation 50 is subject to the same conditions of equation 40, which will determine whether the cumulative knowledge (e.g. a cumulative literature) is overall leading to an increase or a decrease of knowledge.

The peculiarity of equation 50, however, lies in the presence of additional divergence terms, which allow knowledge to grow or decrease independently of the weighted averages of the measured effects. In particular,

$$K_{cum} \geq \overline{K} \iff \overline{d_{Y|X,\tau}} \leq (1 - \overline{K})\overline{d_Y} - \overline{K}(\overline{d_X} + \overline{d_\tau}) \tag{51}$$

with $\overline{K} = \frac{\overline{H(Y)} - \overline{H(Y|X,\tau)}}{\overline{H(Y)} + \overline{H(X)} + \overline{\tau}}$ constituting the K value obtained by the simple averages of each term. This property, combined with the presence of a cumulative hardness component that penalizes the cumulation of diverse methodologies, make K_{cum} rather different from ordinary meta-analytical summaries.

Figure 7 illustrates the differences between meta-analysis and K_{cum}. Like ordinary meta-analysis, K_{cum} depends on the within and between-study variance of effect sizes. However, K_{cum} also decreases if the methodology of aggregated studies is heterogeneous, independent of the heterogeneity of effect sizes. Moreover, K_{cum} can increase even when all included studies report null findings, if the aggregated studies cover different ranges of the explanandum, making the cumulative explanandum larger.

Note that we have not specified how the weights underlying the mixture are calculated. These may consist in an inverse-variance weighting, as in ordinary

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

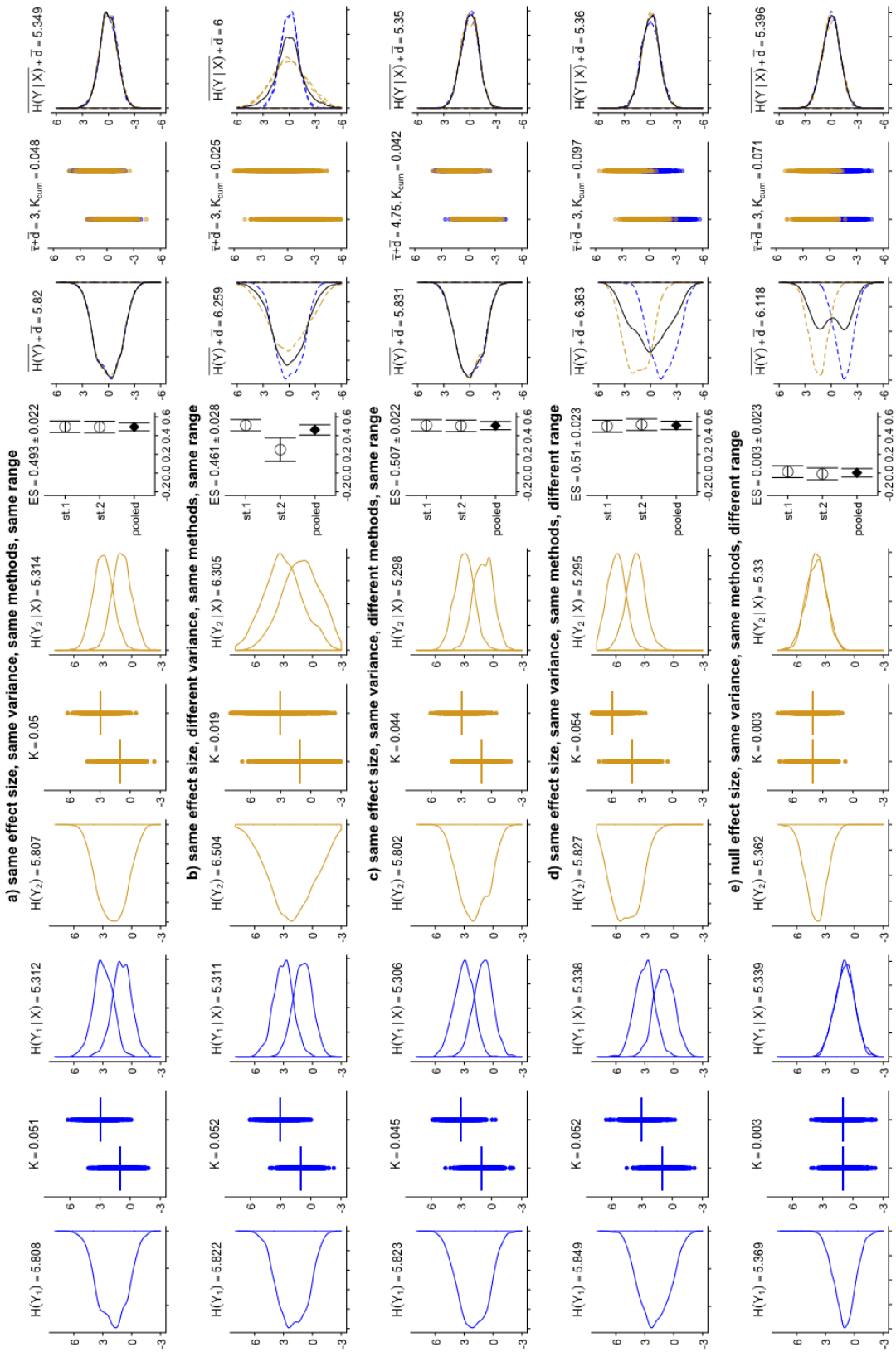


Figure 7. Comparison between meta-analysis and cumulative K analysis. From left to right, the graph shows the simulated data for two imagined studies (blue and golden, respectively), with different assumptions of variance and effect size, then the corresponding meta-analytical summary, and then the corresponding K analysis, with values calculated as in equation 50. Above each figure is indicated the corresponding entropy value, meta-analytical summary effect size, or K value. See text for further discussion.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

meta-analysis, or could be computed based on other epistemologically-relevant variables, such as the relative divergence of studies' methodologies. The latter approach would offer an alternative to the practice of weighting studies by measures of quality, a practice that has largely been abandoned in meta-analysis due to its inherent subjectivity.

4.4 How reproducible is a research finding?

Problem: The concept of "reproducibility" is the subject of growing concerns and expanding research programs, both of which risk being misled by epistemological confusions of at least two kinds. The first source of confusion, is the conflation of the reproducibility of methods and that of results [2]. The former entails that identical results are reproduced if the same data is used, indicating that methods were reported completely and transparently. The latter entails that identical results are obtained if the same methods are applied to new data. Whereas the former is a relatively straightforward issue to assess and to address, the latter is a complex phenomenon that has multiple causes that are hard to disentangle.

The second source of confusion comes from treating the concept of reproducibility as a dichotomy - either a study is reproducible/reproduced or it is not - even though this is obviously a simplification. The extent to which a study's methods or results are reproducible/reproduced is determined not solely by how the research was conducted and reported, but also by contingent characteristics of the research's subject matter and general methodology. How can we distinguish the reproducibility of methods and results and define them in a single, continuous measure?

Answer: The reproducibility of a finding is determined by the relation

$$K_r = KA^{-\boldsymbol{\lambda}\cdot\boldsymbol{d}} \tag{52}$$

in which K_r is the result of a replication study conducted at a study-specific "distance" given by the inner-product of a vector $\boldsymbol{d} : [s, x, \tau]$ of distances and a vector $\lambda : [\lambda_s, \lambda_x, \lambda_\tau]$ of corresponding loss rates.

Explanation: Section 3.3.5 has show how the exponential function in equation 52 describes any amount of decline of a system's K that is due to divergences in subject matter or methodology. In practical terms, a divergence vector will consist in classifiable, countable differences in components of the methods used and/or characteristics of subject matter that, based on theory and prior data, are deemed likely to change the nature of the phenomenon being studied by an amount λ.

Applications of equation 52 to individual cases require measuring study-specific divergences in explanandum and explanans and their corresponding and field-specific loss rates. However, the universality of the function in equation 52 allows to derive general, population-level predictions about reproducibility, as the example below illustrates.

Example: How reproducible is Psychological Science?

The Reproducibility Initiative in Psychology was a monumental project in which a consortium of laboratories attempted to replicate 100 studies taken from recent issues of three main psychology journals. Results were widely reported in the literature and mass media as suggesting that less than 40% of studies had been replicated, a figure deemed to be disappointingly low and indicative of significant research and publication bias in the original studies [49]. This conclusion, however, was questioned on various grounds, including: limitations in current statistical approaches used to predict and estimate reproducibility (e.g. [50–52]), methodological differences between original and replication studies [53], variable expertise of the replicators [54], and variable contextual

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

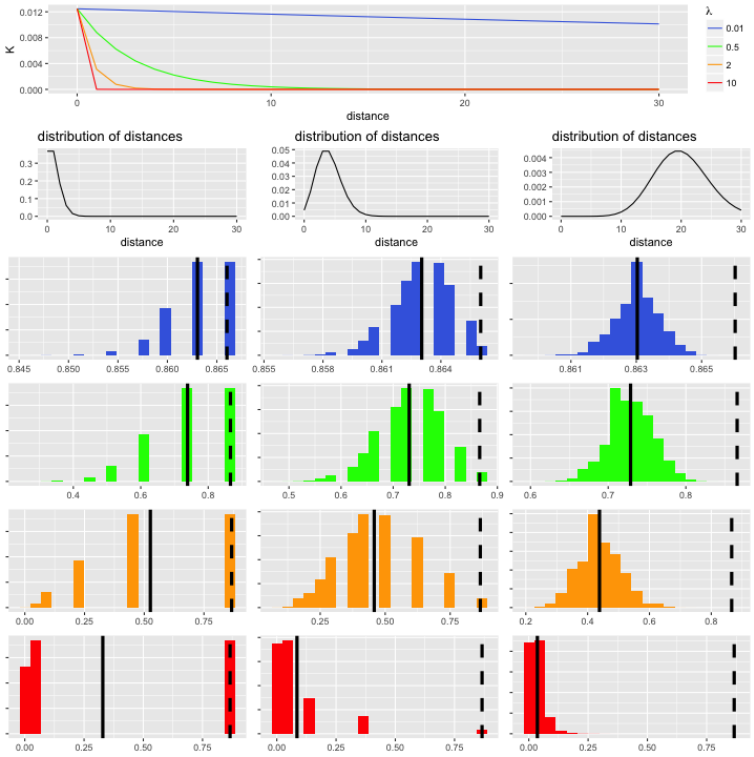


Figure 8. Distribution of results of reproducibility studies, under varying conditions of distances (i.e. number of differences in methodologies) d and their average impact λ. The top panel shows how K declines, as the number of divergences increases, depending on different values of λ. Panels in the second row show the probability distribution of the simulated distances (i.e. Poisson distributions, with mean 1, 5 and 20, respectively). The nine panels below show the distribution of correlation coefficients of reproducibility studies under each combination of number of distances and their impact (the impact is colour coded as in the top panel. For further discussion see text.

sensitivity of the phenomena studied [55, 56]. The common concern of these criticisms, therefore, is a possible divergence in methods and/or phenomena between the original study and its replication.

In theory, each replication study in the RIP could be examined individually using equation 52, but doing so would require field-specific information on the impact of various divergences, data that at present are not available. However, we can use equation 52 to formulate a general prediction about *the shape of the distribution* of results of a reproducibility study, under varying frequencies and impacts of errors.

Figure 8 simulated the distribution of effect sizes (here shown as correlation coefficients derived from the corresponding K) that would be observed in a set of replication studies, depending on their average distances d and impacts λ from an original or ideal study. Distances were assumed to follow a Poisson distribution, with a mean of 1,5 and 20, respectively. The expected impact of these distances was increased moving from the top to the bottom row, with values of λ whose effects are illustrated in the top-most panel. The dotted vertical line in each plot reports the initial value of K (i.e. the left-hand side of equation 52), whereas the solid line shows the mean of the distribution of results.

The figure can be given different interpretations. The distances simulated in Figure 8 may be interpreted as between-study differences in the explanandum or input (e.g.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

cultural differences in the studied populations), between-study differences in methodological choices, or as study-specific methodological errors and omissions, or a combination of all three. The dotted line may represent either the result of the original study or the effect that would be obtained by an ideal study (i.e. a study whose combination of Y, X, τ maximize the K attainable) that is never realised, from which all observed studies are at some distance.

Irrespective of what we assume distances to represent, results suggest that, when distances are few and of minor impact, the distribution of results is compact and right-skewed (top-left). As when the number of such minor-impact distances grows, the distribution tends to be symmetrical and Gaussian in appearance (top-right). However, as the impact of distances increases in magnitude, the distribution tends to become left-skewed if distances are numerous (bottom-right) or bimodal (bottom-left) if few.

This suggests that the conditions typically postulated in analyses of reproducibility (i.e. a normal distribution around the "true" or the "average" effect in a population of studies) are only realized when errors or omissions are numerous and of minor importance. However, when important divergences in explanandum or explanans occur, the distribution becomes increasingly asymmetrical and concentrated around null results, and may either be left-skewed or bimodal, depending on whether the likelihood of such errors is large or small.

Data from the RIP supports these predictions. Before being replicated, studies had been classified by RIP authors based on the level of expertise required to replicate them. As figure 9 illustrates, replication results of studies that were deemed to require moderate or higher expertise are highly concentrated around zero, with a small subset of studies exhibiting medium to large effects. This distribution is markedly different from that of studies that required null or minimal expertise, which was close to normal. Note how the distribution of original results reported by both categories of studies are, instead, undistinguishable in shape. Additional differences between distributions might be explained by a similar classification concerning the stability and/or distances of the explanandum or explanans (e.g. the contextual sensitivity suggested by [56]).

Although preliminary, these results suggest that a significant cause of reproducibility "failures" in the RIP may have been high-impact divergences in the systems or methodologies employed by the replicating studies. These divergences may have occurred despite the fact that many authors of the original studies had contributed to the design of the replication attempts. A significant component of a scientists' expertise consists in "tacit knowledge" [57], which are correct methodological choices that are not codified or described in textbooks and research articles, and that are unconsciously acquired by researchers through practice. Therefore, authors of the original studies might have given for granted, or unwittingly overlooked, important aspects of their own research design when instructing the RIP replicators. The latter, even if professionally prepared, might have lacked sufficient expertise about the systems object of the replication attempt, and may therefore have made "tacit" errors that neither they or the authors of the original studies were able to document.

4.5 What is the value of a null or negative result?

Problem: How scientists should handle "null" and "negative" results is the subject of considerable ambiguity and debate. On the one hand, and contrary to what their names might suggest, "null" and "negative" results undoubtedly play an important role in scientific progress, because it is by cumulation of such results that hypotheses and theories are refuted, allowing progress to be made by "falsification", rather than verification, as Karl Popper famously argued [58]. Null and negative results are especially important in contexts in which multiple independent results are aggregated to test a single hypothesis, as is done in meta-analysis [5].

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

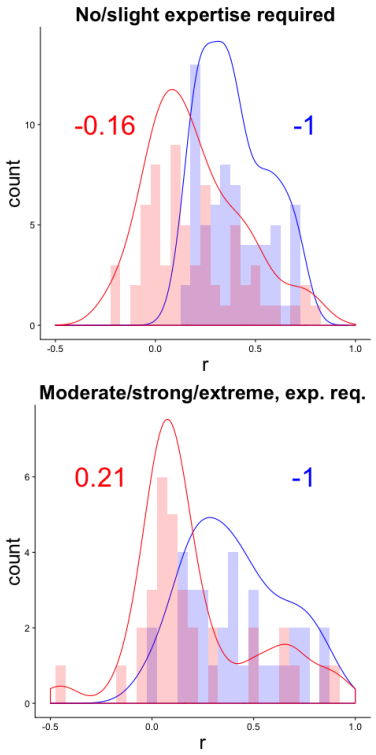


Figure 9. Distributions of correlation coefficients reported by the studies examined in the Reproducibility Initiative in Psychology [49]. Blue: effect sizes originally reported. Red: results of replications. Numbers report the kurtosis of each distribution.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

On the other hand, as Popper himself had noticed, the falsifiability of a hypothesis is typically suboptimal, because multiple "auxiliary" assumptions (or, equivalently, auxiliary hypotheses) may not be controlled for. Moreover, it is intuitively clear that a scientific "discovery" that leads to useful knowledge is made when a pattern is identified, and not merely when a pattern is proved not to subsist.

This is why, if on the one hand there are increasing efforts to counter the "file-drawer problem", on the other hand there are legitimate concerns that these efforts might generate a "cluttered office" problem, in which valuable knowledge is drowned in a noisy plethora of useless null results [59]. The problem is that the value of null and negative results is context-specific. How can we estimate it?

Answer: The knowledge value of a null or negative result is given by

$$K_{null} \leq \frac{h}{H(Y)} \log \frac{|\mathcal{T}|}{|\mathcal{T}| - 1} \tag{53}$$

in which K_{null} is the knowledge gained by the conclusive refutation of a hypothesis, and $|\mathcal{T}|$ is the size of the set of hypotheses being potentially tested (i.e. not controlled for) in the study.

Explanation: Section 3.2.1 described knowledge as resulting from the selection of a $\tau \in \mathcal{T}$, where $\mathcal{T}$ is the a set of possible theories (methodologies) determining a pattern between explanandum and input. These theories have probability distribution $P_T(\tau)$, and are therefore described by the RV T. It can be shown that, because of the symmetry property of the mutual information function

$$K(Y; X, T) = K(T; Y, X) \tag{54}$$

i.e. the information that the set of theories contains about the data is equivalent to the information that the data contains about the theories (see S9 text).

This is indeed how knowledge is attained. A theory τ is selected amongst available alternatives because it best fits an input data Y^{n_Y}, X^{n_X}, and ideally maximizes $k_{adj} - k_{obs}$ (see section 3.3.2). The data is obtained by experiment (or experiences) and the process is what we call learning, as it is embodied in the logic of Bayes' theorem, the MDL principle and generally the objective of any statistical inference method. Since no knowledge, including knowledge about a theory, can be obtained in the absence of a "background" conditioning theory and methodology, a more accurate representation of an experiment entails the specification of an unvarying component m, which quantifies the aspects of the theory and methodology of an experiment that are not subject to uncertainty, and the component for which knowledge is sought, the random variable T, which therefore represents the hypothesis or hypotheses being tested by the experiment. The knowledge attained by the experiment is then given by

$$K(T; Y^{n_Y}, X^{n_X}, m) = \frac{h}{H(Y)}(H(T) - H(T|Y, X, m)) \tag{55}$$

It follows that the experiment is maximally informative when $H(T)$ is as large as possible and $H(T|Y, X, m) = 0$, the latter condition stating that each possible state of T is in a one-to-one correspondence with each of possible state of data Y, X.

Real-life experiments depart from this ideal condition in two ways. First, they usually retain uncertainty about the result, $H(T|Y, X, m) > 0$, because multiple alternative hypotheses are compatible with the same experimental outcome. Second, real experiments usually test no more than one hypothesis at a time. This entails that $H(T|Y, X, m)$ rapidly approaches $H(T)$, as the size of the alphabet of T, increases (see S10 text). These limitations suggest that, assuming maximally informative conditions in which all tested hypotheses are equally likely and one hypothesis is conclusively ruled

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

out by the experiment, we have $H(T) - H(T|Y = y, X = x, m) = \log|\mathcal{T}| - \log(|\mathcal{T}| - 1)$, giving equation 53.

As intuition would suggest, even if perfectly conclusive, a null finding is intrinsically less valuable than its corresponding "positive" one. This occurs because a tested hypothesis is ruled out when the result is positive as well as negative, with $K(T; Y, X, m, T = \tau_1) = K(T; Y, X, m, T = \tau_0)$, and therefore the gain in equation 53 occurs with positive as well as negative results. Positive results, however, also yield knowledge about a pattern. Therefore, whereas a conclusive rejection of a specific pattern in favour of a null yields at most $K(T; Y, X, m, T = \tau_0) = \frac{h}{H(Y)}$, a conclusive rejection of the null hypothesis in favour of the alternative yields $K(T; Y, X, m, T = \tau_1) + K(Y, X, \tau_1) > \frac{h}{H(Y)}$. Perfect symmetry between "negative" and "positive" results is only attained in the ideal conditions mentioned above, in which $H(T|Y, X, m) = 0$ and $H(T) = H(Y)$, and therefore each experimental outcome identifies a theory with empirical value and at the same time refutes other theories. This is the scenario in which "perfect" Popperian falsificationism can operate, and real-life experiments depart from this ideal in proportion to the number $\log(|\mathcal{T}| - 1)$ of auxiliary hypotheses that are not addressed by the experiment.

The departure from ideal conditions is especially problematic in biological and social studies that are testing a fixed "null" hypothesis against a non-specified alternative τ_0 that predicts $K(Y; X, \tau_0) = 0$ as opposed to a generic non-null τ_1 for which $K(Y; X, \tau_1) > 0$. First of all, due to noise and limited sample size, $P_T(\tau_0|Y^{n_Y}, X^{n_X}, m) > 0$. This problem can be substantially reduced by increasing statistical power but can never be fully eliminated, especially in fields in which large sample sizes and high accuracy (resolution) are difficult or impossible to obtain. Moreover, and *regardless of statistical power*, a null hypothesis is inherently more likely to be compatible with multiple "auxiliary" hypotheses, which real-life experiments may be unable to control.

Example: A simulation

To offer a practical example of the theoretical argument made above, Figure 10 reports a simulation. The value of $K(T; X, Y)$, i.e. the informativeness of data for a given hypothesis, was first calculated when a single hypothesis h_1 is at stake, and all other conditions are fixed (Figure 10, top). Subsequently, the alphabet of T (the set of hypotheses in the experiment) was expanded to include a second condition, with two possible states τ_a or τ_b, the former of which produces a null finding regardless of h_1. The state of this latter hypothesis remains undetermined in the experiment. The corresponding value of $K(T; X, Y)$ is measurably lower, even if re-scaled to account for the greater complexity of the explanandum (i.e. the number of tested hypotheses, Figure 10, bottom).

This is a simple illustration of how the value of negative results depends on the number of uncontrolled conditions and/or possible hypotheses. If field-specific methods to estimate the number of auxiliary hypotheses are developed, the field-specific and study-specific informativeness of a null result could be estimated and compared.

The conclusions reached in this section, combined with the limitations of replication studies discussed in section 4.4, may yield new insights into debates over the problem of publication bias and how to solve it. This aspect is briefly discussed in the example below.

Example: Should we publish all negative results?

The literature of the biological and social sciences is rife with articles debating whether publication bias is "a bane or boon in disguise". A vivid example was offered by two

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

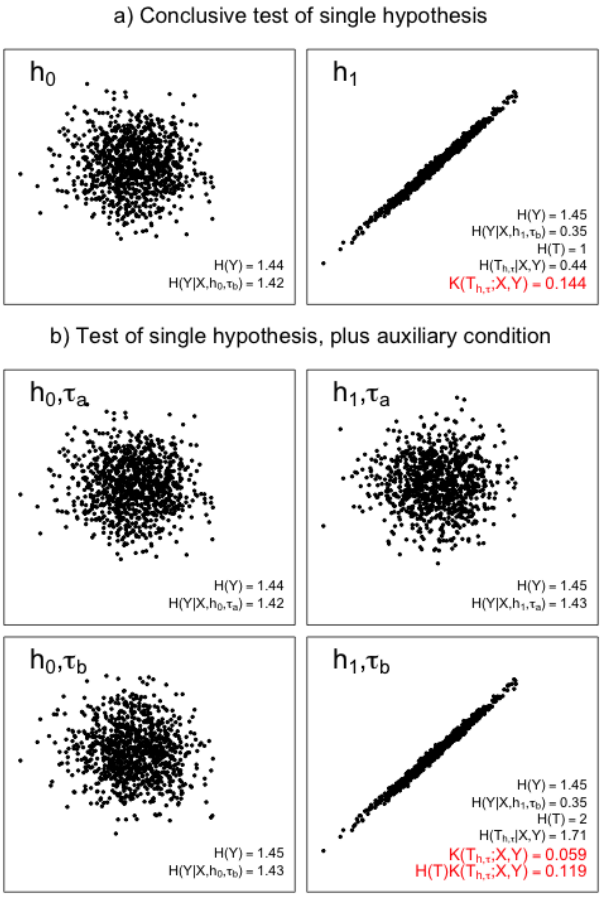


Figure 10. K analysis of the informativeness of data with regards to a hypothesis h, in absence (top) or presence (bottom) of a second condition τ that modulates results of the test. Numbers report all parameters calculated from the analysis. The R code to generate the data and figure is in S11 text. See text for further details and discussion.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

recent studies that used virtually identical methods and arguments but reached opposite conclusions concerning whether "publishing everything is more effective than selective publishing of statistically significant results" [60, 61].

Who is right? Both perspectives may be right or wrong, depending on specific conditions of a field, i.e. of a research question and a methodology. An overlooked assumption behind most analyses of publication bias is that the primary studies subjected to bias are of "similar quality". What this quality specifically consists in is never defined, and its field-specific distribution is probably never homogeneous, and likely to vary in depending on field-specific characteristics. This field-specific heterogeneity, however, cannot be overlooked, because it determines the extent to which $H(T|Y, X, m) > 0$, i.e. the falsifiability of the main hypothesis being tested. Therefore, to properly estimate the true prevalence and impact of publication bias, and determine cost-effective solutions, the falsifiability of hypotheses needs to be estimated on a case-by-case (i.e. field-specific or methodology-specific) basis.

In general, the analysis above suggests that current concerns for publication bias and investments to counter it are most justified in fields in which methodologies are well codified and hypotheses to be tested are simple and clearly defined. This is likely to be the condition of most physical sciences, in which not coincidentally negative results appear to be valued as much or more than positive results [62, 63]. It may also reflect the condition of research in clinical medicine, in which clearly identified hypotheses (treatments) are tested with relatively well-codified methods (randomized controlled trials). However, the value of negative results in other research fields ought to be assessed on a case by case basis.

Methods proposed in this article might help us determine relevant field-specific and study-specific conditions. In particular, the statistical relevance of a null result produced by a study with regards to a specified hypothesis is likely to be inversely proportional to the expected divergence of the study from a standard (or an ideal) methodology and explanandum $\lambda \cdot d$ (section 4.4). This effect is in turn modulated by the complexity and flexibility of a field's methodological choices and magnitude of effect sizes, both quantifiable in terms of the K function proposed in this study.

4.6 How much knowledge do we lose from bias, misconduct and QRPs?

Problem: In addition to relatively well-defined forms of scientific misconduct, studies and policies about research integrity typically address a broader category of "Questionable Research Practices" (QRP). This is a class of rather loosely defined behaviours such as "dropping outliers based on a feeling that they were inaccurate", "failing to publish results that contracted one's previous findings" that, by definition, may or may not be improper, depending on the context [64].

Since QRP are likely to be more frequent than outright fraud, it has long been argued that their impact on the reliability of the literature may be very high - possibly even higher than that of data fabrication or falsification (e.g. [65]). However, besides obvious difficulties in quantifying the relative frequency of proper versus improper QRPs, there is little epistemological or methodological basis for grouping together an extremely heterogeneous set of practices and brand them as equally worrying [4]. Setting aside practices that do not affect the validity of results - which will not be considered here - it is obvious that our concerns for QRPs ought to be proportional to the frequency of their improper use and to the effect that this has in distorting the literature. How can we quantify the impact of misconduct and QRPs?

Answer: The impact on knowledge of a Questionable Research Practice is given by a "bias-corrected" K value

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

$$K_{corr} = K - \frac{h_u}{h_b}B \tag{56}$$

in which $h_u = n_Y H(Y)/(n_Y H(Y) + n_X H(X) - \log p(\tau))$ and $h_b = n_Y H(Y)/(n_Y H(Y) + n_X H(X) - \log p(\tau) - n_\beta H(B))$ are the the hardness terms for the study, without and with bias, respectively, and

$$B = \frac{D(Y|X,\tau||Y|X,\tau,\beta)}{n_Y H(Y) + n_X H(X) - logp(\tau)} \tag{57}$$

is the bias caused by the practice.

Explanation: Equation 56 is derived by a similar logic to that of predictive success, discussed in section 3.3.2. If a research practice is deemed epistemologically improper, that is because it must introduce a bias in the result. This implies that the claim made using the biased practice β is different from the claim that is declared or intended: $K(Y;X,\tau,\beta) \neq K(Y;X,\tau)$. Just as in the case of prediction costs, therefore, we can adjust the K value by subtracting from it the costs required to derive the claimed result from the observed one, costs that are here quantified by B (see equation 22).

Differently from the latter case, however, in the presence of bias the methods employed are of different size. In particular, the biased research has required an additional methodology β, which should not be present in the unbiased research (and/or is not declared in the research article). Following our standard approach, we posit that β is an element of the alphabet of a uniform random variable B. The entropy $H(B)$ represents the set of possible choices that are made in the QRP, and n_β will be the number of times these choices have to be made in the research. For example, a biased research design will have $n_\beta = 1$, and therefore a cost $-\log p(\beta)$ to be added to those of τ. Conversely, if the bias is a post-hoc manipulation of some data or variables, then n_β may be as high as n_Y or even higher. The term $\frac{h_u}{h_b}$ quantifies the relative costs of the biased methodology.

An important property of equation 56 is that the condition holds regardless of the direction of the bias. The term B is always non-negative, independent of how results are shifted. Therefore, a QRP that nullified an otherwise large effect (in other words, a bias against a positive result) would require a downwards correction just as one that magnified it.

4.6.1 Example 1: knowledge cost of data fabrication

The act of fabricating an entire study, its dataset, methods, analysis and results, can be considered an extreme form of ante-hoc bias, in which the claim of an effect was generated by the methods.

Let β represent the method that fabricated the entire study. Clearly, the effect observed without that method is zero, yielding

$$K_{corr} = -B\frac{h_u}{h_b} \leq 0 \tag{58}$$

Hence, a fabricated study yields no positive knowledge and is likely to yield negative knowledge. This latter fact offers the solution to an interesting epistemological conundrum, raised by the fact that a fabricated study may report a claim that is actually true. If independent, genuine studies confirm the made-up finding, then technically no damage has been done to our knowledge.

Equation 56 may shed new light on this conundrum. We can let K represent the genuine knowledge that a body of literature produces. The fabricated study's K is then $-B\frac{h_u}{h_b} \leq K$, because $B = K$ and $h_u > h_b$. The extra information costs of fabricating

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

the entire study generate a net loss of information, even if the underlying claim is correct.

4.6.2 Example 2: knowledge cost of arbitrarily dropping data points

Let's imagine a researcher who collected a sample of n data points, and made a claim $K(Y^n; X^n, \tau) > 0$ without explicitly declaring that during the analysis she had dropped a certain number z of data points which made her results look "better" - i.e. her K appear larger than it is. How egregious was this behaviour?

From equation 56, we derive the minimum conditions under which a bias is tolerable ($K_{corr} > 0$) as

$$K(Y; X, \tau) > B\frac{h_u}{h_b} \tag{59}$$

In the best-case scenario, the researcher identified possible outliers based on a conventional threshold of 3σ, and was therefore confronted with the choice of dropping only 0.3% of her data points, i.e. $n_\beta = 0.003n_y$. This leads to $h_u/h_b \approx 1$ and the simplified condition, $K > B$, in which the bias has to be smaller than the total effect reported. To obtain the full reported effect by dropping no more than 0.3% of data points requires either that the reported effect K is extremely small, and therefore unlikely to be substantively significant, or that the outliers were extremely deviant from the normal range of data. In the latter case, then such data ought to have been removed anyway, and, if retained in the study, would not go unnoticed to the reader. Therefore, arbitrariness in dropping such outliers has a minor impact.

In the worst-case scenario, however, the researcher has inspected each of the n data points and decided whether to drop them or not based on their values. In this case, $n_\beta = n$, yielding

$$\frac{h_u}{h_b} = 1 + \frac{H(B)}{H(Y) + H(X) - \frac{\log p(\tau)}{n}} \approx 2 \tag{60}$$

with the latter approximation derived from assuming that data-selection process required as much information as that contained in the data, and n is large. In this case, therefore the QRP would be tolerable only if $K > 2B$, i.e. the result obtained without the QRP is twice as large as that produced with the QRP. However, if the K inclusive of data was very large, then the researcher would have little improper reasons to drop data points, unless she was biased against producing a result (in which case $K = B$). Therefore, we conclude that under the most likely conditions in which it occurs, selecting data points indiscriminately would be an extremely damaging practice, leading to $K_{corr} < 0$.

4.7 What characterizes a pseudoscience?

Problem: Philosophers have proposed a vast and articulated panorama of criteria to demarcate genuine scientific activity from metaphysics or pseudoscience (S2 Table). However, none of these criteria is accepted as universally valid, and prominent contemporary philosoophers of science tend to endorse a "multicriteria" approach, in which scientific practices are united by a family resemblance, but no single universal trait (e.g. [66–68]).

The multi-criterial solution to the demarcation problem is appealing but has limited theoretical and practical utility. In particular, it shifts the question from identifying a single property common to all sciences to identifying many properties common to some. Proposed lists of criteria typically include normative principles or behavioural standards

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

such as "rigorously assessing evidence", "openness to criticism", etc. These standards are unbojectionable but are hard to assess rigorously. Furthermore, since the minimum number of characteristics that a legitimate science should possess is somewhat arbitrary, virtually any practice may be considered a "science" according to one scheme or another (e.g. intelligent design [69]).

Answer: A pseudoscientific field is characterized by $K_{corr} < 0$, because

$$K < B\frac{h_u}{h_b} \tag{61}$$

where the terms K, B, h_u, h_b are the cumulative equivalent of the terms in equation 56.

Explanation:

Activities such as palmistry, astrology, homeopathy, or psychoanalysis are characterized by having a defined methodology, which contains its own laws, rules and procedures, let's call it ψ. This ψ is what makes these practices appear scientific, and it is believed to produce a $K(Y; X, \psi) > 0$. However, such activities have been recognized as bogus (often many centuries before the concept of science was formalized), because they typically manifest three conditions: 1) they (appear to) produce large amounts of explanatory knowledge, but typically show to have no predictive or causal knowledge; 2) any predictive success or causal power that pseudoscientists attribute to the explanans is more economically explained by well-understood and unrelated phenomena and methodologies; and/or 3) their theories and methodologies are independent from, and often incompatible with, those of well-established and successful sciences (see [68]).

All three properties are contained and quantified in equation 56.

- Condition 1 implies that a field's observed, as opposed to predicted, K is zero, leading to the condition $K_{adj} < 0$ (section 3.3.2) and therefore also to $K_{corr} < 0$ (section 4.6).

- Condition 2 entails that, to any extent that a pseudoscientific methodology (appears to) successfully explain, influence or predict an outcome, the same effect can be obtained with a τ without the specific component ψ. Conscious and unconscious biases in study design (e.g. failure to account for the placebo effect) and post-hoc biases (e.g. second-guessing one's interpretation) fall into this category of explainable effects. We could also interpret K as being the effect produced by standard methods τ, and B as the (identical) effect produced by the pseudoscience, which however has a larger and useless methodology, as it comprises of the sum $-(\log p(\tau) + \log p(\psi))$, leading to condition 56.

- Condition 3, finally can be computed, at least in principle, as an extra methodological cost required by combining the pseudoscientific theory ψ with other standard theories τ. The sum $-(\log p(\tau) + \log p(\psi))$ quantifies a non-theory, in which the two theories are simply kept together with no attempt and unification. However, we could quantify the costs of a theoretical synthesis of the two as a third theory, say υ, that provides a coherent account of the two at the cost of being much longer than the two combined, $\log p(\upsilon) << \log p(\tau) + \log p(\psi)$.The relative length of υ will reflect the extent to which the two theories are incompatible and require additional explanations to be made to cohere. Formal methods to quantify theoretical discrepancies may be developed in future work.

How pseudoscientific is Astrology?

Many studies have been conducted to test the predictions of Astrology, but their results were typically rejected by practising astrologers on various methodological grounds. A

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

notable exception is represented by [70], a study that was designed and conducted with the collaboration and approval of the National Council for Geocosmic Research, a highly prominent organization of astrologers.

In the part of experiment that was deemed most informative, practising astrologers were asked to match an astrological natal chart with one of three personality profiles produced using the California Personality Inventory. If the natal chart contains no useful information about an individual's personality, the success rate is predicted to be random, i.e. 33%, giving $H(Y) = 1.58$. The astrologers predicted that their success rate would be at least 50%, suggesting $H(Y|X, \psi) = 1.58/2 = 0.79$. The astrologer's explanans includes the production of a natal chart, which requires the input of the subject's birth time (hh:mm), date (dd/mm/yyyy) and location (latitude and longitude, four digits each) for a total information of approximately 50 bits. The theory ψ includes the algorithm to compute the star and planet's position and the relation between these and the individual's personality. This could be estimated, but we will avoid doing so for brevity. The alternative, scientific hypothesis according to which there is no effect to be observed, has $h_u = 1$.

Results of the experiment showed that the astrologers did not guess an individual's personality above chance [70]. Therefore, $K = 0$ and equation 61 is certainly satisfied. The K value of astrology from this study is estimated to be

$$K(Y; astrology) = -B\frac{h_u}{h_b} = -\frac{1.58 - 0.79}{1.58}\frac{1.58 + 50 - \log(\psi)\frac{n_\psi}{n_Y}}{1} \leq -25.79 \qquad (62)$$

in which the inequality takes the unspecified costs of ψ into account. The cumulative K of astrology is likely to be even lower, because the experiment offered a conservative choice between only three alternatives, whereas astrology's claimed explanandum is likely to be much larger, as it includes multiple personality dimensions (see section 4.3.3).

4.8 What makes a science "soft"?

Problem: There is ample evidence that scientific practices vary gradually and almost linearly if disciplines are arranged according to the complexity of their subject matters (i.e., broadly speaking, mathematics, physical, biological, social sciences, and humanities) [62, 71–73]. This order reflects what people intuitively would consider an order of increasing scientific "softness", yet this concept has no precise definition and the adjective "soft science" is mostly considered denigrative. This may be why the notion of a hierarchy of the sciences is nowadays disregarded in favour of a partial or complete epistemological pluralism (e.g. [67]). How do we define scientific softness?

Answer: Given two fields studying systems Y_A, X_A, τ_A and Y_B, X_B, τ_B, field A is harder than B if

$$\frac{k_A}{k_B} > \frac{h_B}{h_A} \qquad (63)$$

in which k_A, k_B and h_A, h_B are representatively valid estimates of the fields bias-adjusted cumulative effects and hardness component, given by properties of their systems as well as the field's average level of accuracy, reproducibility and bias.

Explanation: Equation 63 is a re-arrangement of the condition $K(Y_A; X_A, \tau_A) > K(Y_B; X_B, \tau_B)$, i.e. the condition that field A is more negentropicly efficient than field B. As argued below, this condition reflects the intuitive concept of scientific hardness.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

The various criteria proposed to distinguish stereotypically "hard" sciences like physics from stereotypically "soft" ones like sociology cluster along two relevant dimensions:

- 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 these systems' number of relevant variables and the intricacy of their interactions [74].
- Consensus: from the physical to the social sciences, there is a decline in the ability of scientists to reach agreement on the relevance of findings, on the correct methodologies to use, even on the relevant research questions to ask, and therefore ultimately on the validity of any particular theory [75].

(see Table S2 Table, and [73] for further references).

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 $H(Y|X,\tau)$ are high. Consequently, progress is low (section 4.3). A system in which consensus is low is one in which the cumulative methodology $\bar{\tau} + \bar{d}_{\tau}$ expands rapidly as the literature grows. Moreover, higher complexity and particularity of subject matter entails that a given knowledge is applicable to a limited number of phenomena, entailing smaller n_Y. Therefore, all the typical traits associated with a "soft" science lead to predict a lower value of K.

Example: mapping a Hierarchy of the Sciences

The idea that the sciences can be ordered by a hierarchy, which reflects the growing complexity of subject matter and, in reverse order, the speed of scientific progress, can be traced back at least to the ideas of Auguste Comte (1798-1857). The K values estimated in previous sections for various disciplines approximately reflect the order expected based on equation 63, particularly if the re-scaled K values are compared instead, i.e.

$$H(Y_A)k_A h_A > H(Y_B)k_B h_B \tag{64}$$

However, these comparisons cannot be considered complete and conclusive, because the values estimated may not be representatively valid. In addition to making frequent simplifying assumptions, our estimations were usually based on individual cases (not on cumulative evidence coming from a body of literature) and have overlooked characteristics of a field that may be relevant to determine the hierarchy (e.g. the average reproducibility of a literature). Moreover, there may be yet unresolved problems of scaling that impede a full comparison between widely different systems.

Therefore, at present equation 64 could be used to rank closely related fields, whereas methods to compare widely different systems may require further methodological developments of K theory. If produced, this modern hierarchy of the sciences would include a ranking of pseudosciences, thereby considerably extending Comte's original vision.

5 Discussion

This article proposed that K, a quantity derived from a simple function of three main arguments, is a general quantifier of knowledge that can be used in meta-scientific studies. This quantity has many relevant properties. Being derived as a standardization

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

of Shannon's mutual information, this measure is free from distributional assumptions (section 3.2.1). Having the structure of of may statistical measures, K is compatible with ordinary quantifiers of effect size and algorithmic complexity (section 3.2.2). Unlike these ordinary metrics, however, K has a direct physical interpretation as the system-specific efficiency with which information is converted into order or, equivalently, work (section 3.2.3). Furthermore, the K function has properties required by a general quantifier of knowledge, which include Ockham's razor, a dependency on accuracy, different values for causal and correlational knowledge, and a decline with distances between systems (section 3.3). Applied to concrete questions, analyses using the K function yield concise equations that answer meta-scientific questions about studies, research fields, disciplines and knowledge as a whole (see Table 1).

Table 1. Summary of results

Question	Formula	Interpretation
How much knowledge is contained in a theoretical system?	$K = h$	Logico-deductive knowledge is a lossless compression of noise-free systems. Its value is inversely related to its complexity and directly related to the extent of its domain of application.
How much knowledge is contained in an empirical system?	$K = k \times h$	Empirical knowledge is lossy compression. It is encoded in a theory/methodology whose predictions have a non-zero error. It follows that $K_{empirical} < K_{theoretical}$
How much progress is a field making?	$m\Delta X + \Delta\tau < nY\frac{\Delta k}{K}$	Progress occurs to the extent that explanandum and/or explanatory power expand more than the explanans. This is the essence of consilience.
How reproducible is a research finding?	$K_r = KA^{-\boldsymbol{\lambda}\cdot\boldsymbol{d}}$	The difference between the K of a study and its replication K_r is an exponential function of the distance between their systems and/or methodologies.
What is the value of a null or negative result?	$K_{null} \leq \frac{h}{H(Y)} \log \frac{\lvert\mathcal{T}\rvert}{\lvert\mathcal{T}\rvert - 1}$	The knowledge yielded by a single conclusive negative result is an exponentially declining function of the total number of hypotheses (theories, methods, explanations or outcomes) $\lvert\mathcal{T}\rvert$ that remain untested.
What is the cost of research fabrication, falsification, bias and QRP?	$K_{corr} = K - B\frac{h_u}{h_b}$	The K corrected for a questioned methodology is inversely proportional to the methodology's relative description length times the bias it generates (B).
What makes a science "soft"?	$\frac{k_H}{k_S} > \frac{h_S}{h_H}$	Compared to a harder science (H), a softer science (S) yields relatively lower knowledge at the cost of relatively more complex theories and methods.
When is a field a pseudoscience?	$K < B\frac{h_u}{h_b}$	A pseudoscience results from a hyper-biased theory/methodology that produces net negative knowledge. Conversely, a science has $K > B\frac{h_u}{h_b}$.

These results suggest that the theory and method proposed in this article (henceforth indicated as "K theory" for brevity) could find useful applications in meta-research and beyond, at multiple levels of analysis. First and foremost, K theory provides a language to discuss meta-scientific concepts such as progress, bias and reproducibility in terms that are general and abstract, and yet specific enough to avoid confusing over-simplifications. For example, the concept of bias is often operationalized in meta-research as an excess of statistically significant findings [7] or as an exaggeration of findings due to QRPs [76]. Depending on the meta-research contexts, however, these definitions may be too narrow, because they exclude biases against positive findings and only apply to methodologies using null-hypothesis significance testing, or too generic, because they aggregate research practices that may differ from each other in relevant ways. Similar difficulties have been noted in how reproducibility, negative results and other concepts are used (e.g. [2]). The definitions given by K theory may help avoid these limitations because they are both more general and more adaptable to field-specific contexts.

Beyond the conceptual level, K theory offers a framework for meta-research studies, because it contextualizes meta-research results at an appropriate level of generalization. Current meta-research models and empirical studies face a conundrum: they usually

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

aim to draw general conclusions about phenomena that may occur anywhere in science, but these phenomena find contextual expression in fields that vary widely in characteristics of subject matter, theory, methodology and other aspects. As a result, meta-research studies are forced to chose between restricting their conclusions to the specific field or literature they assessed, which may under-emphasize the relevance of findings, or over-generalizing them to an entire field, discipline or science. One of the unfortunate side effects of the latter choice has been the development of a "science in crisis" narrative, which has little empirical support [3]. Undue under-and over-generalizations may be avoided by systematizing meta-research results with K theory, which offers a mid-level understanding of meta-scientific phenomena that is independent of subject matter and yet definable and measurable.

An example of the mid-level generalizations permitted by K theory is the hierarchy of sciences and pseudosciences proposed in section 4.8. At a very coarse-grained level of analysis (i.e. by disciplinary subject matter), this order is likely to reflect the old Comtean hierarchy: at one extreme we may find disciplines like mathematics and physics, which attain relatively high amounts of compression thanks to the extreme regularity and accuracy of their subject matters. At the other extreme are disciplines like ecology or social psychology, in which patterns are relatively short-lived, noise is high, and therefore greater effort is devoted to describing rather than modelling phenomena, and to explaining rather than predicting them. At a finer level of analysis, however, an empirically derived hierarchy of the sciences might reveal that research fields with completely different subject matters have actually very similar characteristics. This could lead us to abandon traditional disciplinary categories (e.g. "physics" or "social psychology") in favour of epistemologically relevant categories such as "high-h" fields, or "low-lambda" systems.

Other classifications and theories about science are conceivable. As an alternative to the "hard-soft" dimension, for example, could be one between two strategies. On the one hand, is what we might call a "τ-strategy", which invests more resources in identifying and encoding rigid laws that allow long-term predictions. On the other hand, we have a "X-strategy", which invests greater resources in acquiring large amounts of contingent, descriptive information, which enables accurate explanations and short-term predictions. Depending on characteristics of the explananda and the amount of resources available for the storage and processing of information, each scientific field might express an optimal balance between τ-strategy and X-strategy.

With adequate developments of its methodology, K theory may also find useful applications in research policy and other areas. In addition to providing a universal measure of effect size to compare findings of different studies or fields, K theory could offer tools to forecast the potential of individual fields to yield reproducible results, to be at risk from bias and misconduct, and the speed with which it can make progress. Furthermore, although the focus of this article has been quantitative research, the K function could in principle be applied to study qualitative forms of knowledge. Indeed, it could be applied to quantify any expression of cognition and learning, including humour, art, biological evolution or artificial intelligence (see S1 text). These extensions of K theory require further theoretical and empirical elaborations, which are beyond the scope of this article and is left to future research.

One of the virtues of K theory is its simplicity. It symthesizes innumerable previous approaches to combining knowledge and information theory, and it does so in a formulation that, to the best of the author's knowledge, is entirely new. Earlier ideas that have much inspired the K function are to be found in Brillouin, who discussed the information value of experiments and calculated the information content of a physical law [23]. Brillouin's analysis, however, did not include factors that are key to the K function, including the standardization on logarithm space, knowledge's decline rate, the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

number of n_y of potential applications of knowledge, and the inclusion of the information costs of the theory τ. The latter is a core component of the Minimum Description Length principle, which was first proposed by Rhissanen [25] and is finding growing applications in problems of statistical inference and computation (e.g. [24, 26]). The methods developed by MDL proponents and by algorithmic information theory are entirely compatible with the K function (and could be used to quantify τ), but differ in important theoretical and mathematical details from it (see section 3.2.2). Within philosophy, Paul Thaggard's "computational philosophy of science" [29] offers numerous insights into the nature of scientific theories and methodologies. Thaggard's ideas could also help to clarify and quantify the contents of τ. Again, however, Thaggard's theory did not offer a general quantifier of knowledge as is proposed in this article.

At least three criticisms of the ideas proposed in this essay may be expected. The first is a philosophical concern with the notion of knowledge, which in this article is defined as information compression by pattern encoding. Critics might argue that this definition does not correspond to the epistemological concept of knowledge as "true, justified belief" [77]. Even Fred Dretske, whose work extensively explored the connection between knowledge and information [28], maintained that "false information" was not genuine information and that knowledge required the latter [78]. The notion of knowledge proposed in this text, however, is only apparently unorthodox. In the K formalism, a true justified belief corresponds to a system for which $K(Y; X, \tau) > 0$. It can be shown that a "false, unjustified" belief is one in which $K(Y; X, \tau \leq 0)$. Therefore, far from contradicting information-theoretic epistemologies, K theory may give quantitative insights to remaining questions such as "how much information is enough"? [78].

A second criticism might be methodological, because entropy is a difficult quantity to measure. Estimates of entropy based on empirical frequencies can be biased when samples sizes are small, and they can be computationally demanding when data is large and multidimensional. Neither of these limitations, however, is critical. With regards to the former problem, as demonstrated in section 3.3.6, powerful computational methods to estimate entropy with limited sample size are already available [35]. Future projects could aim at improve such methods, tayloring them to the requirements of the K function. With regards to the latter problem, we may note that the "multidimensional" K_{md} used in section 4.3 is not computationally demanding (because it is derived from computing uni-dimensional entropies), and may represent a more useful (e.g. psychologically interpretable) measure in many contexts. Furthermore, analytical approaches to estimate the entropy of mixed distributions and other complex data structures are already available and are likely to be developed further (e.g. [79, 80]).

The third criticism may regard the empirical validity of the measures proposed. It was emphasized repeatedly in the text how all the practical examples offered were merely illustrative and preliminary, because they generally relied on incomplete data and simplifying assumptions. In particular, it is generally difficult to quantify exactly the information content of the τ component. This is indeed the major current limitation of the approach proposed, but it is not insurmountable. At least within a meta-research context, it may suffice to measure τ in relative terms. For example one may assess how divergent the methods of two studies are, or what is their relative desciription length. These relative quantifications are certainly attainable, and could become remarkably accurate if they were based on a taxonomy of methods, which provided a fixed "alphabet" $\mathcal{T}$ of methodological choices characterizing a study. Such taxonomies are already being developed in many fields to improve reporting standards (e.g. [81]). Building upon such taxonomies, meta-researchers interested in a particular literature could build a literature-specific alphabet, which would allow to compare the methods used by studies within the same or closely related fields. Cross-fields comparisons may

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

require more refined methods of conversion and re-scaling that will be developed in future research.

6 Supporting information

S1 text 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 [82].

Advances in quantum information theory suggests that our universe may have access to unlimited amounts of information, or at least of information processing capabilities [83] (but see [84] for a critique). However, even if this were the case, there would still be little doubt that information is finite as it pertains to knowledge attainable by organisms. Sensory organs, brains, genomes and all other pattern-encoding structures that underlie learning are finite. The sense of vision is constructed from a limited number of cone and rod cells; the sense of hearing uses information from a limited number of hair cells, each of which responds to a narrow band of acoustic frequencies; brains contain a limited number of connections; genomes a countable number of bases, etc. The finitude of all biological structures is one of the considerations that has led cognitive scientists and biologists to assume information is finite when attempting, for example, to model the evolution of animal cognitive abilities [85]. Even mathematicians have been looking with suspicion to the notion of infinity for a long time [86]. For example, it has been repeatedly and independently shown that, if rational numbers were actually infinite, then infinite information could be stored in them and this would lead to insurmountable contradictions [87].

Independent of physical, biological, and mathematical considerations, the postulate that information is finite is justifiable on instrumentalist grounds, because it is the most realistic assumption to make when analyzing scientific knowledge. Quantitative empirical knowledge is based on measurements, which are technically defined as partitionings of attributes in sets of mutually exclusive categories [88]. In principle this partitioning could recur an infinite number of times, but in practice it never does. Scales of measurements used by researchers to measure empirical phenomena might be idealized as running to infinity, but in practice always contemplate a range of plausible values and are delimited at one or both ends. Values beyond these ends can be imagined as constituting a single set of extreme values that may occur with very small but finite probability.

Therefore, following either theoretical or instrumentalist arguments, we are compelled to postulate that information, i.e. the source of knowledge, is a finite quantity. Its fundamental unit of measurement is discrete and is called the bit, i.e. the "difference that makes a difference", according to Gregory Bateson's famous definition [89]. For this difference to make any difference it must be perceivable. Hence, information presupposes the capacity to dichotomize signals into "same" and "not same". This dichotomization can occur recursively and we can picture the process by which information is generated as a progressive subdivision (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.

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

information required to navigate the world successfully. By "pattern" we intend here simply a dependency between attributes—in other words a relationship that makes one event more or less likely, from the point of view of an organism, depending on another event. By encoding patterns, an organism reduces the uncertainty it confronts about its environment—in other words, it *adapts*. Therefore, postulate 2, just like postulate 1, is likely to reflect an elementary fact of nature; a fact that arguably underlies not just human knowledge but all manifestations of life.

The idea that knowledge, or at least scientific knowledge, is information compression is far from new. For example, in the late 1800s, physicist and philosopher Ernst Mach argued that the value of physical laws lay in the "economy of thought" that they permitted [21]. Other prominent scientists and philosophers of the time, such as mathematician Henri Poincaré, expressed similar ideas [90]. Following the development of information theory, scientific knowledge and other cognitive activities have been examined in quantitative terms (e.g. [23], [91]). Nonetheless, the equivalence between scientific knowledge and information compression has been presented as a principle of secondary importance by later philosophers (includig for example Popper [58]), and today does not appear to occupy the foundational role that it arguably deserves [92].

The reluctance to equate science with information compression might be partially explained by two common misconceptions. The first, is an apparent conflation of lossless compression, which allows data to be reconstructed exactly, with lossy compression, in which instead information from the original source is lost. Some proponents of the compression hypothesis adopt exclusively a lossless compression model, and therefore debate whether empirical data truly are compressible (e.g. [93]). However, science is clearly 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 scepticism seems to lie in an insufficient appreciation for the fundamental role that information compression plays in not only science, but also knowledge and all other manifestations of biological adaptation. Even scientists who equate information compression with learning appear to under-estimate the fundamental role that pattern-encoding and information compression play in all manifestations of life. In their seminal introductory text to Kolmogorov complexity [24], for example, Li and Vitanyi unhesitatingly claim that "science may be regarded as the art of data compression" (pp. 713), that "learning, in general, appears to involve compression of observed data or the results of experiments", and that "in everyday life, we continuously compress information that is presented to us by the environment", but then appear cautious and conservative in extending this principle to non-human species, by merely suggesting that "perhaps animals do this as well", and citing results of studies on tactile information transmission in ants ([24] pp. 711). Is seems that even the most prominent experts and proponents of information compression methodologies can be disinclined to apply their favoured principle beyond the realm of human cognition and animal behaviour.

Indeed, information compression by pattern encoding is the quintessence of biological adaptation, in all of its manifestations. Changes in a population's genetic frequencies in response to environmental pressures can be seen as a form of adaptive learning, in which natural selection reinforces a certain phenotypic response to a certain environment and weakens other responses, thereby allowing a population's genetic codes to "remember" fruitful responses and "forget" erroneous (i.e. non-adaptive) ones. For these reinforcement processes to occur at all, environmental conditions must be heterogeneous and yet partially predictable. Natural selection, in other words, allows regularities in the environment to be genetically encoded. This process gives rise to biodiversity that may mirror environmental heterogeneity at multiple levels (populations, varieties, species, etc.). Such environmental heterogeneity is not

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

exclusively spatial (geographical). Temporal heterogeneity in environmental conditions gives rise to various forms of phenotypic plasticity, in which identical genomes express different phenotypes depending on cues and signals received from the environment [94]. Whether genetic or phenotypic, adaptation will be measurable as a correlation between possible environmental conditions and alternative genotypes or phenotypes. This correlation is in itself a measurable pattern, which results in turn from the pattern-encoding capacities proper to any species capable of adaptation.

As environments are increasingly shaped by biological processes, they become more complex and heterogeneous, and they therefore select for ever more efficient adaptive capabilities—ever more rapid and accurate ways to detect and process environmental cues and signals. Immune systems, for example, allow large multicellular plants and animals to protect themselves from infective agents and other biological threats whose rate of change far out-competes their own speed of genetic adaptation; endocrine systems allow the various parts of an organism to communicate or coordinate their internal activities in order to respond more adaptively and plastically to external conditions. Similar selective pressures have favoured organisms with nervous systems of increasing size and complexity. Animal behaviour and cognition, in other words, are simply higher-order manifestations of phenotypic plasticity, which allow an organism to respond to environmental challenges on shorter temporal scales. Behavioural responses may be hard-wired in a genome or acquired during an organism's life time, but in either case they entail "learning" in the more conventional sense of encoding, processing, and storing memories of patterns and regularities abstracted from environmental cues and signals.

Human cognition, therefore, may be best understood as just another manifestation of biological adaptation by pattern encoding. At the core of human cognition, as with all other forms of biological adaptation, lies the ability to anticipate events and thus minimize error. When we say that we "know" something, we are claiming that we have fewer uncertainties about it because, given an input, we can predict above chance what will come next. We "know a city", for example, if and in proportion to how well we are able to find our way around it, being able to foretell where to go from one street to the next and/or navigating it by means of a simplified representation of it (i.e. a mental map). This ability embodies the kind of information we may communicate to a stranger when asked for directions: if we "know" the place, we can provide them with a series of "if-then" statements about what direction to take given an identifiable point, in multiple steps. In another example, we "know a song" in proportion to how accurately we can reproduce the sequence of words and intonations that recreate it with no error or hesitation or how readily we can recognize it when we hear a few notes from it. Similarly, we "know a person" in proportion to how many patterns about them we have encoded: at first, we might only be able to recognize their facial features; after making superficial acquaintance with them, we will be able to connect these features to their name; when we know them better, we can tell how they will respond to simple questions such as "where are you from?"; eventually we might "know them well" enough to predict their behaviour rather accurately and foretell, for example, the conditions that will make them feel happy, interested, angry, etc.

The examples above aim to illustrate how the concept of "prediction" underlies all forms of knowledge, not just scientific knowledge, and applies to both time (e.g., knowing a song) and space (e.g., knowing a city). Memory and recognition, too, can be qualified as forms of prediction and therefore as manifestations of information compression, whereby sequences of sensory impressions are encoded and recalled (memory) or matched to new experiences (recognition) in response to perception of endogenous or exogenous signals. Language is also a pattern encoding, information compression tool. A typical sentence, i.e. the fundamental structure of human language

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

and thought, expresses the connection between one entity, the subject, and another entity or property, via a relation condition encoded in a verb. It is not a coincidence that the most elementary verb of all—one that is fundamental to all human languages—is the verb "to be." This verb conveys a direct relation between two entities, and thus represents the simplest pattern that can be encoded: "same" versus "not same", as discussed in relation to Postulate 1. Even a seemingly abstract processes like logical deduction and inference can be understood as resulting form pattern-encoding. According to some analyses, computing itself, and all other manifestations of artificial and biological intelligence, may result from a simple process of information compression by pattern-matching [95].

Scientific knowledge, therefore, is most naturally characterized as just one manifestation of human cognition amongst many and, therefore, as nothing more than a pattern-encoding activity that reduces uncertainty about one phenomenon by relating it to information about other phenomena. The knowledge produced by all fields of scientific research is structured in this way.

- Mathematical theorems uncover logical connections between two seemingly unrelated theoretical constructs, generally proving that the two are one and the same.

- Research in the physical sciences typically aims at uncovering mathematical laws, which are rather explicitly encoding patterns (i.e. relationships between quantities). Even when purely descriptive, however, physical research actually consists in the encoding of pattern and relations between phenomena—for example, measuring the atomic weight of a known substance might appear to be a purely descriptive activity, but the substance itself is identified by its reactive properties. Therefore, such research is about drawing connections between properties.

- Most biological and biomedical research consists in identifying correlations or causes and/or in describing properties of natural phenomena, all of which are pattern-encoding activities. Research in taxonomy and systematics might appear to be an exception, but it is not: organizing the traits of a multitude of species into a succinct taxonomical tree is the most elementary form of data compression.

- Quantitative social and behavioural sciences operate in a similar fashion to the biological sciences. Even qualitative, ethnographic, purely descriptive social and historical research consists in data compression, because it presupposes that there are general facts about human experiences, individuals, or groups that can be communicated, entailing that they can be described, connected to each other and/or summarized in a finite amount of text.

- The humanities yield understanding about complex and often unique human experiences, and might therefore appear to have fundamentally different objectives from the natural and social sciences. To any extent that they offer knowledge and understanding, however, these come in the form of information compression. Historical research, or legal analysis, for example, are guided by cases, logic and inference, and therefore follow the principles of economy of thought and compression. The study of works of literature, to make another example, produce knowledge by drawing connections and similarities between texts, identifying general schemata, and/or uncovering new meaning in texts by means of similes and metaphors [96]. Similarities, connections, schemata, similes, and metaphors, which might arguably constitute the basis of human cognition [96], are all manifestations of information compression by pattern-encoding.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

Indeed, arguably any other, non-academic, manifestation of human cognition can be understood as stemming from a process of information compression. The sensual and intellectual pleasure that humans gain from music and art, for example, seems to derive from an optimal balance between perception of structure (pattern that generates predictions and expectations) and perception of novelty (which stimulates interest by presenting new and knowable information) [97]. The sense of humour similarly seems to arise from the sudden and unexpected overturning of the predicted pattern or, more accurately, when an initially plausible explanation of a condition is suddenly replaced by an alternative, unusual, and yet equally valid one [98]. The intellectual and artistic value of a novel or artwork lies in its ability to reveal previously unnoticed connections between events or phenomena in the world (thereby revealing a pattern) and/or to synthesize and effectively communicate highly individual, complex, and ineffable human experiences (compressing the experience by abstracting a relevant essence from it).

S2 text Relation with continuous distribution Indicating with $f(x)$ a probability density function and with $h(X) = -\int f(x) log f(x) d(x)$ the corresponding differential entropy, we have

$$H(X^{\Delta}) \approx h(X) + \log \frac{1}{\Delta} = h(X) + n \tag{65}$$

in which $\Delta = 2^{-n}$ is the size of the length of the bin in which $f(x)$ is quantized, and n corresponds to the number of bits required to describe the function to $n - bit$ accuracy. Evidently, we can always re-scale X in order to have $\Delta = 1$.

Equation 65 applies to any probability density function. Here we will consider in particular the case of the normal distribution, the differential entropy of which is simply $h(x) = \log \sqrt{2\pi e}\sigma_y$. Therefore, if y is a continuous RV, quantized to n bits, for a given x and τ we have

$$\begin{aligned} K(y;x,\tau) = \frac{\log(\sqrt{2\pi e}\sigma_y) + n - \log(\sqrt{2\pi e}\sigma_{y|x,\tau}) - n}{\log(\sqrt{2\pi e}\sigma_y) + n + x + \tau} = \\ \frac{\log\sqrt{2\pi e} + \log\sigma_y - \log\sqrt{2\pi e} - \log\sigma_{y|x,\tau}}{\log\sqrt{2\pi e} + \log\sigma_y^2 + n + x + \tau} = \\ \frac{\log\sigma_y - \log\sigma_{y|x,\tau}}{\log\sigma_y + x + \tau + \log\sqrt{2\pi e} + n} \rightarrow \frac{\log(\sigma'_y - \log\sigma'_{y|x,\tau}}{\log\sigma'_y + x + \tau + \log\sqrt{2\pi e}} = \frac{\log\sigma'_y - \log\sigma'_{y|x,\tau}}{\log\sigma'_y + x + \tau + C} \end{aligned} \tag{66}$$

In which $C = \log\sqrt{2\pi e}$ and σ' corresponds to σ rescaled to a common lowest significant digit (for example, from $\sigma = 0.123$ to $sigma = 123$.

S3 text

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

Proof.

$$\chi^2 = \sum_{i=m} \frac{expected_i - (observed_i)^2}{expected_i} = \sum_{i=m} \frac{n \times (p_{exp}(i) - p_{obs}(i))^2}{p_{exp}(i)} =$$
$$= f\left(\sum_{i=m} \frac{|p_{exp}(i) - p_{obs}(i)|}{p_{exp}(i)}\right) = f\left(\sum_{x,y} \frac{|p(x)p(y) - p(x,y)|}{p(x)p(y)}\right) =$$
$$= g\left(\sum_{x,y} \frac{\log p(x) + \log p(y) - \log p(xy)}{\log p(x) + \log p(y)}\right) = g\left(\sum_{x,y} \frac{\log p(y) - \log p(y|x)}{\log p(x) + \log p(y)}\right) =$$
$$= g\left(\sum_{x,y} \frac{p(y,x)}{p(y,x)} \times \frac{\log p(y) - \log p(y|x)}{\log p(x) + \log p(y)}\right) = g\left(\sum_{x,y} \times \frac{p(y)\log p(y) - p(y,x)\log p(y|x)}{p(y)\log p(x) + p(x)\log p(y)}\right) =$$
$$= h(K(Y;X,\tau)) \quad (67)$$

□

S4 text Akaike's Information Criterion Akaike's information criterion (AIC) contrasts the goodness of fit of a model to the model's complexity, interpreted exclusively as the number of degrees of freedom (parameters) in the model. Specifically, $AIC = 2k - 2ln(L)$ with L representing the likelihood function (the probability of the data) for the model and k the model's number of parameters. A least-squared equivalent is $AIC = 2k - n \times ln(S)$ with S indicating the model's residual sum of squares. When comparing two alternative regression models, model "a" will be better than model "b" if its AIC is lower, i.e.

$$AIC_a < AIC_b \Rightarrow ln\frac{L_a}{L_b} > k_a - k_b \quad (68)$$

It can be shown that in general:

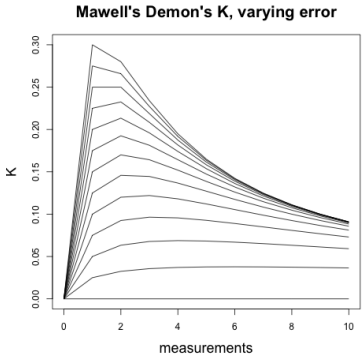
$$K(Y;X,\tau_a) > K(Y;X,\tau_b) \Rightarrow AIC_a < AIC_b \quad (69)$$

Proof. Let τ_a and τ_b be two alternative models for the same data Y^n, then

$$K(Y^{n_Y};\tau_a) > K(Y^{n_Y};\tau_b) \Rightarrow \frac{H(Y) - H(Y|\tau_a)}{H(Y) - H(Y|\tau_b)} > \frac{H(Y) - \log p(\tau_a)}{H(Y) - \log p(\tau_b)}$$
$$\Rightarrow nH(Y|\tau_b)\left(1 + \frac{-\log p(\tau_a)}{nH(Y)}\right) - nH(Y|\tau_a)\left(1 + \frac{-\log p(\tau_b)}{nH(Y)}\right) > -\log p(\tau_a) + \log p(\tau_b) \quad (70)$$

in which the left-hand side quantifies the relative fit of alternative models and the right-hand side contrasts the description lengths of the models themselves. If the only distinguishing feature of the two explanantia (models) is the number of parameters, then $-\log p(\tau_a) + \log p(\tau_b) \equiv k_a - k_b$, and if instead of the entropies on right-hand side we have the estimated probability of strings of length n, we can re-arrange the inequality 70 to obtain a function that converges to equation 68 in the limit of infinite information:

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018



S1 Figure K values plotted against measurements, for an imaginary Maxwell Pressure Demon. Calculations based on equation 18. See text for further details

.

$$
\begin{aligned}
-\log p(y^n|\tau_a)\left(1+\frac{C+k_a}{-\log p(y^n)}\right) - \\
(-\log p(y^n|x^n,\tau_b)\left(1+\frac{C+k_b}{-\log p(y^n)}\right) > k_a - k_b \equiv \\
\left(1+\frac{C+k_a}{-\log p(y^n)}\right)\log\frac{p(y^n|\tau_a)^r}{p(y^n|\tau_b)} > k_a - k_b \quad (71)
\end{aligned}
$$

in which $r = \frac{-\log p(y^n)+C+k_b}{-\log p(y^n)+C+k_a}$, such that

$$
\begin{aligned}
\lim_{n\to+\infty}\left[\left(1+\frac{C+k_a}{-\log p(y^n)}\right)\log\frac{p(y^n|\tau_a)^r}{p(y^n|\tau_b)} > k_a - k_b\right] \\
= log\frac{P(Y|\tau_a)}{P(Y|\tau_b)} > k_a - k_b \\
\equiv log\frac{L_a}{L_b} > k_a - k_b \equiv AIC_a < AIC_b \quad (72)
\end{aligned}
$$

Hence, results obtained with Akaike's information criterion can be derived from the K function, if we assume that the sample size of the explanandum is infinite and the explanantia represent statistical model (pattern) descriptions that are equivalent in all aspects except for the number of parameters. □

S5 text

Proof.

$$
\begin{aligned}
K_{adj} \equiv h\left(\frac{H(Y) - \sum p(yx|\tau)\log\frac{1}{p(y|x,\hat{\tau})}}{H(Y)}\right) = \\
h\left(\frac{H(Y) - \sum p(y,x|\tau)\log\frac{1}{p(y|x,\tau}} - \sum p(y,x|\tau)\log\frac{p(y|x,\tau)}{p(y|x,\hat{\tau})}}{H(Y)}\right) = \\
hK(Y;X,\tau) - D(Y|X,\tau||Y|X,\hat{\tau})\frac{h}{H(Y)} \equiv K_{obs} - D(Y|X,\tau||Y|X,\hat{\tau})\frac{h}{H(Y)} \quad (73)
\end{aligned}
$$

□

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

S6 text The arguments to justify this conclusion are slightly different for distances between explananda and explanantia, and are presented separately. Since the numbers n_y, n_x are not relevant to the argument, we will omit their specification throughout this section.

Proof. **Decline with divergence between explanans and explanandum**. The proof follows directly from our definition of the K function as a standardization of the mutual information function, and by elabotaring on the data processing inequality (DPI). The latter shows that if three random variables X, Y, Z form a Markov Chain, then $I(X;Z) \leq I(X;Y)$ [30]. An important corollary of the DPI shows that it also applies if $Z = f(Y)$, i.e. if each successive RV in the chain is a function of the RV that came before.

Since in the K function $Y = f(X, \tau)$, then $(X_{t_0}, \tau) \rightarrow Y_{t_0} \rightarrow Y_{t_1}...$ also forms a Markov Chain and therefore is subject to the data processing inequality. Following the proof of the inequality given by [30], we consider the mutual information between the explanans and two successive stages of the explanandum $I(X, \tau; Y_n, Y_{n+1})$, which we may simply indicate as $I(X; Y_n, Y_{n+1})$ since τ is a fixed conditioning factor that has no relevance for this part of the proof. The proof of the DPI simply shows that this term can be expanded in two ways:

$$I(X; Y_n, Y_{n+1}) = I(X; Y_n) + I(X; Y_{n+1}|Y_n) = I(X; Y_{n+1}) + I(X; Y_n|Y_{n+1}) \tag{74}$$

and since by Markovity $I(X; Y_{n+1}|Y_n) = 0$, then

$$I(X; Y_{n+1}) = I(X; Y_n) - I(X; Y_n|Y_{n+1}) \leq I(X; Y_n) \tag{75}$$

since $I(X; Y_n|Y_{n+1}) \geq 0$. We re-arrange 75 as

$$I(X; Y_{n+1}) = I(X; Y_n) + (H(Y_n|Y_{n+1}) - H(Y_n|Y_{n+1}X)) \tag{76}$$

and notice that equation 76 can be expanded indefinitely:

$$I(X; Y_{n+2}) = I(X; Y_{n+1}) + (H(Y_{n+1}|Y_{n+2}) - H(Y_{n+1}|Y_{n+2}X))\, I(X; Y_n) + (H(Y_n|Y_{n+2}) - H(Y_n|Y_{n+2}X)) \tag{77}$$

Furthermore, we can show that

$$I(X; Y_{n+1}) - I(X; Y_{n+2}) < I(X; Y_n) - I(X; Y_{n+1}) \tag{78}$$

if $H(Y_n|Y_{n+i}, X) < H(Y_{n+i}|Y_{n+i+1}, X) < H(Y_n|Y_{n+i})$, i.e. that the future relevance of X declines with distance in the Markov chain. In other words, the loss of mutual information along the chain proceeds by decreasing increments, until $H(Y_n|Y_{n+i}, X) = H(Y_n|Y_{n+i}) + \epsilon \approx H(Y_n|Y_{n+i})$, point at which the explanans bears no longer any relevance for determining the states of relative to each other Y. Since the explanans is unchanging, and the system evolves by a Markov process, i.e. following the rules of a transition matrix, this result entails that the value of $I(Y_n; X)$ (and therefore that of $K(Y^{n_Y}; X^{n_X}, \tau)$) declines by proportional decrements and approaches 0 with an exponentially declining curve. An exponential process is indeed how any Markov chain reaches a steady state (see [99]).

This leads us to conclude that, indicating with d_Y the number of "steps", measured in any countable scale, that separate the state Y_{d_Y} from a prior state Y,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

$$K(Y_{d_y}; X, \tau) = K(Y; X, \tau) \times A^{-\lambda_Y d_Y} \tag{79}$$

in which A is an arbitrary basis and λ_Y is a knowledge loss rate specific to the joint system Y, X, τ.

Finally, we may also examine the case in which the DPI allows for equality, and notice that it entails

$$I(X; Y_{n+1}) = I(X; Y_n) \iff I(X; Y_n | Y_{n+1}) = 0 \tag{80}$$

which can only occur in two cases: A) $H(X|Y_{n+1}) = H(X|Y_{n+1}, Y_n)$; B) $H(Y_n|Y_{n+1}, X) = H(Y_n|Y_{n+1})$. Case A implies that the effect of the explanans is constant and invariant along the chain. Perfect stability is rarely, if ever realised in real systems except in logico-deductive systems, in which by definition there is no uncertainty and future behaviour is entirely determined by unchanging relations. Case B, conversely, entails that X has no relevance, but if so, then $K(Y; X, \tau) = 0$ and there is no knowledge about the system to begin with.

Decline with divergence between explanantia In the section above, we have assumed that the theory component of the explanans was a constant. We now relax this assumption and show how, even assuming Y and X to be stationary, a distance between two theories entails a difference in K that can also be described by an exponential curve. The argument in this case verges on the definition of K as resulting from "fixing" the random variable that represents the theory on a specific value of τ (section 3.2.1), combined with a consideration concerning the non-randomness of the value that is thus fixed. Since T results from a sequence of RV $(T_1, T_2 ... T_n)$ that are not necessarily identical nor independently distributed, we will start by making the general case for the random variable T, and then show how this applies to any subset of the random variables composing T.

Let T be a RV with alphabet $\mathcal{T} : \{\tau_1, \tau_2 ... \tau_n\}$ representing the set of all possible theories of finite description length, and let $\tilde{K} : K(Y; X, \tau_i) : \tau_i \in \mathcal{T}$ be the corresponding set of K values. Clearly, $\tilde{K}$ has one maximum, except for the special case in which $K(Y; X, \tau_i) = K(Y; X, \tau_j) \quad \forall \tau_i, \tau_j \in \mathcal{T}$, and all K have exactly the same value irrespective of the theory. Let τ_j be the theory corresponding to the maximum value $K(Y; X, \tau_j)$ in $\tilde{K}$. It follows that for all the remaining $\tau_i \neq \tau_j$, $0 \leq K(Y; X, \tau_i) < K(Y; X, \tau_j)$ and therefore there is a $\lambda_i \in \mathbb{R}$ such that

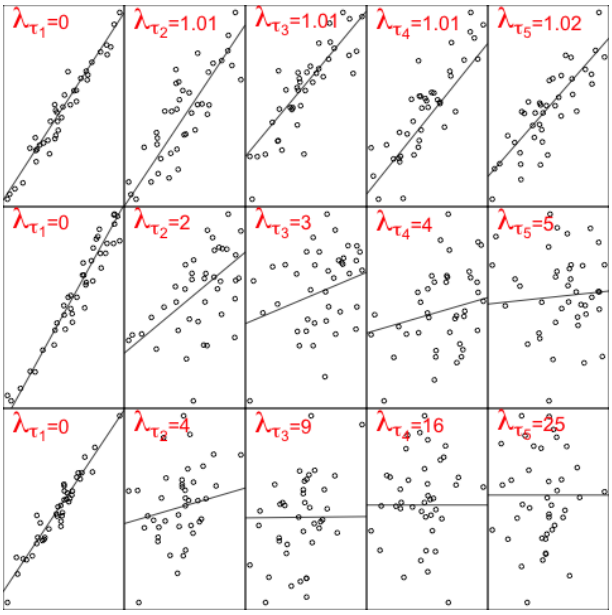
$$K(Y; X, \tau_i) = K(Y; X, \tau_j) \times A^{-\lambda_i d_i} \tag{81}$$

with $d_i = \log \frac{1}{p(\tau_i)} \equiv \tau_i$ being the information required to fix T on the specific theory τ_i (see section 3.2.1).

The relation in equation 81 applies also to the case when T is not fixed on any particular τ, for example because the information concerning which τ is to be fixed is lacking. Indeed, this is the original condition from which K is imagined to be derived. In such case,

$$\begin{aligned} &K(Y; X, T) = h \frac{H(Y) - \sum_{\tau_i \in \mathcal{T}} Pr\{T = \tau_i\} H(Y|X, T = \tau_i)}{H(Y)} = \\ &h \sum_{\tau_i \in \mathcal{T}} Pr\{T = \tau_i\} k(Y; X, T = \tau_i) = h \sum_{\tau_i \in \mathcal{T}} Pr\{T = \tau_i\} k(Y; X, T = \tau_j) \times A^{-\lambda_i d_i} = \\ &hk(Y; X, T = \tau_j) \sum_{\tau_i \in \mathcal{T}} Pr\{T = \tau_i\} A^{-\lambda_i \cdot d_i} \equiv K(Y, X, \tau_j) \times E[A^{-\lambda d}] \equiv K \times A^{-\overline{\lambda}_\tau \overline{d}_\tau} \end{aligned} \tag{82}$$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018



S2 Figure Illustrative examples of how K may vary with λ within a set of alternative methodologies. See text for further details.

in which $\overline{\lambda}_\tau \overline{d}_\tau = -\log_A E[A^{-\lambda d}]$ is a knowledge loss rate specific to the joint system $Y, X, \mathcal{T}$, i.e. the combination of explanans, explanandum, and finite set of theory descriptions.

Since $T : (T_1, T_2 ... T_n)$ is defined as a joint RV, the conclusions above hold for any subsequence T_s of T, with alphabet $\mathcal{T}_s$. In such case, we may indicate with T_c the complementary subset of T for which τ is not varying, and the set of possible K values is $\tilde{K} : \{K(Y; X, \tau_c, \tau_s) : \tau_s \in \mathcal{T}_f\}$ will be described by subset-specific λ_τ and d_τ. Furthermore, to any extent that the effects of each element in the sequence T_s are independent, λ_τ and d_τ could be described by vectors of weights and distances, whose terms are in one-to-one correspondence with the sequence T_s. The decline rate would in this case be given by the inner product of the two vectors.

Figure S2 shows four illustrative examples of sets of K values for different characteristic λ values. The special case in which the set has no unique maximum, which can only occur if all K values are identical, will correspond to the case $\lambda = 0$.

Note that, although we defined the problem by assuming that $K(y, x, \tau_j)$ is the maximum attainable value of K in the set, this assumption is not strictly necessary, because we don't preclude $\lambda < 0$. In general, the conditions $K(Y; X, \tau_i) \leq K(Y; X, \tau_j)$ and $K(Y; X, T) \leq K(Y; X, \tau_j)$ depend on how much above or below average the reference value $K(Y; X, \tau_j)$ is in $\tilde{K}$, which in turn requires making assumptions about how τ_j was determined. As discussed in sections 3.3.54.4, most meta-research analyses are based on the assumption that a study's τ is a random draw from a distribution of possible values subject to Gaussian noise, and is therefore the best representation of an average. This assumption, however, may not be often met in real systems. For example, laboratories that first report an important new discovery may have been - not coincidentally - be operating in conditions where the phenomenon is most likely to be observed, and/or may have developed forms of "tacit" expertise, making their results not exactly reproducible by any other laboratory.

□

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

S7 text Let X^{α} be a RV quantized to resolution (i.e. bin size, or accuracy) α, and let $a \in \mathbb{N}$ (Latin letter a) be the size of the alphabet of X, such that $\alpha_x = \frac{1}{a_x}$. At no cost to generality, let an increase of resolution consist in the progressive sub-partitioning of α, such that $\alpha' = \alpha/n$ with $n \in \mathbb{N}, n \geq 2$ is an increased accuracy. Then:

$$0 < H(X^{\alpha'}) - H(X^{\alpha}) \leq log(n) \tag{83}$$

Proof. If $H(X^{\alpha}) = -\sum_1^a p(x) log p(x)$, with x representing any one of the a partitions, then $H(X^{\alpha'}) = -\sum_1^{a \times n} p(x') log p(x') = -\sum_1^a \sum_1^n p(a)p(n|a) log[p(a)p(n|a)] \equiv H(A) + H(N|A)$, where N and A are the random variables resulting from the partitions. Known properties of entropy tell us that the entropy of the n-partition of α, $H(N|A)$ is smaller or equal to the logarithm of the number n of partitions with equality if and only if the n-partitions of α have all the same probability, i.e. $H(N|A) \leq \log n$. □

Definition: maximal resolution Let X^{α} be a generic quantized random variable with resolution α, and let $\alpha' = \alpha/n$ represent a higher resolution. The measurement error of X^{α} is a quantity $e > 0, e \in \mathbb{Q}$ such that:

$$H(X^{\alpha'}) - H(X^{\alpha}) = log(n), \forall \alpha \leq e \tag{84}$$

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

The effect that a change in resolution has on K depends on the characteristic 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.

From the definitions above follows that, for every empirical system yx for which there is a $\tau \neq \emptyset$ such that $K(Y; X, \tau) > 0$, the system's quantization $Y^{\alpha_Y}, X^{\alpha_X}$ has optimal values of resolution $\alpha*_y$ and $\alpha*_x$ such that:

$$K(Y^{\alpha*_Y}; X^{\alpha*_X}, \tau) > K(Y^{\alpha_Y}; X^{\alpha_X}, \tau), \forall \alpha_Y \neq \alpha*_Y, \alpha_X \neq \alpha*_X \tag{85}$$

Proof. If α is the resolution of Y and $\alpha' = \alpha/n$ is a higher resolution then, assuming for simplicity that τ is constant:

$$K(Y^{\alpha'}; X, \tau) > K(Y^{\alpha}; X, \tau) \iff \frac{H(Y^{\alpha'}) - H(Y^{\alpha'}|X, \tau)}{H(Y^{\alpha}) - H(Y^{\alpha}|X, \tau)} < \frac{H(Y^{\alpha'}) + H(X) + \tau}{H(Y^{\alpha}) + H(X) + \tau} \tag{86}$$

From lemma 83 we know that $H(Y^{\alpha'}) \leq H(Y^{\alpha}) + log(n)$, assuming equality and re-arranging equation 86 we get the condition:

$$H(Y^{\alpha'}|X, \tau) - H(Y^{\alpha}|X\tau) < (1 - K(Y^{\alpha}; X, \tau)) log(n) \tag{87}$$

which if $n = 2$ yields equation 29. Condition 87 is only satisfied when $K(Y^{\alpha}; X, \tau)$ is small and $H(Y^{\alpha'}|X, \tau) - H(Y^{\alpha}|X, \tau) << log(n)$. As resolution increases, however, $K(Y^{\alpha}; X\tau)$ increases by definition, and a point will inevitably be reached at which the increase will stop. Beyond this point, any additional resolution will not increase the numerator of the K function, but will keep increasing the denominator, forcing K to decrease.

The corresponding condition for x is:

$$K(Y; X^{\alpha'}\tau) > K(Y; X^{\alpha}, \tau) \iff \frac{H(Y) - H(Y|X^{\alpha'}, \tau)}{H(Y) - H(Y|X^{\alpha}, \tau)} < \frac{H(Y) + H(X^{\alpha'}) + \tau}{H(Y) + H(X^{\alpha}) + \tau} \tag{88}$$

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

which since $X^{\alpha'} \leq X^{\alpha} + log(n)$ leads to

$$H(Y|X^{\alpha},\tau) - H(Y|X^{\alpha'},\tau) > \log(n)K(Y;X^{\alpha},\tau) \quad (89)$$

Combining equations 86 and 88 yields the general condition:

$$K(Y^{\alpha'_Y};X^{\alpha'_X},\tau) > K(Y^{\alpha_Y};X^{\alpha_X},\tau) \iff$$
$$H(Y^{\alpha'_Y}|X^{\alpha'_X},\tau) - H(Y^{\alpha_Y}|X^{\alpha_X},\tau) < (1-K(Y^{\alpha_Y};X^{\alpha_X},\tau))\log\left(\frac{n_y}{n_x^{\frac{1}{1-K(Y^{\alpha_Y};X^{\alpha_X},\tau)}}}\right) \quad (90)$$

in which $n_Y = \alpha'_Y/\alpha_Y$ and $n_X = \alpha'_X/\alpha_X$, respectively. The right hand side of equation 90 is increasingly negative as K grows, whereas the left hand side is lower bounded by $-H(Y^{\alpha}|X^{\alpha_X},\tau)$, which restricts the range of conditions for growth K to grow.

The only scenario in which K never ceases to grow is one in which $H(Y^{\alpha'_Y}|X^{\alpha'_X},\tau) - H(Y^{\alpha_Y}|X^{\alpha_X},\tau) = 1 - K(Y^{\alpha_Y};X^{\alpha_X},\tau) = 0$ for every level of resolution, which entails $e = 0$ and thus a non-empirical system (definition 6). Two simulations illustrate how K may change as a function of resolution depending on characteristics of the system (in this case, of the shape of the pattern).

□

S8 text

Proof. to simplify the notation, let $\tau \equiv \tau_{Y|X}$ be the theory and $\tau' \equiv \tau'_{Y|X,X',\tau_{Y|X}}$

$$K(Y^{n_Y};X^{nX},X'^{nX},\tau,\tau') > K(Y^{n_Y};X^{nX},\tau) \rightarrow$$
$$\frac{n_Y H(Y) - n_Y H(Y|X,X',\tau,\tau')}{n_Y H(Y) + n_X H(X) + n_X H(X') - \log p(\tau) - \log p(\tau')} > \frac{n_Y H(Y) - n_Y H(Y|X,\tau)}{n_Y H(Y) + n_X H(X) - \log p(\tau)} \rightarrow$$
$$(n_Y H(Y) + n_X H(X) - \log p(\tau))(n_Y H(Y) - n_Y H(Y|X,X',\tau,\tau')) -$$
$$-(n_Y H(Y) + n_X H(X) + n_X H(X') - \log p(\tau) - \log p(\tau'))(n_Y H(Y) - n_Y H(Y|X,\tau)) > 0 \rightarrow$$
$$(n_Y H(Y) + n_X H(X) - \log p(\tau))(n_Y H(Y|X,\tau) - n_Y H(Y|X,X',\tau,\tau')) >$$
$$(n_X H(X') - \log p(\tau'))(n_Y H(Y) - n_Y H(Y|X,\tau)) \rightarrow$$
$$n_Y H(Y|X,\tau) - n_Y H(Y|X,X',\tau,\tau') > (n_X H(X') - \log p(\tau'))K(Y;X,\tau) \rightarrow$$
$$(n_Y H(Y) - n_Y H(Y|X,X',\tau,\tau')) - (n_Y H(Y) - n_Y H(Y|X,\tau)) > (n_X H(X') - \log p(\tau'))K(Y;X,\tau) \rightarrow$$
$$k' - k > \frac{n_X H(X') - \log p(\tau')}{n_Y H(Y)} kh \quad (91)$$

□

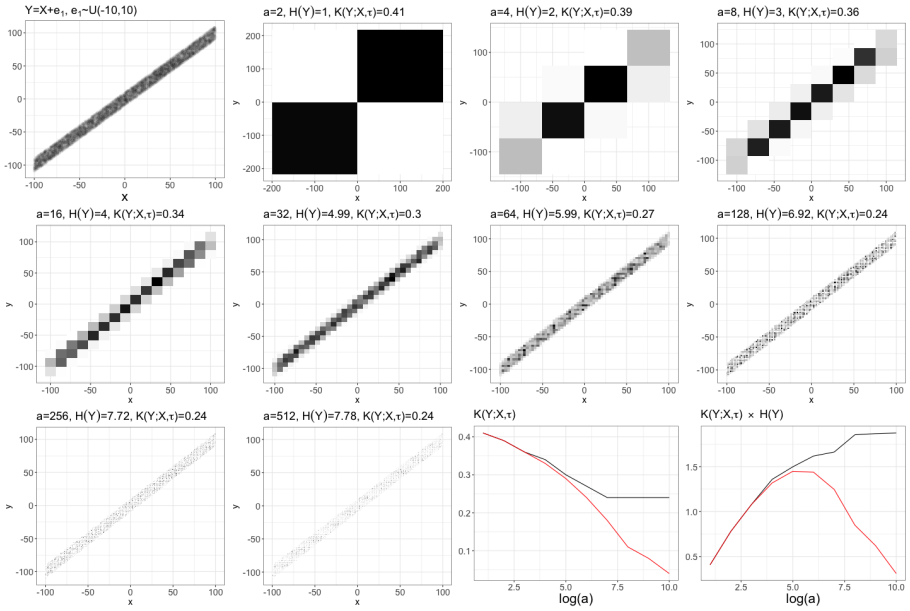
S9 text

Proof.

$$K(Y;XT) = \frac{h}{H(Y)}(H(Y) - H(Y|XT)) = \frac{h}{H(Y)}(H(Y) - H(YXT) + H(XT)) =$$
$$\frac{h}{H(Y)}(H(Y) - H(YXT) + H(X) + H(T)) = \frac{h}{H(Y)}(H(T) - (H(YXT) - H(Y) - H(X)) =$$
$$\frac{h}{H(Y)}(H(T) - (H(T|YX)) = K(T;Y,X) \quad (92)$$

□

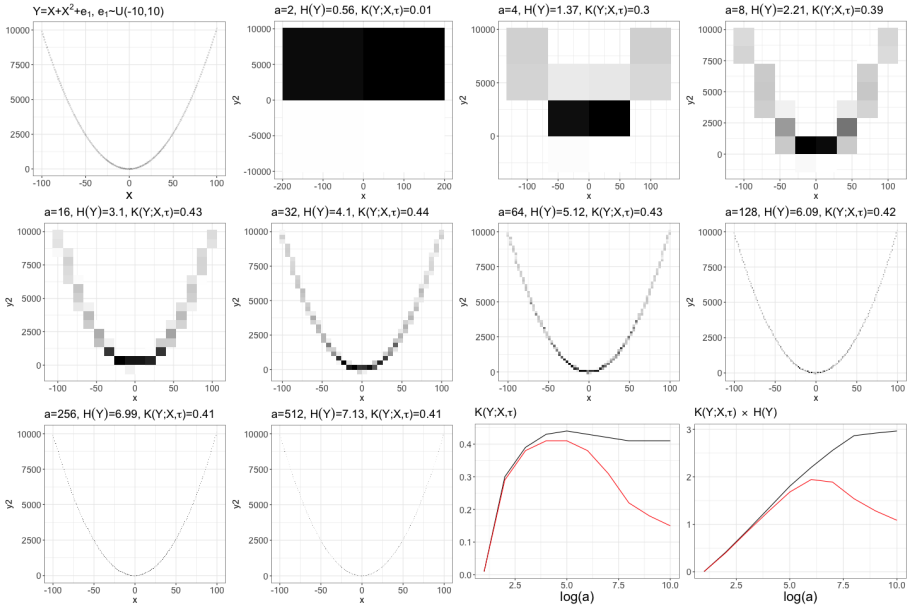
PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018



S7 Figure A Illustrative example of how, as the resolution with which explanandum and explanans is changed, K varies dependong in the shape of the pattern. The figures and all the calculations were derived from a simulated dataset, in which the pattern linking explanandum to explanans was assumed to have a noise with uniform distribution. Black line: entropies and K values calculated by maximum likelihood method (i.e. bin counting). Red line: entropies and K values calculated using the "shrink" method described in [35] (the R code used to generate the figures is provided in S11 text). Note how the value of K and its re-scaled version $H(Y)K$ are maximized at a single optimal resolution.

.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018



S7 Figure B Illustrative example of how, as the resolution with which explanandum and explanans is changed, K varies dependong in the shape of the pattern. The figures and all the calculations were derived from a simulated dataset, in which the pattern linking explanandum to explanans was assumed to have a noise with uniform distribution. Black line: entropies and K values calculated by maximum likelihood method (i.e. bin counting). Red line: entropies and K values calculated using the "shrink" method described in [35] (the R code used to generate the figures is provided in S11 text). Note how the value of K and its re-scaled version $H(Y)K$ are maximized at a single optimal resolution.

.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

S10 text

Proof. Let T be a RV or alphabet $\mathcal{T} = \{\tau_1, \tau_2 ... \tau_z\}$, probability distribution $p(\tau)$ and entropy $H(T) = -\sum_i p(\tau_i) \log p(\tau_i)$. Let T' be a random variable derived from T by removing from its alphabet the element $\tau_j \in \mathcal{T}$ of probability $p(\tau_j)$. Then

$$H(T') = \frac{1}{1 - p(\tau_j)} \sum_{i \neq j} p(\tau_i) \log \frac{1}{p(\tau_i)} - \log \frac{1}{1 - p(\tau_j)} \quad (93)$$

When $|\mathcal{T}| = 2$, $H(T') = 0$ independent of the probability distribution. Otherwise, the value rapidly approaches $H(T)$ as $p(\tau_j)$ decreases (e.g. as the alphabet of T increases in size). Note that under specific conditions $H(T') > H(T)$ - for example, if T equals $p(\tau_j) = 0.9, p(\tau_k) = 0.05, P(\tau_k) = 0.05$). This entails that the uncertainty about a condition might momentarily increase, if the most probable case is excluded. However, the effect is circumscribed since, as more elements are removed from the alphabet, $H(T')$ tends to 0. □

S2 Table

Table 2. 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 recurring to deities or non-observables entitites	Comte 1830 [20]
methodologism	follows rigorous methods for selecting hypotheses, acquiring data, and drawing conclusions	fails to follow the scientific method	e.g. Pearson 1900, Poincare 1914 [90, 100]
verificationism	builds upon verified statements	relies on non-verifiable statements	Wittgenstein 1922 [101]
falsificationism	builds upon falsifiable, non-falsified statements	produces explanations devoid of verifiable counterfactuals	Popper 1959 [58]
methodological falsificationism	generates theories of increasing empirical content, which are accepted when surprising predictions are confirmed	protects its theories with a growing belt of auxiliary hypotheses, giving rise to "degenerate" research programs	Lakatos 1970 [102]
norms	follows four fundamental norms, namely: universalism, communism, disinterestedness, organized scepticism	operates on different, if not the opposite, sets of norms	Merton 1942 [103]
paradigm	is post-paradigmatic, meaning it: 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 [104]
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 [66–68]

S2 Table

S11 text R code used to generate all figures and analyses.

7 Acknowledgments

Marco del Giudice gave helpful comments about the analysis of gender differences in personality.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

Table 3. Properties variably possessed by sciences

Principle	Property or properties	Author, [ref]
scientific hierarchy	simplicity, generality, quantifiability, recency, human relevance	Comte 1830 [20]
consilience	ability to subsume disparate phenomena under general principles	Whewell 1840 [105]
lawfulness	nomoteticity, i.e. interest in finding general laws, as opposed to idioteticity, i.e. interest in characterizing individuality	Windelband 1894 [106]
data hardness	data that resist the solvent influence of critical reflection	Russel 1914 [107]
empiricism	ability to calculate in advance the results of an experiment	Conant 1951 [108]
rigour	rigour in relating data to theory	Storer 1967 [109]
maturity	ability to produce and test mechanistic hypotheses, as opposed to mere fact collection	Bunge 1967 [110]
cumulativity	cumulation of knowledge in tightly integrated structures	Price 1970 [111]
codification	consolidation of empirical knowledge into succinct and interdependent theoretical formulations	Zuckerman and Merton 1973 [75]
consensus	level of consensus on the significance of new knowledge and the continuing relevance of old	Zuckerman and Merton 1973 [75]
core cumulativity	rapidly growing core of unquestioned general knowledge	Cole 1983 [112]
invariance	contextual invariance of phenomena	Humphreys 1990 [74]

References

1. Ioannidis JPA, Fanelli D, Dunne DD, Goodman SN. Meta-research: Evaluation and Improvement of Research Methods and Practices. PLOS BIOLOGY. 2015;13(10). doi:10.1371/journal.pbio.1002264.

2. Goodman SN, Fanelli D, Ioannidis JPA. What does research reproducibility mean? Science Translational Medicine. 2016;8(341). doi:10.1126/scitranslmed.aaf5027.

3. Fanelli D. Is science really facing are producibility crisis, and do we need it to? PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA. 2018;115(11):2628–2631. doi:10.1073/pnas.1708272114.

4. Fiedler K, Schwarz N. Questionable Research Practices Revisited. SOCIAL PSYCHOLOGICAL AND PERSONALITY SCIENCE. 2016;7(1):45–52. doi:10.1177/1948550615612150.

5. Song F, Parekh S, Hooper L, Loke YK, Ryder J, Sutton AJ, et al. Dissemination and publication of research findings: an updated review of related biases. Health Technology Assessment. 2010;14(8):1+. doi:10.3310/hta14080.

6. Wikipedia contributors. Journal of Negative Results in Biomedicine — Wikipedia, The Free Encyclopedia; 2018. Available from: `https://en.wikipedia.org/w/index.php?title=Journal_of_Negative_Results_in_Biomedicine&oldid=831902354`.

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

8. Gorroochurn P, Hodge SE, Heiman GA, Durner M, Greenberg DA. Non-replication of association studies: "pseudo-failures" to replicate? GENETICS IN MEDICINE. 2007;9(6):325–331. doi:10.1097/GIM.0b013e3180676d79.

9. Lipsey MW, Wilson DB. Practical meta-analysis. Applied social research methods series. Sage Publications; 2001.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

10. Higgins J, Thompson S. Quantifying heterogeneity in a meta-analysis. STATISTICS IN MEDICINE. 2002;21(11):1539–1558. doi:10.1002/sim.1186.

11. Shrier I, Platt RW, Steele RJ. Mega-trials vs. meta-analysis: Precision vs. heterogeneity? CONTEMPORARY CLINICAL TRIALS. 2007;28(3):324–328. doi:10.1016/j.cct.2006.11.007.

12. Schnitzer SA, Carson WP. Would Ecology Fail the Repeatability Test? BIOSCIENCE. 2016;66(2):98–99. doi:10.1093/biosci/biv176.

13. Voelkl B, Wurbel H. Reproducibility Crisis: Are We Ignoring Reaction Norms? TRENDS IN PHARMACOLOGICAL SCIENCES. 2016;37(7):509–510. doi:10.1016/j.tips.2016.05.003.

14. Goodman S, Greenland S. Why most published research findings are false: Problems in the analysis. PLOS MEDICINE. 2007;4(4):773. doi:10.1371/journal.pmed.0040168.

15. Moonesinghe R, Khoury MJ, Janssens ACJW. Most published research findings are false- but a little replication goes a long way. PLOS MEDICINE. 2007;4(2):218–221. doi:10.1371/journal.pmed.0040028.

16. Almeida RMVR. The role of plausibility in the evaluation of scientific research. REVISTA DE SAUDE PUBLICA. 2011;45(3):617–620. doi:10.1590/S0034-89102011000300021.

17. Miller J, Ulrich R. Optimizing Research Payoff. PERSPECTIVES ON PSYCHOLOGICAL SCIENCE. 2016;11(5):664–691. doi:10.1177/1745691616649170.

18. Park IU, Peacey MW, Munafo MR. Modelling the effects of subjective and objective decision making in scientific peer review. NATURE. 2014;506(7486):93+. doi:10.1038/nature12786.

19. Smaldino PE, McElreath R. The natural selection of bad science. Royal Society Open Science. 2016;3(9). doi:10.1098/rsos.160384.

20. Comte A. Cours de philosophie positive. vol. 6 vols. Paris: Rouen first, then Bachelier; 1830-1842.

21. Mach E. The economical nature of physical inquiry. In: Popular Scientific Lectures by Ernst Mach [1895]. The Open Court Publishing Co.; 1882. p. 186–213.

22. Shannon CE. A mathematical theory of Communication. The Bell system technical journal. 1948;27:379–423.

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

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

25. Rissanen J. Modeling by shortest data description. Automatica. 1978;14(5):465–458. doi:doi:10.1016/0005-1098(78)90005-5.

26. Grünwald PD. The Minimum Description Length Principle. Adaptive computation and machine learning. MIT Press; 2007.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

27. Hutter M. Universal Artificial Intelligence: Sequential Decisions Based on Algorithmic Probability. Texts in Theoretical Computer Science. An EATCS Series. Springer Berlin Heidelberg; 2004.

28. Dretske FI. Knowledge and the Flow of Information. Bradford Books. MIT Press; 1983.

29. Thagard P. Computational Philosophy of Science. Bradford Books. A Bradford; 1988.

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

31. Maruyama K, Nori F, Vedral V. Colloquium: The physics of Maxwell's demon and information. Rev Mod Phys. 2009;81:1–23. doi:10.1103/RevModPhys.81.1.

32. H BC. The thermodynamics of computation - a review. International Journal of Theoretical Physics. 1982;21(12):905–940.

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

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

35. Hausser J, Strimmer K. Entropy Inference and the James-Stein Estimator, with Application to Nonlinear Gene Association Networks. JOURNAL OF MACHINE LEARNING RESEARCH. 2009;10:1469–1484.

36. Wiles A. Modular Elliptic Curves and Fermat's Last Theorem. Annals of Mathematics. 1995;141(3):443–551.

37. Olive KA, Agashe K, Amsler C, Antonelli M, Arguin JF, Asner DM, et al. REVIEW OF PARTICLE PHYSICS Particle Data Group. CHINESE PHYSICS C. 2014;38(9). doi:10.1088/1674-1137/38/9/090001.

38. type; 2018 [cited 17]. Available from: `https://eclipse.gsfc.nasa.gov/eclipse.html`.

39. Myles S, G WJ. 8: Orbital Ephemerides of the Sun, Moon, and Planets. In: Urban, Seidelmann, editors. Explanatory Supplement to the Astronomical Almanac. 3rd ed. University of Virginia; 2013.

40. type; 2018 [cited 17]. Available from: `https://eclipse.gsfc.nasa.gov/SEsearch/SEdata.php?Ecl=+30000426`.

41. Chavalarias D, Cointet JP. Phylomemetic Patterns in Science Evolution-The Rise and Fall of Scientific Fields. PLOS ONE. 2013;8(2). doi:10.1371/journal.pone.0054847.

42. Wilson EO. Consilience: The Unity of Knowledge. Knopf Doubleday Publishing Group; 2014.

43. Nonacs P, Hager R. The past, present and future of reproductive skew theory and experiments. BIOLOGICAL REVIEWS. 2011;86(2):271–298. doi:10.1111/j.1469-185X.2010.00144.x.

44. Reeve H, Starks P, Peters J, Nonacs P. Genetic support for the evolutionary theory of reproductive transactions in social wasps. PROCEEDINGS OF THE ROYAL SOCIETY B-BIOLOGICAL SCIENCES. 2000;267(1438):75–79. doi:10.1098/rspb.2000.0969.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

45. Hyde J. The gender similarities hypothesis. AMERICAN PSYCHOLOGIST. 2005;60(6):581–592. doi:10.1037/0003-066X.60.6.581.

46. Del Giudice M, Booth T, Irwing P. The Distance Between Mars and Venus: Measuring Global Sex Differences in Personality. PLOS ONE. 2012;7(1). doi:10.1371/journal.pone.0029265.

47. type; 2012 [cited 3 May 2018]. Available from: http://journals.plos.org/plosone/article/comment?id=10.1371/annotation/2aa4d091-db7a-4789-95ae-b47be9480338.

48. Booth P Tom; Irwing. Sex differences in the 16PF5, test of measurement invariance and mean differences in the US standardisation sample. Personality and Individual Differences. 2011;50(5):553–558.

49. Collaboration OS. Estimating the reproducibility of psychological science. Science. 2015;349(6251). doi:10.1126/science.aac4716.

50. Patil P, Peng RD, Leek JT. What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. Perspect Psychol Sci. 2016;11(4):539–44. doi:10.1177/1745691616646366.

51. Etz A, Vandekerckhove J. A Bayesian Perspective on the Reproducibility Project: Psychology. PLoS One. 2016;11(2):e0149794. doi:10.1371/journal.pone.0149794.

52. Stanley DJ, Spence JR. Expectations for Replications: Are Yours Realistic? Perspect Psychol Sci. 2014;9(3):305–18. doi:10.1177/1745691614528518.

53. Gilbert DT, King G, Pettigrew S, Wilson TD. Comment on "Estimating the reproducibility of psychological science". Science. 2016;351(6277):1037. doi:10.1126/science.aad7243.

54. Bench SW, Rivera GN, Schlegel RJ, Hicks JA, Lench HC. Does expertise matter in replication? An examination of the reproducibility project: Psychology. JOURNAL OF EXPERIMENTAL SOCIAL PSYCHOLOGY. 2017;68:181–184. doi:10.1016/j.jesp.2016.07.003.

55. Ramscar M. Learning and the replicability of priming effects. CURRENT OPINION IN PSYCHOLOGY. 2016;12:80–84. doi:10.1016/j.copsyc.2016.07.001.

56. Van Bavel JJ, Mende-Siedlecki P, Brady WJ, Reinero DA. Contextual sensitivity in scientific reproducibility. Proc Natl Acad Sci U S A. 2016;113(23):6454–9. doi:10.1073/pnas.1521897113.

57. Polanyi M. Personal Knowledge. Taylor & Francis; 2012.

58. Popper KR, Popper KR. The Logic of Scientific Discovery. Harper Torchbooks. HarperCollins Canada, Limited; 1959.

59. Nelson LD, Simmons JP, Simonsohn U. Let's Publish Fewer Papers. PSYCHOLOGICAL INQUIRY. 2012;23(3):291–293. doi:10.1080/1047840X.2012.705245.

60. de Winter J, Happee R. Why Selective Publication of Statistically Significant Results Can Be Effective. PLOS ONE. 2013;8(6). doi:10.1371/journal.pone.0066463.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

61. van Assen MALM, van Aert RCM, Nuijten MB, Wicherts JM. Why Publishing Everything Is More Effective than Selective Publishing of Statistically Significant Results. PLOS ONE. 2014;9(1). doi:10.1371/journal.pone.0084896.

62. Fanelli D. 'Positive' Results Increase Down the Hierarchy of the Sciences. PLoS ONE. 2010;5(4):1–10. doi:10.1371/journal.pone.0010068.

63. Fanelli D. Positive results receive more citations, but only in some disciplines. SCIENTOMETRICS. 2013;94(2):701–709. doi:10.1007/s11192-012-0757-y.

64. Fanelli D. How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. PLoS One. 2009;4(5):e5738. doi:10.1371/journal.pone.0005738.

65. Steneck N. Fostering integrity in research: Definitions, current knowledge, and future directions. SCIENCE AND ENGINEERING ETHICS. 2006;12(1, SI):53–74. doi:10.1007/PL00022268.

66. 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. p. 111–128.

67. Dupre JA. The Disorder of Things. Metaphysical foundations of the disunity of science. Harvard University Press; 1993.

68. 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. p. 9–28.

69. Fuller S. Dissent Over Descent: Intelligent Design's Challenge to Darwinism. Icon; 2008.

70. Carlson S. A double-blind test of astrology. Nature. 1985;318(5):419–425.

71. Braxton JM HL. Variation among academic disciplines: Analytical frameworks and research. In: Higher education: handbook of theory and research.. vol. XI. New York: Agathon Press.; 1996. p. 1–.

72. 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.

73. Fanelli D, Glanzel W. Bibliometric Evidence for a Hierarchy of the Sciences. PLoS ONE. 2013;8(6):1–11. doi:10.1371/journal.pone.0066938.

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

75. 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.

76. Fanelli D, Costas R, Ioannidis JPA. Meta-assessment of bias in science. Proc Natl Acad Sci U S A. 2017;114(14):3714–3719. doi:10.1073/pnas.1618569114.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

77. Steup M. Epistemology. In: Zalta EN, editor. The Stanford Encyclopedia of Philosophy. summer 2018 ed. Metaphysics Research Lab, Stanford University; 2018.

78. Dretske FI. Epistemology and Information. In: van Benthem; Paul Thagard; John Woods PADMGJ, editor. Philosophy of Information. vol. 8 of Handbook of the philosophy of science. North Holland; 2008.

79. Michalowicz JV, Nichols JM, Bucholtz F. Calculation of Differential Entropy for a Mixed Gaussian Distribution. ENTROPY. 2008;10(3):200–206. doi:10.3390/entropy-e10030200.

80. Abdolsaeed Toomaj RZ. Some New Results on Information Properties of Mixture Distributions. Filomat. 2017;31(13):4225–4230.

81. type; 2018 [cited 2018-05-23]. Available from: https://www.equator-network.org/reporting-guidelines/.

82. Lloyd S. Computational Capacity of the Universe. Phys Rev Lett. 2002;88:237901. doi:10.1103/PhysRevLett.88.237901.

83. Hardy L. Quantum ontological excess baggage. STUDIES IN HISTORY AND PHILOSOPHY OF MODERN PHYSICS. 2004;35B(2):267–276. doi:10.1016/j.shpsb.2003.12.001.

84. Ten Yong T. Failure of ontological excess baggage as a criterion of the ontic approaches to quantum theory. STUDIES IN HISTORY AND PHILOSOPHY OF MODERN PHYSICS. 2010;41(4):318–321. doi:10.1016/j.shpsb.2010.04.002.

85. Fitch WT. Toward a computational framework for cognitive biology: Unifying approaches from cognitive neuroscience and comparative cognition. Physics of Life Reviews. 2014;11(3):329 – 364. doi:http://doi.org/10.1016/j.plrev.2014.04.005.

86. Kleene SC, Beeson MJ. Introduction to Metamathematics. Ishi Press International; 2009. Available from: https://books.google.co.uk/books?id=HZAjPwAACAAJ.

87. Chaitin GJ. Meta Maths!: The Quest for Omega. A Peter N. Névraumont book. Vintage Books; 2006.

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

89. Bateson G. Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology. University of Chicago Press; 1972.

90. Poincaré H, Maitland F. Science and method. London: T. Nelson and Sons; 1914.

91. Moles AA, Cohen J. Information Theory and Esthetic Perception. Illini Books. University of Illinois Press; 1968.

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

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018

94. West-Eberhard MJ. Developmental Plasticity and Evolution. OUP USA; 2003.

95. Wolff JG. Unifying computing and cognition. CognitioniResearch.org; 2006.

96. Slingerland E. What Science Offers the Humanities: Integrating Body and Culture. Cambridge University Press; 2008.

97. Schmidhuber J. Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990-2010). IEEE Transactions on Autonomous mental Development. 2010;2(3):230–247. doi:10.1109/TAMD.2010.2056368.

98. Hurley MM, Dennett DC, Adams RB. Inside Jokes: Using Humor to Reverse-engineer the Mind. MIT Press; 2011.

99. Gallager RG. In: Finite State Markov Chains. Boston, MA: Springer US; 1996. p. 103–147. Available from: `https://doi.org/10.1007/978-1-4615-2329-1_4`.

100. Pearson K. The Grammar of Science. 2nd ed. London: A. and C. Black; 1900.

101. Wittgenstein L. Tractatus Logico-Philosophicus. New York: Harcourt, Brace & company, Inc.; 1922.

102. Lakatos I. Falsification and the Methodology of Research Programs. In: Criticism and the Growth of Knowledge. Cambridge: Cambridge University Press.; 1970. p. 91–97.

103. Merton RK. The Normative Structure of Science. In: The Sociology of Science: Theoretical and Empirical Investigations. Chicago: University of Chicago Press; 1942(1973). p. 267–180.

104. Kuhn TS. The Structure of Scientific Revolutions. 2nd ed. Chicago: The University of Chicago Press; 1970.

105. Whewell W. The philosophy of the Inductive Sciences: Founded Upon Their History. London: J.W. Parker; 1840.

106. Windelband W. History and Natural Science. Theory & Psychology. 1894 (1998);8(1):5–22.

107. Russell B. Our knowledge of the external world as a field for scientific method in philosophy. Chicago: The Open Court Publishing Co.; 1914.

108. Conant JB. Science and Common Sense. New Haven: Yale University Press; 1951.

109. Storer NW. Hard sciences and soft - Some sociological observations. Bulletin of the Medical Library Association. 1967;55(1):75?84.

110. Bunge M. The maturation of science. In: Lakatos I, Musgrave A, editors. Problems in the Philosophy of Science. vol. 3. Amsterdam: Norh-Holland Publishing Company; 1967. p. 120?137.

111. 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.

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

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.1968v4 | CC BY 4.0 Open Access | rec: 25 Jun 2018, publ: 25 Jun 2018