The “Tau” of Science

How to Measure, Study, and Integrate Quantitative and Qualitative Knowledge

Daniele Fanelli

Department of Methodology, London School of Economics and Political Science
email@danielefanelli.com

January 7, 2022

Abstract

Scientists’ ability to integrate diverse forms of evidence and evaluate how well they can explain and predict phenomena, in other words, *to know how much they know*, struggles to keep pace with technological innovation. Central to the challenge of extracting knowledge from data is the need to develop a metric of knowledge itself.

A candidate metric of knowledge, K , was recently proposed by the author. This essay further advances and integrates that proposal, by developing a methodology to measure its key variable, symbolized with the Greek letter τ (“tau”). It will be shown how a τ can represent the description of any phenomenon, any theory to explain it, and any methodology to study it, allowing the knowledge about that phenomenon to be measured with K .

To illustrate potential applications, the essay calculates τ and K values of: logical syllogisms and proofs, mathematical calculations, empirical quantitative knowledge, statistical model selection problems, including how to correct for “forking paths” and “P-hacking” biases, randomised controlled experiments, reproducibility and replicability, qualitative analyses via process tracing, and mixed quantitative and qualitative evidence.

Whilst preliminary in many respects, these results suggest that K theory offers a meaningful understanding of knowledge, which makes testable metascientific predictions, and which may be used to analyse and integrate qualitative and quantitative evidence to tackle complex problems.

Contents

1	Introduction	2
1.1	The K function	3
1.2	Gaps in the current quantification of τ	4
2	Overview of methods proposed	5
2.1	T as a set of nested hypergraphs	5
2.2	The nesting operator	5
2.3	Calculating the size of T	7
2.4	The recursive function $D()$	7
2.5	The cumulation operator	8
2.6	Divergence between two τ and other operations	9
3	Implications of this methodology	9
3.1	Generalization of K	10
3.1.1	A recursive interpretation of K	11
3.1.2	Knowledge as lossy information compression	11

3.1.3	The total K of a τ	11
3.2	How to measure the K of anything	12
4	Results - examples of applications	12
4.1	Abstract (logical-deductive) knowledge	13
4.1.1	Example 1.1: Modus ponens	13
4.1.2	Example 1.2: Syllogism	14
4.1.3	Example 1.3: Logical proof and creation of new knowledge	15
4.1.4	Example 1.4: Mathematical knowledge	17
4.2	Empirical, quantitative knowledge	19
4.2.1	Example 2.1: Predicting solar eclipses, with conditioning	19
4.2.2	Example 2.2: Statistical model selection	20
4.3	Metascientific knowledge	24
4.3.1	Describing a study's theory	24
4.3.2	Describing a study's methodology	26
4.3.3	Example 3.1: The K of randomized controlled experiments	26
4.3.4	Example 3.2 Measuring and studying replicability	27
4.4	Empirical qualitative knowledge	30
4.4.1	Example 4.1: Process tracing	30
4.5	Merging qualitative and quantitative knowledge	32
5	Discussion	33
6	Extended Methods	34
6.1	T as a set of hypergraphs	34
6.1.1	Principled definition of τ	34
6.1.2	Size of T	35
6.1.3	Knowledge hypergraph	35
6.1.4	The nesting operator	36
6.1.5	Nesting of entire τ	37
6.1.6	Size of T and description length of τ	38
6.2	D function	39
6.2.1	D() as a universal descriptor of τ	39
6.3	Operations on τ	42
6.3.1	The cumulation operator	42
6.3.2	Divergence between τ	44
6.3.3	Conditioning of τ	45
6.3.4	Conceptualization	45
6.4	Knowledge as information compression	46
6.5	The total K of a τ	47
7	APPENDIX A1	48

1 Introduction

The scientific and societal progress of humanity depends on its ability to tackle problems of increasing complexity, by integrating ever more diverse sources of evidence and evaluating how well theories, hypotheses and interventions succeed in explaining and predicting observations. Technological innovations have empowered a growing population of researchers to collect, process and share data of all types. However, scientists' ability to integrate that evidence and evaluate how well they can explain and predict phenomena, in other words, *to know how much knowledge they have or lack*, has lagged behind.

Central to the challenge of extracting knowledge from data is the need to develop a metric of knowledge itself. A candidate metric of knowledge, K , was recently proposed in [1], and this essay develops a methodology to measure its key variable, symbolized with the Greek letter τ (“tau”).

A τ , this essay will argue, can represent the description of any phenomenon, any theory to explain it, and any methodology to study it, allowing the knowledge about that phenomenon to be measured with K . In addition to offering a meaningful metric of knowledge, this approach makes testable predictions concerning, for example, the replicability of research findings. Finally, the essay will suggest how a K value can be calculated for qualitative evidence and how it can be used to combine qualitative and quantitative evidence to evaluate competing theories.

The essay is structured as follows. Section 1.1 overviews the K metric, and section 1.2 illustrates the limitations in measuring τ that this essay addresses. Section 2 offers a concise description of the key methods proposed, which are laid out in full mathematical detail in section 6. Section 3 illustrates some new insights derived from this approach, including a generalization of K and a substantive interpretation of its meaning. The “results” section 4 gives practical examples of how a K value can be derived from various forms of knowledge, including: logic 4.1, empirical quantitative knowledge 4.2, meta-scientific knowledge 4.3, empirical qualitative knowledge 4.4, and mixed quantitative and qualitative knowledge 4.5.

1.1 The K function

The approach that we shall call “ K theory” [1] proposes that all expressions of knowledge, in scientific research and beyond, can be represented and quantified as relations between information quantities, and therefore that a theory and methodology to quantify knowledge can be derived from a quantity K calculated by the “ K function”

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

that expresses how much of the information contained in an *explanandum* Y is compressed by an explanans consisting of an information input X and an information conditioning factor τ , which may encode both theoretical and methodological factors. The quantifiers n_Y and n_X represent the number of repetitions, i.e. independent instantiations of the corresponding random variables, and the information of any random variable Z is Shannon’s entropy, which is the expected value of the information content of each instantiation of the variable, i.e.

$$H(Z) = \sum p(z) \log \frac{1}{p(z)} \equiv E[-\log p(z)] \quad (2)$$

Equation 1 is proposed to characterise any structure composed of entities and relations, which in the broadest sense is referred to as a “system”. To any extent that such a system expresses knowledge, it reduces the uncertainty about some random variable Y by processing the side information X via a structure τ .

The formulation in 1 is more general than it might appear, due to several known properties of information. First, any random variable Z can result from the joining of any number of other random variables, independently distributed or otherwise, and therefore X, Y, τ stand for any number of inputs, explananda and structure components. Second, for any discrete random variable Z with alphabet size $|\mathcal{Z}|$, $H(Z) \leq \log |\mathcal{Z}|$, with equality when Z is uniformly distributed. Thus, the distinction in equation 1 between information quantities for Y, X versus τ are mainly formal and intend to symbolize the fact that τ describes a single instantiation of an ensemble, whereas Y, X will more typically represent random variables [1].

Third, theorems in algorithmic information theory show that if z^n is a string of length n generated by a source (a random variable) Z , then

$$\lim_{n \rightarrow \infty} \frac{C(z^n)}{n} = H(Z) \quad (3)$$

where $C(z^n)$ is the Kolmogorov complexity of the string, i.e. the length of the shortest program that, fed into a Universal Turing Machine, would print the string and halt [2]. Kolmogorov complexity is an intrinsic property of any string (and by extension of any object described by a string) and, whilst technically uncomputable in the general case, it can be approximated empirically or calculated exactly for simpler objects, including graphs [3]. Thus, the terms in equation 1 are information quantities in the most general sense, and they may represent random variables or descriptions of objects [1].

Manipulations of equation 1 were shown to offer mathematical answers to questions in meta-research and in philosophy of science, raising the prospect that K theory might constitute a general paradigm for metascience [1]. Further, since the K function is derived from two simple postulates about the nature of information and knowledge, the same theory would seem applicable to qualitative forms of knowledge and indeed any manifestation of cognition [1].

1.2 Gaps in the current quantification of τ

As proposed in [1] the theory/methodology component τ encodes all the conditions that are necessary to observe a specified relation between X and Y . In the context of scientific knowledge, for example, τ was said to encode both the underlying theory (i.e. mathematical relations, models etc.) as well as the methodology (e.g. a listing of the experimental conditions and the sequence of procedures) underlying a particular claim.

Formally, τ is an instantiation of a uniform random variable T with an alphabet of size $|\mathcal{T}|$, and thus $-\log p(\tau)$ is the information required to “choose” a τ amongst a set of $|\mathcal{T}|$ alternatives. It was suggested in [1] that, this quantity, whilst hard or impossible to calculate exactly, could be estimated partially and in relative terms, by assuming that τ is a sequence of independent and uniformly distributed random variables, such that

$$\log \frac{1}{p(\tau)} = \sum_i \log \frac{1}{p(\tau_i)} \quad (4)$$

where each τ_i would correspond to particular elements of the whole (e.g. each step in a sequence of methodological decisions).

As will be shown, this approach is not fundamentally incorrect but is incomplete, due to limitations including the following:

1. Any information describing the *structure* of τ is not formally computed in 4. For example, if $\tau = (\tau_1, \tau_2)$ is a sequence of two RV, then $\log p(\tau_1, \tau_2) = \log p(\tau_2, \tau_1)$ and the order does not make any difference. However, it *could* be the case that the order of the RV is important, in which case we should expect the description to be longer than otherwise, because it should include the additional information needed to encode the order. Intuitively, this information will be encoded in just another addendum, but this intuition was not formalized in [1].
2. The description length of τ depends on potentially arbitrary decisions pertaining to how a system is described. This is a known limitation of any measure of descriptive complexity [4, 5], but for our purposes it acquires particular relevance since how a system is described affects the quantity of K as well as, from a theoretical perspective, the actual knowledge that K is intended to quantify. Intuitively, if the same encoding procedure is used, then the relative description length of two τ 's should be preserved, but this, again, was not formalized in [1].
3. The concept of complexity, which τ is intended to capture, has at least two meanings that are connected and yet distinguishable: 1) the *ontic* complexity of the object being described; 2) the *epistemic* complexity reflected in how much is knowable or describable about that object [5]. Terms in equation 1 may capture both concepts depending on the problem, but in either case their quantification refers to a description that will be typically partial and incomplete, as it omits elements that are either not known or constant across systems. For example, in quantifying the complexity of a mathematical equation we would not care to include a description of all elementary operations, and yet these quantities are implicitly contributing to the value of K .

This article will integrate the theory and methodology proposed in [1] by defining a general approach to quantifying the description length of the τ component and of K , and showcase its implementation in various forms of knowledge.

2 Overview of methods proposed

To quantify the τ we shall follow a three-pronged strategy, which combines ideas and techniques from multiple methodologies, including graphical knowledge representation [6, 7, 8], minimum description length principle [9, 10] and probabilistic graphical modeling [11, 12]. The key methodological innovations proposed in this essay are: 1) a particular language of graphical knowledge representation, which allows to represent the τ of any object; 2) A formula to calculate the size of the set T based on that graphical language; 3) a recursive description function D that formalizes the multilevel nature of phenomena and their descriptions, allowing a formal estimation of $\log p(\tau)$ and K in the absence of complete information on τ . These three strategies and their implications will be briefly overviewed below, whereas more detailed technical justifications are given in the Extended Methods section 6.

2.1 T as a set of nested hypergraphs

It follows from the definition knowledge that any τ , can be described by a labelled nested hypergraph - that is, a graph in which edges may connect more than two vertices, and in which both vertices and edges are identified by symbols (see 6.1.1). The symbols we will use in the τ graphs are just pointers that identify vertices or relations has having same or different underlying τ describing them. The sameness or difference of any element of a τ , in other words, is encoded in lower-level descriptions consisting of other graphs (other τ). The τ , therefore, is an endlessly recursive structure of structures (Figure 1).

To describe and estimate the size of a τ we shall use a simple two-dimensional graphical language consisting of ordinary labels and relations, very similar to those used in knowledge representation (e.g. [6, 7]). An example of a very simple τ is:

which represents a “theory” that two inputs labelled A, B are related to the explanandum labelled C by the relation labelled r . We can also represent these hypergraphs in text following the convention that relations (hyper-edges) are represented by lower-case letters, concepts by upper-case letters or numbers, hyperedges are read from left to right, and tails of hyperedges are read counterclockwise starting from the top left. The graph above can thus be described as $ABrC$ or, using parenthesis to avoid confusion, $((AB)rC)$.

In a graph representing a τ , like the one above, relations are not necessarily causal or even statistical, and the graphs are not necessarily acyclic, as they are in formal treatments of causation (e.g. [13]). Indeed, “relation” must be intended here in the broadest and most generic possible sense, because the nature or meaning of any relation is specified separately, by the nested description of the relation itself. Similarly, the labels have no intrinsic meaning, other than the information they convey by being different from other symbols, thus pointing to different underlying description.

2.2 The nesting operator

To nest a lower-level description within a graph element, we use a nesting operator and a nesting graphical symbology. Following from the example given above, we may use

Figure 1: Pictorial representation of τ as a recursive multi-level structure, of structures, of structures, etc. Mathematically, the structures are represented as adjacency tensors, which connect vertices (represented by elements on each side of the cubes) via relations (represented by elements within the cubes). Note how both relations and nodes in any one of the tensors are also tensors, and how one of the nodes may consist in a “null” element, here depicted as an empty square (a “null” description), which allows relations of lower arity to be represented in higher dimensions. The relations depicted have only arity 2 or 3, which is reflected in the number of dimensions of the tensors. Whilst an arity of 2 or 3 is common, relations can have arity of any size, and the resulting tensor will have a corresponding number of dimensions.

Figure 2: Complexity of τ in relation to variables in the hypergraph representing it. Each plot shows the value of equation 5 when varying one parameter, indicated as x , and holding the other parameters constant. Parameters legend: d = number of vertices ; r = number of relations, a = maximum arity

to represent graphically the fact that the vertex A contains a lower-level description 12pA, which describes A as being the explanandum of two inputs 1, 2 and the relation p . Using the nesting operator, we can express this as $((12pA) \circledast A)BrC$. The graph can be “denested” by reversing this operation. This we can represent as a reversion of the nesting operator itself: $((12pA) \circledast A)BrC = (12pA)BrC$, which gives a single τ resulting from the joining of the two graphs above:

We must stress again that the labels are a mere representational device and have no intrinsic meaning. They are mere pointers to further graphs, that is, to other τ . The τ is, in other words, just a structure of structures of structures...

The mathematical representation of the τ is that of an adjacency tensor, of adjacency tensors, of adjacency tensors, etc. Where by “tensor” we mean the multidimensional equivalents of a two-dimensional adjacency matrix (see section 6.1.2), as depicted in Figure 1.

2.3 Calculating the size of T

Applying a modified version of Minimum Description Length techniques proposed for graphs (e.g. [14]) it is possible to derive an optimal encoding of the labelled adjacency tensor, and from it the size of the set T of which a given τ is member. Specifically, the description length of any τ with d vertices, r relations of maximal arity a , label multiplicities m and relation symmetries $s(e)$ is given by the equation:

$$\log |T| = \log \binom{d^a}{r} + \log \binom{d}{m_{v1}, m_{v2} \dots m_{vp}} + \log \binom{r}{m_{e1}, m_{e2} \dots m_{eq}} - \sum_{e_i=1}^q \log s(e_i) \quad (5)$$

(see 6.1.6 for more details), in which the terms on the right-hand side quantify, respectively, the description of the structure of the graphs, the diversity of labels of vertices and of relations, and any symmetries in the relations that reduces the total description length. This equation quantifies the size of a single-level description. To accommodate the nested nature of the structure, a recursive function is introduced.

The resulting value varies considerably depending on an interplay between all parameters characterizing the graph (Figure 2).

2.4 The recursive function $D()$

Using the nesting operator $\circledast$ for graphs, we define a “description” recursive function as:

$$D(\tau) = 0 \text{ if } T = \emptyset \quad (6)$$

$$D(\tau) = \log \frac{1}{p(\tau)} + \sum_{\tau_i \supset \tau} D(\tau_i) \text{ otherwise} \quad (7)$$

where $p(\tau) = \frac{|\theta|}{|T|}$ is calculated using equation 73 (see 6.2 for a complete definition). This function formalizes the fact that, to obtain the complete description length of a τ , we should, at least in principle, add the description length of all nested components. Note that the D function is analogous to the summation in equation 4, but now, unlike in [1], it encodes the structural properties of the τ at each level of description.

The total description length of a τ is calculated recursively, from the top-most level of description downwards. In principle, any description could go down infinitely many steps, either by repeating a pattern or genuinely containing new descriptions every time. In practice, however, all descriptions have an end, which in our representation corresponds to the level below which no further descriptions are given.

For example, in the simple nested graph given above, i.e. $((12pA) \supset A)BrC$, the τ to be quantified is the entirety of the graph, and thus its D value is given by the recursive application of equation 5 to each level of description in turn. Since in our example we only gave a description of A :

$$\begin{aligned} D(\tau) &= 7.34 + D(A) + D(B) + D(C) + D(r) = \\ &= 7.34 + 7.34 + D(1) + D(2) + D(p) + D(A) + D(B) + D(C) + D(r) = \\ &= 14.68 + \sum_{\tau_i \supset \tau} D(\tau_i) \end{aligned} \quad (8)$$

where the value 7.34 is derived from equation 5 assuming no symmetry in the relation. If no lower level description is available for the other terms, then $D(\tau) = 14.68 + 0 + 0 + 0 + 0 + 0 + 0 + 0 = 14.68$. Otherwise, we either continue adding the known terms or we could stop at the result $D(\tau) < 14.68$, which is a lower bound of the true value.

2.5 The cumulation operator

Information is always generated by the encoding of differences at one level of description, between objects that are the same at another level. For example, results of a classic trial of picking coloured balls from a jar entails that, at one level, all balls from the jar are the same and, at another level, that they differ in some property that is being recorded (differences in colour). We formalize this process with a ‘‘cumulation’’ operator $\oplus$, which forms strings of τ ’s that are connected by a common higher-level description. This operation can be applied recursively to τ structures of any complexity that are nested into any number of levels, element-by element (that is, tensor cell by tensor cell), level by level as:

$$D(\bigoplus^n \tau) = \log \frac{1}{p(\tau_1)} + \sum_{\tau_a \supset \tau_1}^n D(\bigoplus^n \tau_a) \text{ if } T_1 = T_2 \dots = T_n \neq \emptyset \quad (9)$$

$$D(\bigoplus^n \tau) = nH(W) + \sum D(\tau_w) \text{ otherwise} \quad (10)$$

where $T_i = T_j \iff d_i = d_j, r_i = r_j$ etc. means that the two τ at that level of description have the same structure, that is, identical adjacency matrix and analogous distribution of labels (pointers) and $H(W)$ is the entropy of the random variable generated by the distribution of divergences (different lower-level descriptions).

Graphically, this corresponds to taking the top-most level of description and: if the two graphs are homologous in the sense above, then add the description length of that structure and proceed to

do the same with each corresponding subset of vertices and relations within it; if the two graphs are different, then construct a string of the two descriptions as described in section 6.2.1.

For example, let us cumulate two τ , $(Ar([BpD](\succ B))$ and $Ar([BCqD](\succ B))$, which are identical at higher levels, but not in their description of B

then

$$\begin{aligned}
 D(Ar([BpD](\succ B)) \oplus As([BCqD](\succ B))) &= \\
 3 + D(A \oplus A) + D(r \oplus s) + D(B \oplus B) &= \\
 3 + [2 \times 0 + D(A)] + [2 \times 1 + D(r) + D(s)] + [2 \times 1 + D(BpD) + D(BCqD)] &= \text{etc.} \quad (11)
 \end{aligned}$$

(see 6.3.1 for more details).

2.6 Divergence between two τ and other operations

From the description length of the cumulation of τ we can derive a measure of divergence between τ , which reflects the extent to which two or more structures “branch” out and away from each other. Formally, in the case of two τ , the measure of divergence is given by

$$D(\tau_i | \tau_j) = D(\tau_i \oplus \tau_j) - D(\tau_j) = d + \sum_{a \not\sim \tau_j} D(\tau_a) \quad (12)$$

where $\tau_i \oplus \tau_j$ is the cumulation normalized by setting $n = 1$ in equation 9, and d is the number of divergent nested τ . This number can be directly observed and derived from the graphical representations of two τ , moving from the top-most level down and counting the number of elements that are homologous at that level but have different lower-level descriptions.

Equation 12 states that the description length of a τ conditioned on another τ equals the number of diverging elements, plus the nested descriptions of these diverging elements. If we ignore the latter component or assume that it is constant for all τ ’s, then this measure of divergence has all properties of a metric of distance.

Other operations that we define in this essay include conditioning (6.3.3) and conceptualization (6.3.4). In future work, it may be possible to derive further operations, either specific to τ , like the ones above, or derived from ordinary graph and hypergraph theory (e.g. [15, 16]).

3 Implications of this methodology

The combination of graphical language, operators and quantifiers described above allows to circumvent numerous conceptual and practical obstacles to the quantification of τ , leading to a more general interpretation and application of τ and K . In particular, these methods allow to:

1. Encode and quantify the structural properties of any τ .
2. Express the multilevel and multifaceted nature of any describable object.
3. Accommodate the arbitrary nature of any description, whilst avoiding subjectivity in measuring the complexity of that description and its divergence from other descriptions.
4. Separate and combine epistemic and ontic complexity. The knowledge that is measured is often if not always derived from incomplete descriptions and, in this sense, always reflects epistemic complexity. However, the recursive formalism of τ formally encompasses the complete description

of an object as a (potentially even infinite) series of additive components, and in this sense it can represent the ontic complexity of a phenomenon. In either case, a K value can be derived.

These properties lead to a generalization of K and new insights into its meaning.

3.1 Generalization of K

Section 6.2 shows that the D function 7 quantifies not just the information content of τ but also of symbols, random variables and their instantiations. It follows that equation 1 may be reformulated and generalized as

$$K(Y^{n_Y}; X^{n_X}, \tau_{xy}) = \frac{D(\bigoplus^{n_Y} \tau) - D(\bigoplus^{n_Y} \tau')}{D(\bigoplus^{n_Y} \tau)} \equiv \frac{D(\tau) - D(\tau')}{D(\tau)} = K(\tau, \tau') \quad (13)$$

where the last equivalence follows simply from the fact that a string of τ is just another τ , and which says that K quantifies the amount of redundancy (i.e. of structure, of predictability) present in a system (a τ) that can be in two states (e.g. before/after, with/without, etc.), across n instantiations (i.e. events, experiences, measurements).

If the τ in equation 13 is completely described, then K represents an exactly measurable property of the system being examined. The value may be relative to the description strategy (and scientific progress may be characterised as a process of finding the optimal value of K for a given system), but is otherwise objective.

However, in the most common scenario, τ is not fully quantified, because a set of background conditions and/or nested descriptions of concepts and relations is omitted. This entails that the K values estimated about a system will typically be *upper bounds* of the actual values. Formally, if $K(Y; X, \tau_i, \tau_j)$ is the exact value of K when both additive components τ_i, τ_j of the τ are measured, but we only calculate one of the, $K(Y; X, \tau_i)$, then

$$K = K(Y; X, \tau_i, \tau_j) < K(Y; X, \tau_i) \quad (14)$$

and the calculated value of K is an upper boundary of the true value. Similarly, if $K(Y; X, \tau_k, \tau_j)$ is an alternative that we are comparing, it is clear that a derived measure like the difference ΔK

$$\Delta K = K(Y; X, \tau_k, \tau_j) - K(Y; X, \tau_i, \tau_j) < K(Y; X, \tau_k) - K(Y; X, \tau_i) \quad (15)$$

and the difference in K values that we calculate is an upper boundary of the true difference.

Any conclusion drawn from estimates of K will remain valid in relative terms, unless $-\log p(\tau_j) = \infty$ (in which case $K = 0$), but this case is excluded by the first postulate in of K theory (information is finite), which is really a postulate about the inevitable finiteness of any description.

Relation to Kolmogorov complexity and physical entropy

The general and relative validity of K makes it applicable to exact but incomputable quantities like Kolmogorov complexity, or exactly measurable physical quantities, like Boltzman's entropy. In the former case, since the Kolmogorov complexity is the most succinct description of an object, then $D(\tau) \geq C(\tau)$ where $C()$ is the Kolmogorov complexity of the system, giving

$$K = \frac{D(\tau) - D(\tau')}{D(\tau)} \leq \frac{C(\tau) - C(\tau')}{C(\tau)} \quad (16)$$

showing that the K values we calculated are lower or equal to those obtained when $C()$ is known.

Similarly, instead of representing Shannon's entropy, the terms in the K function could represent physical entropy, in particular either Boltzman's entropy measure, $S = k_B \log w$ where W is the number of distinguishable microstates and k_B is Boltzman's constant, or the more general Gibb's entropy $k_B H(W)$, where W is the random variable describing the ensemble at some coarse-grained level. In either case, the constant multiplication term cancels out, such that

$$\frac{k_B n_Y H(Y) - k_B n_Y H(Y|X, \tau)}{k_B n_Y H(Y) + k_B n_X H(X) - k_B \log p(\tau)} = \frac{n_Y H(Y) - n_Y H(Y|X, \tau)}{n_Y H(Y) + n_X H(X) - \log p(\tau)} \quad (17)$$

which suggests that the substantive interpretation of K extends to physical entropy.

3.1.1 A recursive interpretation of K

Equation 13 generalizes K to a measure of structure in a system. It is also possible to understand the meaning of the number by noticing that

$$K = \frac{n_Y H(Y) - (n_Y H(Y|X, \tau) - cK)}{n_Y H(Y) + n_X H(X) + \log p(\tau) + c} \quad (18)$$

in other words, if we expand the explanans by c bits, we need to save $c \times K$ bits of residual ignorance to keep the K value stable.

Alternatively, we need to expand the explanandum by $cK/(1 - K)$ bits:

$$K = \frac{n_Y H(Y) + c \frac{K}{1-K} - n_Y H(Y|X, \tau)}{n_Y H(Y) + c \frac{K}{1-K} + n_X H(X) + \log p(\tau) + c} \quad (19)$$

Thus, the number K is a measure *the fractional costs of expanding the explanans*. Put another way, K is a measure of what proportion of our ignorance needs to be reduced to justify an increase in complexity of our theory/methodology. The lower the K , the lower the costs of doing so. K is a measure of the potential for knowledge growth within a system.

3.1.2 Knowledge as lossy information compression

The recursive property of K also allows to prove that K grows by lossy abstraction of patterns from data - that is, by describing sets of phenomena at higher levels and discarding all lower-level information (proof in 6.4).

This proof validates the second postulate of K theory, which states that knowledge is information compression. It also further confirms the first postulate, i.e. that information is finite, by showing that the compression attained is lossy (see also another argument in [1], section 2.3.6).

In other words, we can show mathematically why, in order to “know” anything, we have to “condition out” the greatest part of a potentially infinite τ , limiting our understanding to a small part of it. K is the measure of how well we are doing with a given abstraction.

3.1.3 The total K of a τ

Combined with the recursive quantification of τ , the recursive interpretation of K allows us to calculate the total knowledge contained in a τ of any complexity. Let $\bar{\tau} = \tau_1 \supset \tau_2 \supset \tau_3 \dots$ be a nested τ of arbitrary depth and complexity, and let $\underline{\tau} = \tau \supset \dots \supset \tau_2 \supset \tau_1$ be the de-nested version (that is, the union of all its nested components), then (see 6.5 for more details) the total K is given by a continued, recursive fraction, such that

$$K(\bar{\tau}) = \dots \frac{\Delta\tau_4 + \sum_1^3 D(\tau) \frac{\Delta\tau_3 + \sum_1^2 D(\tau) \frac{\Delta\tau_2 + D(\tau_1) \frac{\Delta\tau_1}{D(\tau_1)}}{\sum_1^2 D(\tau_i)}}{\sum_1^4 D(\tau_i)} = \frac{\sum \Delta\tau_i}{\sum D(\tau_i)} = \frac{\sum \Delta\tau}{D(\bar{\tau})} \quad (20)$$

which shows that the total K of any multilevel τ results from the sum, at the numerator, of the number of bits yielded by all statements (all sub- τ 's) contained in the nested τ and, at the denominator the sum of the descriptions of the individual nested τ 's, which by definition is the description of the total τ .

Further, unless the τ is completely disjoined,

$$K(\bar{\tau}) \geq \frac{\sum \Delta\tau}{D(\underline{\tau})} = K(\underline{\tau}) \quad (21)$$

showing that the K calculated from the nested version of the τ is an upper bound of the one calculated with the de-nested version. The latter, if calculated completely, would constitute the true, exact calculation of the total knowledge expressed by a given τ .

3.2 How to measure the K of anything

In light of equation 13, we see that in order to estimate the K of a knowledge claim or system we typically have two steps to consider:

1. Describe the system in terms of a τ that is adequate (i.e. sufficiently deep) for one's objectives.
2. Identify which elements of it constitute the explanandum, thereby defining the knowledge claim being analysed.

Any τ can be described progressively, following any combination of a “top down” and “bottom-up” approach. In the former, we start with the simplest possible τ . This is:

because no structure can be simpler without entailing $K = 0$. From this unambiguous top-most level, we proceed by expanding the terms, adding elements to the τ and/or nested descriptions.

The “bottom-up” approach starts by describing elements of the system in any order, via small, separate graphs, and then structures them by some combination of nesting them in one another, joining them together in more complex τ , or by cumulating them in strings.

Decisions about what constitutes the explanandum depend on the objective of the analysis and what is or can be known about the system. For example, let τ be a chain of three elements and two relations:

A generic τ' might have both z and y as explananda, which we can represent as:

leading to

$$K = \frac{D(y) + D(z) - D(y|z, r) - D(z|x, r)}{D(y) + D(z) + D(x) + D(r) + 7.75} \quad (22)$$

However, it could be the case that z is not defined and/or $z|x, r$ is not quantified, such that $D(z|x, r) = D(z)$ and

$$K = \frac{D(y) - D(y|z, x, r)}{D(y) + D(z) + D(x) + D(r) + 7.75} \quad (23)$$

which is still a valid K , which measures the knowledge about a single explanandum y using a theory comprising multiple entities. Naturally, such a theory could be simplified by nesting and reducing the input to a single variable x , but this again is a methodological and descriptive choice to be made in the context of an analysis.

In another possible case, τ' is what remains unexplained about a partial description of τ , or a generating process for it. In this case the explanandum is constituted by the whole description length of τ itself.

Practical examples for these and other cases will be give in the Results section.

4 Results - examples of applications

This section will offer practical examples of how the methods described above can be applied to different types of knowledge.

Examples are presented in a similar order to the examples given in [1], that is: logic 4.1, empirical quantitative knowledge 4.2; meta-scientific knowledge 4.3, empirical qualitative knowledge 4.4, and mixed quantitative and qualitative knowledge 4.5.

This order or presentation reflects the increasing complexity of subject matter, theories and methodologies, and the decreasing order of theoretical and methodological consensus and codification. As such, this order reflects what is commonly intended as the “hardness” of different sciences

(see [1] section 3.8 and references within). This order also corresponds to increasing difficulties encountered in describing the τ , which, in more complex and less codified fields, requires developing specific coding schemes and logics that require a dedicated analysis to conduct. Therefore, the earlier examples in logic and mathematics will be more complete and specific than later ones, the complete development of which is left to future work.

4.1 Abstract (logical-deductive) knowledge

Knowledge stemming from any kind of tautology is characterised by having no empirical error, i.e. $H(Y|X, \tau) = 0$. A preliminary K value for logic was calculated by [1] by assuming $H(Y) = 1$ bit, which corresponds to the uncertainty about a true/false proposition, obtaining the general formula:

$$K = \frac{1}{1 + \frac{-\log p(\tau)}{n}} \quad (24)$$

In light of the methods developed above, we can now ascribe to any logical statement z a value $D(z)$ that reflects the information value of the statement, which may correspond to how novel and interesting the information is, often defined as its “surprisal”.

Depending on how z is defined and described, $D(z)$ may be estimated as the subjective surprisal (for example using Bayesian techniques of elicitation, see [17]) or its descriptive complexity, calculated from 5. In either case, it will often be the case that $D(\neg z) \approx 0$, since the surprisal of $\neg z$ will be low by definition, whereas the complexity of $\neg z$ is given by $-\log(1 - p(\tau_z))$ which will be very small unless z is extremely simple.

4.1.1 Example 1.1: Modus ponens

The modus ponens is the most elementary logical proposition, of the form “if s , then p ” (in symbols, $\forall s, p$), where s is the condition and p the conclusion.

Proposition, condition and conclusion correspond, respectively, to the τ , the input and the explanandum. The τ for the implication proposition is:

which has $D \geq 3$

Plugging in the K function the values for single events we get the amount of knowledge:

$$K(p; \tau, s) = \frac{D(p)}{D(p) + D(s) + D(\Rightarrow) + 3} < \frac{1}{1 + \frac{D(s)+3}{D(p)}} \quad (25)$$

which increases with the surprisal of p and decreases with the surprisal of s .

Note that the implication says nothing about the truth value of p when s is not true ($\neg s$), something that we can express as $D(p|\neg s, \tau) = D(p)$ (as opposed to $D(p|s, \tau) = 0$).

If we evaluate across a cumulation of n conditions $s_1 \oplus s_2 \dots s_n$, indicating with S the random variable that associates a probability distribution π to s being either true or false, then we get a measure of expected knowledge across multiple (potentially varying) conditions as:

$$\begin{aligned} K(p^n; \tau, S^n) &= \frac{nD(p) - nE[D(p|S, \tau)]}{nD(p) + nE[D(s)] + D(\Rightarrow) + 3} \\ &< \frac{D(p)(1 - \pi(\neg s))}{D(p) + H(S) + D(s) + D(\neg s) + \frac{3}{n}} < \frac{\pi(s)}{1 + \frac{H(S)+D(s)+3/n}{D(p)}} \end{aligned} \quad (26)$$

which, since $H(S) \leq 1$, says that the knowledge obtained is roughly proportional to the probability that s is true (s occurs rather than $\neg s$).

If instead of the implication we have a biconditional $\Longleftrightarrow$ such that $D(p|\neg s, \tau) = D(p|s, \tau) = 0$, then both s, p are random variable, S, P , and the general formula is

$$K(P^n; \tau, S^n) < \frac{H(P) + D(p)}{H(P) + H(S) + D(p) + D(s) + \frac{D(\Leftrightarrow) + 3}{n}} < \frac{1}{1 + \frac{H(S) + D(s) + 3/n}{H(P) + D(p)}} \quad (27)$$

which generalizes 25. If no description or surprisal for p, s is given, then K reduces to $K(P^n; \tau, S^n) < 1/(2 + 3/n)$ which depends solely on the relatively complexity of the τ .

4.1.2 Example 1.2: Syllogism

Syllogisms are the basis of reasoning and can be construed as expansions of the modus ponens, in which the condition s and the conclusion p acquire more complex structures.

A classic example studied by most high-school students is Mill's syllogism: "Socrates is a man, all men are mortal, therefore Socrates is mortal". This is easily translated into the graphical language proposed. Mill's syllogism, for example, may be written out in nested form as:

This nested representation is compact but not economical, because it isolates and multiplies unnecessarily concepts that are identical (for example, a nested description of Ma is contained both in A and B , so the total D of this nested representation would include $2 \times D(Ma)$). In order to quantify more accurately the minimum complexity of the syllogism, therefore, we de-nest the graph obtaining:

which completely describes the syllogism, less the description of the lower-level concepts "Socrates", "Man" and "Mortal". The size of the syllogism's τ is $D(\tau) = 49.28 + D(So) + D(Ma) + D(Mo) + D(A) + D(B) + D(s) + D(p) + D(is) + D(\wedge) + D(\Rightarrow)$, which simplifies to $49.28 + D(So) + D(Ma) + D(Mo) + D(is) + D(\wedge)$ because A, B, s, p merely represent the higher-level concepts described within the τ itself.

This τ expresses not just the inference $s \rightarrow p$, but also the meaning that the inference has in terms of the underlying concepts (here So, Ma, Mo). Note also how all of the concepts and relations on the left-side of the concept p and including p could be nested into the relation "is" that connects "Socrates" to "Mortal", which correctly reflects the fact that the longer reasoning entailed by the syllogism is a lower-level, deeper description of the concept that Socrates is mortal.

The knowledge acquired via the syllogism is the knowledge contained in the conclusion p that "Socrates is Mortal" which, indicating its τ as τ_p is quantified as:

$$K(Mo; So, \tau_p) = \frac{1}{1 + \frac{D(So) + D(is) + 3}{D(Mo)}} \quad (28)$$

where the values $D(Mo), D(So)$ are the surprisals of the concepts "mortal" and "Socrates". Therefore, to quantify the *upper boundary* of the amount of knowledge obtained via the syllogism, we de-nest τ_p , replacing it with the full τ of the syllogism:

$$K(Mo; So, \tau) = \frac{1}{1 + \frac{D(So) + D(is) + D(Ma) + 49.28}{D(Mo)}} \geq \frac{1}{1 + \frac{D(So) + 49.28}{D(Mo)}} \quad (29)$$

where in the last step we arbitrarily decide to omit the relation is and the concept Man (but might have included it, depending on its relevant to an analysis).

The values $D(So), D(Mo)$ can be determined in at least two possible ways:

1. Uninformed: where we give each of them a standard entropy value, typically that of one bit. The resulting K might be used to compare, for example, the potential knowledge value of two logical constructs, ignoring their contents.
2. Structural: by quantifying the description lengths of these concepts. This is in principle an objective approach to estimating surprisal since, all else equal, concepts with longer descriptions are by definition rarer, more contingent, more surprising, less probable etc.
3. Subjective: by eliciting how surprising the explanandum is given the explanans, e.g. how much more surprising is it to know that Socrates is mortal relative to anything else being mortal?. Techniques of elicitation developed in the context of qualitative Bayesian analyses might be used [17]. In absence of information, or interest in such information, a standardized surprisal value of 1 bit can be given to both, as in the original treatment in [1].

4.1.3 Example 1.3: Logical proof and creation of new knowledge

The set of all statements that have been proven to be true in a particular logical language is defined by logicians as the “theory” of that language. New true statements are derived via proofs, which consist in chains of other true statements. Just as we derived, in previous sections, the K of a statement and the K of a syllogism as an expansion of a statement, we can calculate the K value of a proof which, all else equal, will be inversely related to the length and complexity of the proof itself (see [1] section 3.1).

To make a simple example, we shall take the language of boolean logic and the proof that $(y \wedge \neg x) \vee (y \wedge x) \iff y$. One proof of this statement consists in the following sequence of statements: $(y \wedge \neg x) \vee (y \wedge x) \iff y \wedge (\neg x \vee x)$; $y \wedge (\neg x \vee x) \iff y \wedge T$; $y \wedge T \iff y$. For the proof to occur at all, each of the statements involved in this proof must be already available to the proving system. If we write down each of the statements and then join them we obtain the τ of the minimum knowledge possessed by a system capable of producing the proof, which is

and it has $D > 118.58$. This structure represents the amount of logical “theory” encoded by the system, and it is an information processing device at multiple levels. At an elementary level - which is the one already seen in previous examples, it is a τ that, for any combination of values (true/false) of x, y determines the truth value of the remaining 7 higher-level concepts (except for the concept T , which is fixed).

This structure, processes information at a higher level, too, because we could nest in each higher-level concept their corresponding lower-level description. This means that this τ accepts as inputs not just x, y truth/false values, but also any τ corresponding to the description of a higher-level concept (e.g. $x \wedge y = TRUE$). Information processing occurs at this higher level via the equivalence relation $\leftrightarrow$, which connects a particular sub- τ to another sub- τ .

Furthermore, if we allow information to flow backwards, from the head of the hyper-edge to its tail, this structure also encodes the “meaning” of each higher-level concept. For example, the meaning of the symbol $y \wedge x$ is the sub-graph connecting y, x via $\wedge$, of which the symbol is the concept (see section 6.3.4).

How much knowledge does this system possess? The total K is obtained from the summation of all the statements that the system can process. This is equivalent to the total number of relations, r . Each statement will have one or more inputs, taken from the total possible inputs corresponding to the total set of concepts encoded in it. This is the total number vertices, d . Setting $n = 1$, the maximum total K for the system is:

$$K(\text{total } \tau) = \frac{r}{r + d + D(\tau)} = \frac{11}{11 + 10 + 118.53 + \sum_{r,d} D(i)} \leq 0.079 \quad (30)$$

where $D(i)$ are the lower-level descriptions, if present.

Logical proofs are structures in which the theory (the complete τ above) relates τ inputs to τ outputs. The proof in question, for example, can be described as the linear sequence of steps:

which has $D = 20.53$ and in which τ is the structure described above, and $\tau_0, \tau_1, \dots$ are the steps in the process, described as conditionings of the τ to specific sub-components. Specifically the steps are:

- τ_0 = input the proposition $(y \wedge \neg x) \vee (y \wedge x)$:

- τ_1 : τ_0 after application of the distribution property: $(y \wedge \neg x) \vee (y \wedge x) = y \vee (\neg x \wedge x)$:

- τ_2 : first step in order to apply the complement property, in which we nest the relevant subcomponent:

- τ_3 : τ_2 after application of the complement property: $y \vee (\neg x \wedge x) = y \vee T$:

- τ_4 : τ_3 after application of final equivalence $y \vee T = y$

Therefore, the total complexity of the proof results from the sum of the higher-level structure and the description of each passage described above:

$$D(\text{proof}) > 20.53 + 118.58 + 41.96 + 30.735 + 15.56 + 10.58 = 237.945 \quad (31)$$

The K provided by the proof is obtained from the general formula 27, in which we assume $H(S) = H(P) = 1$ and let s be the input τ_0 and p be the statement that is proven $(y \wedge \neg x) \vee (y \wedge x) \iff y$. We get

$$K(\text{proof of } p) < \frac{1}{1 + \frac{1 + D((y \wedge \neg x) \vee (y \wedge x)) + 237.945/n}{1 + D((y \wedge \neg x) \vee (y \wedge x) \iff y)}} \quad (32)$$

As discussed in section 4.1.1, the surprisals of $D((y \wedge \neg x) \vee (y \wedge x))$ and $D((y \wedge \neg x) \vee (y \wedge x) \iff y)$ can be given a standard value or can be estimated either structurally or structurally.

This example also offers a first instance of the quantification of learning. Once the statement is proven true, the system can encode the new connection. By doing so, it expands its knowledge with a new statement that can be used in further proofs. If so, then the new theory τ will be:

which has $D > 126.53$.

Has the K of the system is increased? Even without calculating it, we can prove that it has, by applying the recursive property of K . Since we have added a new relation, the total number of new statements has expanded by 1 and we know from equation 19 that K will increase if $1 > 0.079(126.53 - 118.58) = 0.62$, which is true. Hence, as the theory expands to include new true statements, the total knowledge quantified by K increases, as would be expected.

4.1.4 Example 1.4: Mathematical knowledge

Just as mathematics is an extension of logic, the τ of mathematical knowledge is an extension of that used in logic, in which however concepts and relations have more complex descriptions.

A natural number n is a τ described by a set containing a number n of elements. Its description length is therefore $D(n) = \log n$, which represents the minimum information required to distinguish each element in the set (see 6.2). This quantity is also the minimum length of the binary encoding of that number. Encodings of rational numbers, real numbers, floating points etc. are extensions of this logic, equal to imagining larger sets of elements.

Therefore, any sequence of mathematical operations can be represented as a nested τ whose concept are ultimately encodings of sets (numbers) and in which relations are operations that transform them. For example, let our problem be to find the value of x given that $x + 3 = y, z = 7 - 2, y = z$. The sequence of mathematical operations that give the solution might be described graphically as:

and the sequence of steps that solves the equation is described by:

where the τ is:

with $D(\tau) = 74.67$, for a total complexity $D = 20.53 + D(\tau) + \sum D(\tau_i) + D(7) + D(2) + D(3) + D(5) = 95.2 + 35.33 + 2 + 6.33 + 7.33 + 7.33 + \log(7) + \log(2) + \log(5) + \log(3) + \log(2) = 163.23$.

The total K expressed by this τ can be calculated as for logic:

$$K(\text{total } \tau) = \frac{6}{6 + 6 + 163.23 + \sum D(i)} < \frac{6}{6 + 6 + 163.23} = 0.034 \quad (33)$$

where $D(i)$ is the encoding of the numbers and operations involved.

4.2 Empirical, quantitative knowledge

As shown in [1], the fundamental difference between empirical knowledge and abstract, logical-deductive knowledge is that the former is always characterised by non-zero residual ignorance:

$$\text{empirical knowledge} \iff H(Y|X, \tau) > 0 \quad (34)$$

which entails that there is noise in the system and/or that the validity of a knowledge claim cannot be deduced or proved logically, and must be verified against measurements. It follows that other factors, whilst not exclusive to empirical knowledge, will affect the K of empirical knowledge more significantly:

- **Conditioning.** Empirical knowledge claims are, to a lesser or greater degree, only applicable to subsets of all possible phenomena. The more the universe is “narrowed down” to a subset of phenomena to explain (explananda), the more we are likely to identify strong patterns within it, increasing the K at the cost of generality.
- **Explananda expansion:** the same input and τ may be applicable to multiple different phenomena. Applying the same τ to multiple phenomena entails that the τ is actually slightly more complex, but the expansion is compensated by the additive growth of the bits of phenomena being compressed.
- **Theory maximization.** Since noise is bound to remain around any knowledge claim, multiple possible “theories” can be produced to explain, incompletely, a set of observations. A balance typically needs to be found between the complexity of a theory and its explanatory power.
- **Methodology optimization.** Just as we narrow the range of possible empirical explananda, we narrow down the set of possible choices and actions that may be performed in order to obtain the pattern. Again, an optimal balance may be found between the level of detail and specificity of a methodology and its generalisability.

Empirical scientific research can be understood as a process of maximizing the K by optimizing the factors above, i.e. by finding the minimum set of conditions, for the broadest range of phenomena, to which the simplest possible theory applies with the simplest number of methodological steps.

4.2.1 Example 2.1: Predicting solar eclipses, with conditioning

One of the first examples of empirical knowledge given in [1] had calculated the knowledge expressed in predicting a solar eclipse on Earth using the JPL DE405 solar system ephemeris. The explanandum and residual ignorance were derived from the total surface of the Earth and its the residual error, both in square kilometres, as $s_{Earth} = 510072000$ and $e_{Earth} = 4$, respectively; and the input, was approximated to the 6910 bits needed to encode all the circa 416 physical constants, parameters and initial conditions required, assuming an accuracy of 5 digits on average.

However, there was no simple way to estimate the τ itself, comprised of 33 equations of motion and 21 computations. We can now estimate the complexity of the JPL DE405 calculations. An exact calculation would require the drawing of the τ of all equations, which would be too laborious for the scope of this essay. However, we can estimate the *minimum* complexity of this system as follows.

We shall assume that each of the 416 parameters is involved in exactly one operation, and that each operation requires two variables (that is, each relation has arity 3). The minimum structure of this system will be that of a tree, in which each result of a binary operation becomes one of the variables in the next operation up, until a result (the explanandum) is given. This is the τ we shall assume, and it is likely to be a conservative estimate of the real structure, which must include operations of varying arity and which probably makes multiple uses of the same inputs and values.

It is easy to show that the total number of vertices of such a tree is $|V| = 416 \times 2 - 1 = 831$, and the number of relations is $r = 416 - 1 = 415$. Let us approximate to 416 and assume that only four fundamental operations are involved, with equal frequency, such that the multiplicities of each

relation is exactly $m_{ei} = 416/4 = 104$, with symmetry in two of them. Plugging into equation 5, we get $D(\tau) = 16557$ bits circa.

Thus, the *upper bound* of the K for locating a total solar eclipse on Earth at a given time (assuming $n_1 = 1$) is:

$$K(E; JPL) = \frac{\log(510072000) - \log(4)}{\log(510072000) + \frac{6910+16557}{n}} \quad (35)$$

which is < 0.001 for $n = 1$. However, as pointed out in the original analysis, $n \gg 6910 + 16557$ because the system is deemed virtually perfect and it has/will be used millions of times, making the costs of the explanans negligible.

However, the original analysis ignored the conditioning component, implicitly assuming that we are only interested in finding eclipses on Earth. What if our objective is to predict eclipses on all planets in the solar system, using JPL?

For the purposes of illustration, let's assume that the JPL system is completely incapable of making predictions for any planet other than our own (this is not true, but we will assume it to simplify the example). In the most naive condition, we do not distinguish between the planets and treat all of the surface of the eight planets as one single surface on which to locate eclipses. Indicating with s_i the logarithm of the surface of planet i , with e_i the logarithm of the prediction error, and indicating with $J = (6910 + 16557)/n$ the explanans (which from now on is a constant) we have

$$K(\sum s_i; JPL) = \frac{\log(\sum s_i) - \log(\sum e_i)}{\log(\sum s_i) + J} \approx \frac{\log s_{Earth} - \log e_{Earth}}{\log(\sum s_i) + J} \quad (36)$$

Let's expand slightly the τ and posit instead that we do distinguish between the eight planets prior to implementing the JPL system on any particular planet. Instead of considering the sum of the surfaces, we consider a cumulation of them (see 6.3.1), which requires $H(W) \leq \log 8$ bits plus the individual value for the surface of each planet. This quantification is equivalent to imagining us drawing one planet at random, and applying the JPL to see if we can predict an eclipse on it. Applying equation 101 to this explanandum, we get

$$K(s_i; JPL) = \frac{H(W) + \sum \log s_i - (H(W) + \sum \log e_i)}{H(W) + \sum \log s_i + J} = \frac{\log s_{Earth} - \log e_{Earth}}{H(W) + \sum \log s_i + J} \quad (37)$$

which is smaller than the value in equation 36, due to $\sum \log s_i > \log(\sum s_i)$ and the additional information term $H(W) > 0$ needed to distinguish amongst the eight planets. However, since we have the information to distinguish the planets, we also have the information to selectively apply the τ only to the Earth.

We can picture this selection as a filter, which for any one of the n possible requests of prediction on a planet will only accept the fraction m that are about the Earth. Methodologically, this is represented by a conditioning of the input (see 6.3.3). The result is a conditioned τ

$$K(s_i; JPL) = \frac{\log s_{Earth} - \log e_{earth}}{\log \frac{n}{m} + \log s_{Earth} + J} \quad (38)$$

which is the same as equation 35 but now with an explicit conditioning factor. If we were to compare the K of JPL to another system applicable across the planetary system, we would use equation 37. When comparing two systems on different planets, we would use 38. However, when comparing systems conditioned on Earth, the conditioning factor is a constant that can be ignored (see 6.3.3), giving back equation 35.

4.2.2 Example 2.2: Statistical model selection

The central problem of statistical model selection is to identify the model that, amongst a set of candidate alternatives, best accounts for the data. The model might be sought to merely explain the data, but more commonly it is taken as an approximate representation of the process that generates the data, i.e. a theory about the phenomenon being measured.

Statistical models are generally subject to a complexity/bias tradeoff: making them more complex reduces the risk of missing important information in the data, but it does so at the risk of over-fitting a particular data set, reducing the model’s ability to fit future data. To identify the optimal model, metrics of model selection typically penalize a measure of model fit (e.g. the negative log-likelihood) by a measure of the complexity of the model.

However, most criteria of model selection equate a model’s complexity with the the number of parameters estimated by the model (or some function of it). The simplest and most popular example is the Akaike Information Criterion (AIC, [18, 19]), which for model m is calculated as

$$AIC_m = 2L_m + 2k_m \quad (39)$$

where $L = -\sum \log p(y|m)$ is the the negative of the log-likelihood (i.e. the log of the joint probability of the data given the model) and k_m is the total number of parameters estimated in the model. Following this criterion, a model i is preferable to a model j , if

$$AIC_i < AIC_j \iff L_j - L_i > k_i - k_j \quad (40)$$

Although derived as a rather sophisticated estimator of the distance from the “true” (unknown) model, the AIC and similar information approaches does not quite capture all of the factors by which a model’s fit should be penalized. Two factors that have proven particularly challenging are:

1. Model’s functional form. This extends the concept of model complexity by equating it with the model’s ability to fit a broader range of data sets. Models with the same number of parameters might have different degrees of “pliability” and therefore adaptability to data, and information criteria like the AIC are blind to this property. Methods based on the Minimum Description Length principle expand the criteria of model selection by including, in their most sophisticated version, a term fthat quantifies a model’s complexity in terms of the breadth of data sets that it can accommodate [9, 10]. Quantifying the functional form, however, is hard to do theoretically and computation-intensive empirically [20].
2. Bias due to researcher’s “degrees of freedom”. Whenever arbitrary choices are made in an analysis, in a way that is not random but rather directed at improving the model’s fit, the ability of the model to fit new data is over-estimated, resulting in upwardly biased estimates of effects and poor generalizability. In the context of statistical analyses, this is a known problem of over-fitting, and it becomes even more pernicious and hard to tackle when it is extended to the entirety of a research process. Researchers may take innumerable arbitrary decisions both before and after running a statistical analysis, leading to two separate sources of bias:
 - Ex-ante: decisions are made *prior* to viewing any results, and are determined by the analysts’ own knowledge of what types of approaches will best apply to that particular data set. These sources of ex-ante bias have been variously and pictoresquely indicated in the literature as “forking paths” or “multiverse” problems [21, 24]).
 - Post-hoc: decisions change the analysis and revise previous strategies based on how well the model fits the data. Whilst in principle this is a legitimate way to “explore” the data, it becomes problematic in the context of hypothesis testing or confirmatory research, because it inflates the risk of false positives. When these post-hoc decisions are undisclosed, they are often collectively defined as “P-hacking”, because they are typically guided by the researcher’s wish to reject the null hypothesis with low P-values [22]).

Using τ as a measure of complexity and K as a metric of model selection might offer an alternative, which can incorporate both issues listed above.

K as a metric of model selection

Given a data set of n observations of a response variable Y and explanatory variables $X_1, X_2, \dots$, the terms of K for a statistical model are defined as follows:

- τ : Technically describes all other components of the model, i.e. the model's structure, operations, and variables (explanandum, input and parameters). For example, the model $y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2$ translates into a τ

with $D = 106.12$.

- Explanandum: this will be the log-likelihood of the response variable given the null (or intercept-only) model m_0 . This follows from the fact that

$$nH(Y|m_0) \sim \sum_n \log \frac{1}{p(y_i|m_0)} \equiv L_0 \quad (41)$$

where we may chose the basis of the logarithm as 2 or e .

- Residual ignorance: follows from above as the log-likelihood given a candidate model m_1 :

$$nH(Y|m_i) \sim \sum_n \log \frac{1}{p(y_i|m_1)} \equiv L_i \quad (42)$$

(a more accurate estimation would be obtained by using deviance statistics instead, although doing so would simply introduce constants at the numerator, which do not affect conclusion.

- Input: is the distribution of one or more independent explanatory variables

$$n_X H(X) = n_X \sum H(X_i) \quad (43)$$

where the entropy is calculated using the same base as the explanandum.

- The term n_X represents the number of measurements needed for each explanatory variable. It is usually assumed that $n_X = n_Y = n$, and that each variable needs to be measured on each new individual, but that is not necessarily the case, because some measurements could be re-used in a variety of different contexts, depending on their stability. Variables that are constant and re-usable will have a smaller n_X and therefore a lower impact on the total size of τ . In practice, the two most common and useful values are $n_X = n_Y = n$ and $n_X = 1$. This modulation of n bears some analogy with some formulations of AIC [23], although the logic and theory behind it is different.
- Model parameters: are in effect just another component of the input $H(X)$, which, once estimated, is fixed for all n observations. Hence for this component it is always $n_X = 1$, and the corresponding values are a function of the number of parameters times a constant C_β that quantifies the encoding of each parameter.

The resulting K is

$$K = \frac{L_0 - L_i}{L_0 + n_{X1}H(X_1) + n_{X2}H(X_2) + 106.12 + 4C_\beta} \quad (44)$$

where C_β may follow different methods and be more or less precise (see [9]) but, once chosen, will be constant.

When comparing any two models on the same data but different numbers k of parameters and different explanatory variables will yield preference of model i over j when $K_i > K_j$, which re-arranging gives

$$L_j - L_i > K_j \frac{D(\tau_j) - D(\tau_i)}{n} \quad (45)$$

where L is the negative log-likelihood and

$$D(\tau_i) - D(\tau_j) = \log \frac{p(\tau_j)}{p(\tau_i)} + n_{X_i} [\sum H(X_i) - \sum H(X_j)] + (k_i - k_j)C_\beta \quad (46)$$

is the total difference in description length between the two τ .

Note that the difference in the number of parameters ($k_i - k_j$) is now just one of the penalties for model complexity, in addition to those resulting from the difference in structural complexity and in the entropy of the variables. The relative weight that these different penalties have on model selection depends on the values of n_X .

Further, note that in 45 the overall penalty is not only inversely proportional to sample size n_Y , in line with ordinary log-likelihood-based methods, but also inversely proportional to the K value of one of the terms of comparison. In other words, the “bar” for increasing model complexity is raised higher and higher as the model fit improves.

$D(\tau)$ as a measure of functional form

Equation 45 encompasses the structural complexity of a model, captured within $D(\tau)$, that may help approximate the “pliability” of a model. To take a very simple example used in the literature (e.g. [20]), two models like $y = a + b$ and $y = e^{ab}$ have the same number of parameters and would therefore be considered of equal complexity by ordinary information metrics like AIC, and yet the former function is less pliable than the latter.

Such ambiguity would not subsist if the complexity of the two models was measured as a τ , because the structure of the latter, by requiring two operations, is clearly larger than the former. Graphically, we may represent the two functions as:

has $D > 18.83$, whereas a minimal representation of the second function

is

has $D > 30.74$, showing unambiguously that the latter is

more complex.

Preliminary results suggests that D does indeed correlate with empirical measures of a model’s adaptability, but a rigorous analysis is left to future work.

K as a bias-adjusted metric of model selection

The biases resulting from researcher’s degrees of freedom can be quantified in terms of the number of bits entailed by the choices that researchers made. For example, if a researcher faces the decision, after seeing an analysis, to retain or drop a subset of m data points in its data set, and she does so completely arbitrarily, then that choices entails $m \times 1 = m$ bits of information, where the 1 is the one bit entailed by a completely arbitrary binary choice. A more extended but still preliminary analysis of how ex-ante and post-hoc biases can be quantified is given in [1].

One of the unique advantages of K is that, unlike any other metric of model selection or measure of bias, it is able to penalize model selection for ex-ante and post-hoc biases separately. If we let B quantify the cost of post-hoc choices and A quantify that of ex-ante choices, then equation 45 is expanded to:

$$L_j - L_i > K_j \frac{D(\tau_j) - D(\tau_i) + A_j - A_i}{n} + (B_j - B_i) \quad (47)$$

which, all else equal, penalizes the result derived from a larger set of ex-ante and post-hoc arbitrary choices, evaluated separately. Unlike the cost of ex-ante choices, the cost of post-hoc choices is not

divided by n or multiplied by K_j , suggesting that P-hacking should be a greater concern than the forking-paths problem.

Deciding which ex-ante methodological decisions are arbitrary and need correcting is a non-trivial matter [24]) and pre-registering studies to transparently document any post-hoc re-specification is not a common nor entirely uncontroversial practice [25]. However, to any extent that accurate information on the choices made by researchers is available, the K of their results can be penalized by the extra degrees of freedom involved, which may help derive a more accurate estimate of the “strength” and generalisability of a result.

4.3 Metascientific knowledge

One of the primary objectives of K theory is to offer a theoretical and methodological paradigm for Metascience, which is the field that studies the processes and practices of scientific research in order to identify areas of improvement.

Metascientific applications of K theory have already been presented in [1]. However, a complete treatment required a consistent quantification of a study’s theory and methodology, which was not attainable at the time. This section will first illustrate how a study’s theory and methodology can be translated into a quantifiable τ , and it will then offer an example of how this approach can be used to measure, study and forecast a study’s replicability - which is one of key concerns of metascientists.

4.3.1 Describing a study’s theory

Previous examples have already offered instances of τ ’s that represent theories. Whilst there may be different philosophical approaches to characterizing the nature and role of theory in research [26, 27], all theories share the fundamental feature of being a semantic structure that defines what objects are relevant (often giving rise to a specific ontology), how they relate to each other, and thus defines all true statements about those objects and relations.

Therefore, virtually any τ underlying a knowledge claim may be thought as representing a “theory”. The τ defining the relation between a specific explanans and explanandum is a circumscribed theory about that system. By extension, the conceptually broader kind of theory that characterises all the assumptions and implications of a study, and/or the even more general and abstract paradigms that define an entire research program, field or discipline (e.g. [28]), ultimately must consists in larger, more complex τ ’s.

In practice, it may be very difficult or impossible to fully describe the theory underlying a field, but the process is analogous to that shown in previous sections for formal logic statements: we translate each true statement composing a theory into a τ , and then quantify the theory as a whole as the τ that results from denesting (that is, uniting) all these τ . Just like the τ used in the proofs above, the theory constituting a background to a study is a larger structure from which all statements relevant to that study are derived (e.g. the passages in a proof, but also the predictions made by a particular study, the description of its explanans and explanandum, the conditions in which these apply etc.).

A theory τ , however, is likely to have distinctive features, including:

- **Predictive:** implicitly or explicitly, a theory states what the relation between entities is expected to be in future observations. The relation between explanatory and predictive knowledge was examined in [1]. Here it is worth remarking that, in the context of predictions, the residual ignorance is enlarged by a measure of divergence between the observed and the predicted distribution. In the case of random variables, this divergence may be given by the cross-entropy between the observed and predicted outcomes, $y, \hat{y}$:

$$H(Y\hat{Y}|X, \tau) = \sum p(y|x, \tau) \log \frac{1}{p\hat{y}} = H(Y|X, \tau) + B \quad (48)$$

where $B = \mathcal{D}(Y|X\tau||\hat{Y}|X\tau)$ is the Kullback-Leibler divergence between observed and predicted distributions (see [29, 1]). In light of equation 13, $K < 0$ entails $D(\tau') > D(\tau)$, i.e.

$$D(\tau') = D(\hat{\tau}') + D(\tau'|\hat{\tau}') \quad (49)$$

where $\hat{\tau}'$ is the τ expected, and τ' is that observed. When observations are contrary to predictions (e.g. the theory predicts $T > C$ but $T < C$ is observed instead), then $K < 0$.

- **Incomplete:** Unlike for the examples encountered so far, it may be the case that the τ of a large and complex theory is not fully connected. If so, the graph representing such a theory would have separate components, each relating inputs and explananda that are distinct from each other. For example, a theory may define separate relations between two distinct pairs of phenomena:

Since $D(\tau) > D(XrY) + D(ZpW)$, and since this τ has only two possible explananda (Z, W), clearly a larger K would be attained by finding patterns or new connections within the τ , which would, respectively, give it a more compact description and expand the τ 's explanatory power.

- **Contradictory:** It may also be the case that an empirical theory is not fully consistent and coherent, in which case its united τ may lead to contradictions. For example, a theory may contain two opposite statements about a correlation between two quantities. Conditioning on just these two statements, a contradictory theory may be represented as:

whereas an incoherent theory may be represented as containing a random element, corresponding to the arbitrariness about which version of the theory is chosen to explain each empirical observation:

Even though it may yield non-zero knowledge, this theory is sub-optimal nonetheless, for at least two reasons. First, $D(\tau)$ is larger than an equivalent measure without the contradiction, because it will include the entropy of the random variable for r and its two lower-level descriptions. Second, $E[H(Y|X, \tau_i)]$ is smaller than the average explanatory power of the two theories, because either theory is applied randomly to any observation; and/or it could be equal to zero if the two version of the theory cover every possible observation. Thus, K would increase if the contradiction is resolved, which may be attained, for example, by conditioning the theory to a subset of observations (see 4.2.1), or by reducing it to a theory that does not entail uncertainly, or by expanding it into a theory with an additional input input that determines which relation should be expected for a given observation.

The literature offers several attempts to formalize the nature of theories, including a graphical calculus to combine and compare them (e.g. [27]). The graphical methods developed here are entirely compatible with these approaches, but they are also more complete, because τ graphs not only represent theories, but also connect them to empirical consequences and, with K , allow to evaluate a theory's explanatory power.

4.3.2 Describing a study's methodology

If a study's theory describes the relation between concepts, a study's methods describe the conditions under which those concepts and relations will be observed. Therefore, the description of a study's methods will consist in an expansion of the theory τ aimed at describing how each concept or relation is defined, obtained, conditioned, etc.

The methods used in a study are usually not entirely idiosyncratic, but rather follow general schemes. These schemes constitute the broader "methodology" that characterises a research field, and which gave the basis of the specific methods used in a study. Therefore, in order to compare different studies within a field, we propose to use general schema to describe the τ of a particular methodology, at least at higher levels.

In principle, all studies pertaining to a field are following the same methodology and testing the exact same theory. In practice, however, each study will inevitably diverge, at some level, from the general scheme, because it is bound to have to confront, explain and control the contingencies and particularities of the instantiation of the phenomena being studied. Therefore, a corpus of studies will be characterised by a set of theory/methodology descriptions which are identical at higher levels, but that eventually diverge from each other at lower levels.

4.3.3 Example 3.1: The K of randomized controlled experiments

Randomized controlled experiments are a suitable example to make, because their methodology is well-codified.

In its most succinct form, the theory of an RCE is a statement that factor X is causally related to factor Y , e.g. $\boxed{X} \rightarrow \diamond \rightarrow \boxed{Y}$

In the logic of RCE methodology, X is a binary variable, generated by the randomization process that allocates the condition of interest (treatment T) to one part of the sample, and not to the other (the controls C). A 50/50 randomization generates exactly one bit of information. The nested description of both T and C will consist in a series of steps that are, ideally, identical except for one single detail of interest. Prior to implementing the procedure, an RCE will typically specify two sets of conditions. One are the subject conditions S , consisting in a series of characteristics that subjects need to meet (e.g. human participants may need to be of a specified age range or gender etc.). The other are experimental conditions (e.g. room temperature, characteristics of experimenter, etc.) C . Finally, the outcome is usually measured indirectly, via some measurement procedure M .

Therefore, the theory τ above can be expanded into the following general scheme for RCE methodologies:

where S, C are sets of subject and experimental conditions, and P_1, P_2, M are further describable in terms of chains of steps, linked by relations that are usually either binary (when one step simply follows from the other) or ternary (if the step requires a parameter, for example when a certain amount of time needs to pass between one step and the next).

To quantify the K of such a study we proceed as follows:

- The input X is where information is generated. Assuming a 50/50 allocation to treatment and controls, $H(X) = 1$. This entails that a maximum of one bit can be extracted from the explanandum.
- The explanandum Y is a description of the measured outcome. Since the maximum achievable result is when X and Y are perfectly correlated, $H(Y) = H(X)$. Thus, in this example of 50/50 randomization, $H(Y) = 1$. This equates the prediction that, if the treatment works perfectly well, then the outcomes measured via M for treatments and controls will be perfectly segregated, with all values lying below the median value for one group and above for the other.
- The residual ignorance is either the entropy or the cross-entropy of the result observed, depending on whether the trial is testing a directional prediction or not (usually, it is). In either case, this term is measured from the observed distribution of values of the outcome, i.e. the relative frequencies above vs. below the median, averaged over treatments and controls. Thus, if all treatments are, say, above the median, and all controls below, as the theory predicts, then $E[H(Y\hat{Y}|X, \tau)] = 0$, whereas if there is no difference then $E[H(Y\hat{Y}|X, \tau)] = H(Y)$, and if the observed results are the opposite of those predicted then $E[H(Y\hat{Y}|X, \tau)] > H(Y)$.
- The repetitions n_Y are equated with the sample size, with $n_Y = n_X = n$

The total K will be:

$$K = \frac{1 - H(Y\hat{Y}|X\tau)}{1 + 1 + D(S) + \frac{c + D(C) + D(P_1) + D(P_2) + D(M)}{n}} \quad (50)$$

where $c = 31.74 + 16.56 + 7.34 + D(\text{and}) + D(\text{then}) + D(\rightarrow) + D(\text{via})$ is the complexity of the scheme. Note that the scheme above is merely one possible scheme, and in principle any scheme will work, because any scheme that is applied consistently across studies contributes a constant that can be ignored. Differences in the size of τ and the number of divergences between studies will be determined by the lower-level descriptions of each term.

4.3.4 Example 3.2 Measuring and studying replicability

One of the major concerns driving Metaresearch is that many sciences might be facing a “crisis” of replicability, purportedly caused by questionable research practices [30]. However, consensus on the true prevalence and causes of irreproducibility of results has been hampered by the difficulty of measuring replicability and comparing it across widely different methodologies and research questions.

Current measures of replication assume that phenomena measured in science, and consequently the truth of any scientific claim, either exists or not. Consequently, metrics of replicability aim to establish the extent to which replication results are compatible with the results originally reported. This compatibility is established using standard but arbitrary statistical thresholds, such as whether the replication result rejects the null at the 0.05 level, whether it falls within the 95% confidence interval or prediction interval of the original finding, whether a pooled summary of all results still rejects the null, etc. These criteria are implemented identically in all replication studies, regardless of the original study’s field and methodology, with the implicit assumption that “true” findings should replicate by these standards at least 95% of the time (e.g. [31, 32, 33, 34, 35, 36, 37]).

The above approach to measuring replicability has been repeatedly criticised in the literature as it implies a series of false dichotomies [38, 39, 40] and it ignores the enormous diversity of phenomena and methodologies that could be examined. It is intuitively clear, for example, that results from fields dealing with highly complex phenomena and less codified methods might be less likely to be replicated exactly, regardless of the truthfulness of the original report [41, 42, 43, 44, 45, 46, 47, 48, 49, 50, 40]. Although less explicitly discussed in the literature, it seems equally intuitive that studies with more complex methodologies, and studies that require a greater amount of conditions to be satisfied will, all else equal, replicate less successfully. Finally, it is intuitively clear that complexity of subject matter and complexity of methodology are partially correlated. Crucially, this difficulty in replicating

is manifested in lower effect sizes found in replication results relative to the original, which has nothing to do with the validity of the original finding.

The K function allows most of these limitations to be solved. It gives a new and universal metric of replicability and a model through which replicability can be explained and forecast. The basis of the method was proposed in [1] which analysed the effects of random divergences on the system and concluded that replication and original study are connected by the relation:

$$K_r = K_o 2^{-\lambda \cdot d} \quad (51)$$

K_r and K_o are the (expected) K values of replication and original study, and $\lambda \cdot d = [\lambda_{Y1}d_{Y1} + \lambda_{Y2}d_{Y2} \dots \lambda_X d_X \dots \lambda_{\tau_1} d_{\tau_1} \dots]$ is the inner product of two vectors of values of divergences d and “weights” λ [1].

Taking the logarithm and re-arranging equation 51 we get a linear model for reproducibility and replicability as:

$$\log \frac{H(Y_r) - H(Y_r|X, \tau_r)}{H(Y_o) - H(Y_o|X, \tau_o)} = \log \frac{D(\tau_r)}{D(\tau_o)} - \sum \lambda_i d_i \quad (52)$$

where τ_r, τ_o are the τ of replication and original study, d are divergences obtained as in section 6.3.2) and λ are weights that can be estimated by maximum likelihood. Applied, for example, to the case of randomized experiments described in section 4.3.3, assuming that the same study scheme is used, and removing all constant terms, the equation simplifies to:

$$\log \frac{1 - H(Y_r|X, \tau_r)}{1 - H(Y_o|X, \tau_o)} = \log \frac{D(S_r) + (1/n_r)[D(C_r) + D(P_{1r}) + D(P_{2r}) + D(M_r)]}{D(S_o) + (1/n_o)[D(C_o) + D(P_{1o}) + D(P_{2o}) + D(M_o)]} - \sum \lambda_i d_i \quad (53)$$

where the left-hand side is the empirical estimation of replicability, and the right-hand side includes the two main factors that are predicted to affect replicability: the relative size and specification of the τ and the effects of random divergences between the two τ .

A new, universal metric of replicability

Even if we ignore (or lack) all information about the τ and divergences on the right-hand side of equation 52, we obtain an empirical estimator of replicability as

$$\Lambda = \log \frac{k_r}{k_o} \quad (54)$$

where $k = (H(Y) - H(Y|X, \tau))/H(Y)$ is a standardized measure of the effect measured in the study. This metric circumvents all criticisms and limitations of current reproducibility metrics, because it is

- Non-binary
- Non bounded, since $\Lambda \in (-\infty, +\infty)$ and is negative when reproducibility is low, and positive when reproducibility studies yield stronger results than the original ones. Hence, it accommodates a broader range of cases compared to reporting the percentage of replicated versus not.
- It is a scale-free metric and, just like K , is obtainable from results of any study (any metric), with no requirement to make additional assumptions.
- It can still be converted to an intuitive ratio (or a percentage), via

$$2^\Lambda = \frac{k_r}{k_o} \quad (55)$$

which expresses the average knowledge loss or gain observed across replications of an original study. Unlike the percentages used in previous reproducibility studies, this ratio has a precise and intuitive meaning and is not derived from arbitrary criteria and artificial cut-off points.

A general theory of reproducibility and replicability

To any extent that the τ and divergences are measurable, it is possible to derive from the right-hand side of equation 52 specific predictions about what to expect from a replication, and how to improve replicability. In particular, the equation tells us that:

- if $\tau_r = \tau_o$, the replication is exact in all aspects, which is the assumption under which reproducibility is evaluated, then

$$\Lambda = 0 \iff K_r = K_o \quad (56)$$

and we should expect the replication to yield the same results as the original.

- If $D(\tau_r) = D(\tau_o)$ and $d \neq 0$, the methods have identical description lengths (same amount of complexity), but random or uncontrolled divergences. This is the condition of “replicability” and is expressed as

$$\Lambda = - \sum \lambda_i d_i < 0 \quad (57)$$

which predicts that, on average replications should yield smaller effect sizes than the original study, in proportion to how many random divergences separate them from it. These random divergences may characterize the subjects (e.g. the two populations in which the studies are conducted) as well as any other aspect of the methodology.

- If $\tau_r \neq \tau_o$ and $d = 0$, the methods in the replication study differ *non-randomly* from the original, then

$$\Lambda = \log \frac{D(\tau_r)}{D(\tau_o)} \quad (58)$$

and the replication might yield either larger or smaller effects than the original, depending on whether it includes more or fewer details and specifications than the original study.

Depending on characteristics of the τ examined, some factors will be more relevant than others. In the example of the RCE equation 53, $D(S_r)/D(S_o)$ are multiplied by their respective sample size, and therefore we would predict subject conditions S to exert a greater influence, leading to the specific prediction that $\Lambda \propto \frac{D(S_r)}{D(S_o)}$ for large n_r .

A more generalised prediction is that replicability will be lower the smaller the K_o of the original study. This follows because the K of the original study reflects both the complexity of subject matter (large residual ignorance) and methodology (large τ).

Precision in testing these predictions is limited by the fact that, in most replicability studies the τ of the original study is not actually known, or at least is known incompletely. This is due to at least two reasons: first, because space in publications is limited and authors usually lack guidance concerning the level of detail they need to provide and, second, because relevant factors might be unknown or overlooked by the authors who reported the original finding - factors including characteristics of the observations collected, conditions in which the study was carried out, and the “tacit” component of methodological expertise that went into conducting aspects of the study.

However, it may be possible to use equation 51 to infer the amount of information missing in τ_o , and therefore to derive testable predictions that concern exclusively relations between observables. Such methods and tests will be presented in future work.

4.4 Empirical qualitative knowledge

One of the most ambitious promises of K theory is to overcome the artificial distinction between “quantitative” and “qualitative” research, that is, between numerical and non-numerical forms of evidence and knowledge. This goal is made considerably closer by the methods proposed in this essay.

Many methods of qualitative research entail the creation of categories and the identification of relation between them. In Grounded Theory [51], for example, the researcher collects information about a particular phenomenon (via interviews, or examining texts, images, and any other relevant source), and constructs categories that are progressively revised and refined, until “saturation” is reached (the researcher no longer encounters notable new patterns) and a theory is constructed in terms of higher-level connections between these categories. Similarly, in Thematic Analysis, which is another widely used qualitative methodology, the researcher codes a corpus with the objective of identifying recurrent thematic patterns [52].

Further more, graphical methods, whilst still not widespread in qualitative research, have been recently proposed to aid qualitative methodologies including: process tracing [53], text analysis [54], narrative data [55], sociological theory [56], relationships between causal theories [27].

To any extent that phenomena and/or theories of qualitative studies are describable in terms of categories, relations between them, and graphs, then a τ can be used to describe them and a K value can be derived. And the τ derived from any number n of sources from a corpus can be fed into equation 13 to explore a variety of questions about how much knowledge is obtained by applying a particular theory to a corpus. Questions might include:

- How similar is one source to another?

$$K(\tau_2; \tau_1) = \frac{D(\tau_2) - D(\tau_2 \oplus \tau_1)}{D(\tau_2) + D(\tau_1)} \quad (59)$$

- How well does a particular theory explain a given corpus?

$$K(\sum \tau; \bigoplus \tau) = \frac{\sum D(\tau) - D(\bigoplus \tau)}{\sum D(\tau)} \quad (60)$$

- How distant is new source from a corpus?

$$K(\tau_i; \bigoplus \tau) = \frac{D(\bigoplus \tau \oplus \tau_i) - D(\bigoplus \tau)}{D(\bigoplus \tau \oplus \tau_i)} \quad (61)$$

- In progressively collecting new data, how much new information is obtained? How close is researcher’s saturation?

$$1 - K(\tau_{n+1}; \bigoplus^n \tau) = \frac{D(\bigoplus^{n+1} \tau | \bigoplus^n \tau)}{D(\bigoplus^{n+1} \tau)} \quad (62)$$

4.4.1 Example 4.1: Process tracing

Process tracing is one of the most widely used methodologies to conduct case studies of complex phenomena. Whilst different formalizations of process tracing exist, the essence of the methodology lies in the “systematic examination of diagnostic evidence selected and analyzed in light of research questions and hypotheses posed by the investigator” [57]. Specifically, evidence is collected and evaluated in relation to either a descriptive hypothesis about how a particular case or process unfolded (descriptive inference) or a hypothesis about the causes that led to a particular event (causal inference) [57].

The use of graphs has already been proposed in the process tracing literature to facilitate the evaluation of hypotheses. In particular [53] argued for the use of Direct Acyclic Graphs (DAG, a popular tool for causal analysis [13]) to represent, first, an abstract causal chain hypothesised to

explain a general class of phenomena and then, second, to represent an “event-history map” that reflects how a particular case matches the causal graph. Researchers would then evaluate how well the evidence pertaining to specific cases matches the general scheme.

Another relevant attempt to formalize process tracing entails the use of Bayesian methods [17]. Bayesian process tracing involves the formulation of specific and mutually exclusive hypotheses, then the assignment of priors to these hypotheses, then the identification of relevant evidence to evaluate the likelihood of evidence under each hypothesis, in order to compare alternative hypotheses using likelihood ratios and Bayes factors. Bayesian process tracing is a promising but still somewhat controversial method due to known limitations, which include the difficulty in formulating rigorous models for complex phenomena and the arbitrariness in estimating likelihoods [17].

K theory methodologies may offer an alternative, simpler and more general approach to formalizing process tracing, that maintains some of the benefits of graphical and Bayesian approaches whilst avoiding some of their limitations. In the K approach to process tracing, researchers do not construct formal DAGs but a τ that captures events and relations between them, and then evaluate the balance between explanatory power and plausibility of the explanation using K .

The τ could be obtained inductively or deductively. In the inductive approach, facts are first gathered and a τ is built to connect them via a hypothesised process that may include any number of steps and relations, both observable and not. In the deductive construction, a “theory” τ is first formulated, and then evidence of specific cases is connected to it. In either case, describing the explanatory scheme can suggest what observable evidence should be collected to test and refine the theory.

The relative value of explanatory schemes is evaluated with a K function that has as explananda each event comprising the evidence and as input all the events or conditions that are known to be true or necessary. Each event is given an information weight $-\log p(y)$, where the probability could, as discussed before, be derived in three ways:

- Uninformed: in which we ascribe a uniform probability, thus attributing 1 bit of information to each item.
- Subjective: using techniques of elicitation developed in Bayesian process tracing, each item is given a weight based on a “decibel” scale [17]. The researcher is asked to evaluate how loudly the data “speaks” in favour of a hypotheses, using sound volume as a sensorial analog. The decibel scale is a logarithmic scale, and so this method effectively elicits the equivalent of log-likelihoods.
- Structural: if each event is given a prior description based on common schemes, then a surprisal measure is also given by its description $D(y)$. The surprisal relative to a hypothesis, $D(y|x, \tau)$ could then either be given a standard value of 0, as in the Uninformed method or, with a suitably constructed language, as a divergence between the description of the event without and with the explanatory hypothesis.

For example, let’s assume that we have three events A, B, C and the hypothesis

and we compare this to a τ with both B and C as explananda, where

$$D(B|A, \tau) = 0 \text{ if BrA is observed} \quad (63)$$

$$D(B|A, \tau) = D(B) \text{ otherwise} \quad (64)$$

These are qualitative statements, not necessarily statistical ones, and they do not entail any pre-defined description. Any agreed-upon criterion for what counts as evidence of that particular explanandum will do, provided it is used consistently.

We do the same for BsC and applying equation 13 gives

$$K(B, C; A, \tau) = \frac{D(B)n_{ArB} + D(C)n_{BsC}}{n[D(A) + D(B) + D(C)] + D(r) + D(s) + 12.5} \quad (65)$$

where $n \geq 1$ is the total number of cases examined and $n_{ArB} \leq n$ and $n_{BsC} \leq n$ are the number of times the relevant evidence is observed in these cases. Thus, observing ArB but not CsB would count as partial evidence in one particular case.

The nested descriptions $D(B)$ and $D(C)$ thus play the function of “weights” that reflect the relative informativeness of that particular item of evidence in an absolute sense, not in relation to any hypothesis. These values should be determined in advance of the analysis, and may be derived from any one of the usual methods: uninformed equal value, subjective surprisal, or quasi-objective structural description length.

Let’s assume we have a competing explanation:

such that the K is

$$K(B, C; A, D, \tau) = \frac{D(B)n_{ArB} + D(C)n_{DsC}}{n[D(A) + D(B) + D(C) + D(D)] + D(r) + D(s) + 16.4} \quad (66)$$

where n_{DsC} is now the number of times DsC is observed. Due to the greater complexity of τ_2 , equivalent amount of evidence would favour τ_1 , whereas a situation of partial evidence would call for an evaluation of the weighted sum of evidence relative to that of the additional inputs and larger size of τ_2 .

Therefore, K might constitute an objective basis on which to evaluate the relative merits of two or more competing hypotheses in the context of multiple and complex sources of evidence, avoiding virtually any posterior subjectivity. The process of constructing competing hypothesis would also help identify sources of evidence that are lacking or particularly influential, guiding further data collection.

In the more subjective version of the semi-objective approach, we would have

$$D(B) - D(B|A, \tau) = \log \frac{p(B|A\tau)}{p(B)} \quad (67)$$

and the researcher would evaluate how much “louder” the evidence is without the offered explanation. Operating in relative terms might offer some advantages over the method of “fully inhabiting the competing hypotheses” followed by evaluation of their absolute “loudness”. But even if one adopted a fully subjective weighting of evidence, the K method would retain some advantages over alternative process tracing methods, including the use of a graphical language that does entail causality or any particular structure (e.g. does not need to be acyclic), a quantifier that expresses the relative merits of equally supported theories, and the ability to evaluate the explanatory success of a hypothesis cumulatively, across multiple K studies, even when different types of evidence are obtained in each case. Doing so simply entails a summation of the evidence across cases. Given evidence E from s cases:

$$K = \frac{\sum_s \left[\sum \log \frac{1}{p(E_{s,i})} - \sum \log \frac{1}{p(E_{s,i}|\tau, A_1)} \right]}{\sum_s \left[\sum \log \frac{1}{p(E_{s,i})} \right] - \log p(\tau)} \quad (68)$$

where evidence cumulates but the explanation is constant, so that, all else equal, the more evidence from more cases is successfully explained by a theory, the higher the value of K for that theory.

4.5 Merging qualitative and quantitative knowledge

Since all terms of the K function are amenable to description, a K can be obtained from any combination of qualitative and quantitative evidence. Taking again the process tracing example discussed in section 4.4.1, let B represent the result of a regression analysis, and let C represent how well a given explanatory scheme describes a corpus of m texts, measured with equation 60. We re-describe them formally as τ , τ_B , τ_C with corresponding K values,

$$K_B = \frac{D(\tau_B) - D(\tau'_B)}{D(\tau_B)} = \frac{n_Y[H(Y) - H(Y|X, \tau_{YX})]}{D(\tau_B)} \quad (69)$$

and

$$K_C = \frac{D(\tau_C) - D(\tau'_C)}{D(\tau_C)} = \frac{\sum_m D(\tau_i) - D(\bigoplus^m \tau_i)}{\sum D(\tau_i)} \quad (70)$$

Since $K \leq 1$, we can replace the n_B, n_C, n terms in equation 65 with the K values, and we get

$$K(\tau_B, \tau_C; \tau_A, \tau) = \frac{D(\tau_B)K_B + D(\tau_C)K_C}{D(\tau_A) + D(\tau_B) + D(\tau_C) + D(r) + D(s) + 12.5} = \frac{n_Y [H(Y) - H(Y|X, \tau_{YX})] + m[E[D(\tau_i)] - D(\bigoplus^m \tau_i)/m]}{D(\tau_A) + D(\tau_B) + D(\tau_C) + D(r) + D(s) + 12.5} \quad (71)$$

Thus, the numerator is now a computation of the net number of bits being compressed by the regression equation and the qualitative scheme. Like all other K , this metric will increase with the number of instances being compressed - that is, with increasing sample size of the regression result and/or number of texts examined in the corpus. It will also increase with the accuracy of measurement - that is, with the amount of information extracted in each case, in the form, respectively, of the entropy of the response variable and the complexity of the description scheme.

5 Discussion

This essay completes the paradigm proposed in [1], by showing how the knowledge about any phenomenon can be described, then translated in an information quantity, and then plugged into the K function in order to evaluate how much of the phenomenon's information is being compressed, and therefore how much knowledge is possessed about that phenomenon. This work achieved two main objectives: first, it defined a consistent and coherent mathematical approach, that yielded important insights into the meaning and uses of K itself (section 6); second, it offered practical examples of the methodology, to illustrate how to calculate a K from any type of knowledge claim.

The central idea in this paradigm is that all knowledge results form the conditioning of a τ . The τ is a recursive, potentially infinite structure of structures (Figure 2.1). Knowledge results from an abstraction, i.e. a reduction of this τ to a finite structure that involves a limited set of categories and relations between them. Each phenomenon that can be described, measured and explained is represented, usually incompletely, by a smaller τ of finite sizes, which in turn can be partially described by a nested labelled hypergraph. Since all particular τ 's are defined as conditions of a larger τ , there is ultimately just one τ , from which all particular knowledge is derived by conditioning.

The mathematics of τ yielded a number of new insights into how K is interpreted and calculated (see 3.1). In substantive terms, K measures the proportional costs of expanding the explanans - that is, of making one's explanations, theories or methods more complex than they already are. Put in simpler words, K measures how much room remains for improving one's understanding of some phenomenon or class of phenomena (see 3.1.1). For a given description τ of a system, K measures the proportional reduction of its description length attained by some process or computation. A larger K results from abstracting patterns from a given set of particular phenomena, shedding all lower-level information (see 3.1.2). These and other results obtained in this essay are consistent with the assumptions of K theory - namely that informations is finite and knowledge is information compression - validate this approach as a coherent paradigm to understand knowledge.

The methodology to measure the τ of any system is in principle simple and may be mastered by any researcher, regardless of their specific expertise: it mainly requires translating into a graph the key categories and relations of interest - that is, the concepts used to describe a phenomenon, the theory proposed to explain it, and the methodology used to measure it. Simple graphical operations can then be implemented to combine, separate, nest and de-nest the graphs, and the resulting D values can be plugged into the K function.

However, the practical implementation of these methods in real research examples may not be quite as straightforward. In particular, in order to measure and compare the D of theories and methodologies it is necessary to use a single, consistent scheme. Defining the appropriate scheme

might prove to be challenging in proportion to the complexity and lack of codification of a field’s methodology. Further, even in simple forms of knowledge like those of logic, a complete evaluation of K ideally requires an evaluation of the information content (surprisal) of the lowest-level concepts (see for example section 4.1). Again, this may be harder to do in proportion to how complex and imprecisely defined these concepts are. Finally, this and several other hidden methodological decisions are made in calculating K , and therefore the methodology itself may need more work to be codified and reduce its own arbitrariness. Therefore, future studies will aim to assess to what extent the methods proposed in this essay can be usefully applied to real cases of research, in order to track the progress made and answer other meta-scientific questions in ways that are sufficiently accurate, consistent and valid.

A second limitation to point out is that, whilst results of this essay corroborate several tenets of K theory, the theory itself can and should be subject to further empirical testing. The theory makes numerous meta-scientific predictions, derived from the general prediction that the magnitude of K reflects the “hardness” and “scientificity” of any knowledge claim (see more details in [1]). The most developed predictions concern reproducibility, which were further articulated in this essay (section 4.3.4) and future work will aim to verify them on empirical data.

To the best of the author’s knowledge, none of the specific ideas and methods presented in this essay have been proposed before. However, these ideas were inspired and/or elaborated from concepts and methods used in several broad and expanding literatures, including complexity science (e.g. [4, 58]), graphical knowledge representation (e.g. [6, 7, 8, 27]), graph theory (e.g. [15, 16]), probabilistic graphical modelling (e.g. [11, 12, 13]), and algorithmic information theory (e.g. [2, 10, 14]).

Future work could seek further integration with these various literatures. This would probably expand and integrate K theory and might also lead to new insights and applications in other fields that build on these same literatures. In particular we speculate that measures of τ and K may help any research that attempts to give computational accounts of elusive phenomena, including cognition (e.g. [59]), art (e.g. [60]), humour (e.g. [61]), and consciousness (e.g. [62, 63]).

6 Extended Methods

6.1 T as a set of hypergraphs

6.1.1 Principled definition of τ

The quantity $-\log p(\tau)$ is defined as the information required to identify the particular theory/methodology component τ in a “universe” set T of which τ is an element [1]. To quantify it, therefore, we need to define the set T . Following the ideas laid out in [1], we reason as follows:

1. Every knowledge claim concerns the relation between the information in an input X and that in an explanandum Y . The τ is the structure that defines such a relation. The triple Y, X, τ is what we call a “system”.
2. Since all knowledge reduces to patterns of information and vice-versa, then each entity and/or relation that constitutes a knowledge claim must in turn be ultimately knowable (describable, understandable) in terms of other entities and relations, i.e. other systems.
3. Since patterns are fundamental to relations, the τ component of any such description is the locus of the knowledge about the τ being described.
4. The “description” of any τ , therefore, must fundamentally consist in the encoding of its relation with other τ , relation that is also described by τ ’s.
5. These descriptions are separable from the τ that they describe, and yet connected to them. They are structures that acquire the meaning of describing a τ once they are conditioned on them.
6. We define this conditioning as a “nesting” of the description within the entity it refers to. In other words, a τ is “described” by its connections to other τ ’s, which are either nesting it or nested within it.

7. The description is recursive and can in principle be infinite. In practice, however, a description is never infinite and it ends when no further lower- or higher-level description is given.

From the first statement we conclude that every knowledge claim can be represented as a graph that connects inputs and explananda; from the subsequent ones that the graph must have a “nesting” property that makes its description arbitrarily recursive. These conclusions are given mathematical form in section 6.1.2, where we shall define T as a set of nested hypergraphs whose description length is given by a recursive function, which we define in section 6.2.

6.1.2 Size of T

What type of graph can represent a τ ? Since the input X and explananda Y that a τ connects may consist in any number of random variables, and since the type of relations connecting them is variable, then the τ must be represented as a nested labelled hypergraph (directed or undirected) in which vertices represent the concepts and hyperedges (or hyperarcs) represent the relations, with arity ≥ 1 .

6.1.3 Knowledge hypergraph

Mathematically, a hypergraph $H = \langle V, E \rangle$ is defined by a set of vertices $V : \{v_1, v_2 \dots v_m\}$ and a set $E : \{e_1, e_2 \dots e_n\}$ of hyperedges and/or hyperarcs. Hyperedges are unordered subsets of nodes (that is, $E \subseteq \mathbb{P}(V)$, where $\mathbb{P}(V)$ is the power set of V) whereas hyperarcs are ordered pairs $\langle S, v \rangle$ where $S \subseteq \mathbb{P}(V)$ is the “tail” and $v \in V$ is the “head”. A labelled hypergraph includes a vocabulary composed of a set of vertex labels T_V and a set of hyperedge labels T_E , and labelling functions l_V, l_E that associate each vertex and edge with one or more elements of the vocabulary. The hypergraphs that we shall consider are “nested” because both vertices and hyperedges may in turn consist of hypergraphs.

Crucially, we shall draw no distinction between labels and nested descriptions of vertices: labels in a graph are mere pointers that indicate which description applies to which entity. The nested descriptions are mere structures, and the τ is just a structure of structures, in principle endlessly recursive. This is the key difference between an ordinary hypergraph and a τ . Although we shall often use the two terms interchangeably, a τ is a recursive structure of structures, whereas a graph is a mathematical object that describes a τ , usually incompletely.

Therefore, we shall posit that any τ is universally describable by a particular type of directed labelled hypergraph, defined as follows:

- A hypergraph and each of its hyperarcs represents a τ , which connects one or more inputs to one or more explananda, embodying the flow of information between them.
- The graph’s hyperarcs do not necessarily have just one head, but are all always an *ordered* subset of nodes. The order of nodes allows to identify the tail (inputs) and heads (explananda), and any ordering convention can work, assuming it is applied consistently.
- Depending on the type of relation, the order of inputs and/or explananda may be partially or totally irrelevant, giving rise to a *symmetry*, defined below.
- We express the symmetry $s(e_i)$ of a hyperarc e_i as the factorial of the number of vertices whose order does not need to be specified. For example, a hyperarc with a single head $\langle S, v \rangle$ for which the order of the inputs is irrelevant will have symmetry equal to $s(e_i) = |S|!$; a hyperarc that represents a completely unordered relationship (that is, in which explanandum and input are equivalent) will have symmetry $s(e_i) = (|S| + 1)!$, whereas a hyperarc representing a fully ordered set of vertices (i.e. a fully specified sequence of entities) will have symmetry $s(e_i) = 1$.
- Whereas the vocabulary $\mathcal{T} = T_V \cup T_E$ of typical hypergraphs consists of ordinary sets (in which each distinctive label is present once), we shall typically describe the total set of labels of vertices and hyperedges as *multisets*, $\{m_{v1} \times v_1, m_{v2} \times v_2 \dots m_{vp} \times v_p\}$ and $\{m_{e1} \times e_1, m_{e2} \times e_2 \dots m_{eq} \times e_q\}$ of the same size as V and E - that is, in which the m terms represent the multiplicities of each label, and $\sum m_{vi} = |V|$ and $\sum m_{ei} = |E|$.

An example of a hypergraph representing an elementary τ is:

in which the two inputs labelled A, B are related to the explanandum labelled C by the relation labelled r . It is usually not necessary to represent the order of the elements of the hyperarchs explicitly (e.g. with numbers or letters next to each arrow). We shall simply adopt a convention that vertices connecting to a relation are ordered from the top down, both for tails and for vertices. A τ can also be described in text as $ABrC$, following the convention of reading the order of vertices from left to right, and using lower-case letters for relations and upper-case letters and numbers for vertices.

6.1.4 The nesting operator

To represent lower-level descriptions, we shall typically insert a complete graph into the element of another. For example, take the τ

We could add a nested description of A as the explanandum of two other inputs, 1, 2 and a relation q . This will be graphically represented as showing a blown-up version of the node A next to its graph, with inside it a self-contained graph with A as the explanandum:

To symbolize mathematically the nested relation between two graphs we will adopt a new operator, $\odot$, whose shape is chosen to symbolizes how the graph on the left of the symbol is contained inside the one on the right of it (and how, conversely, the other contains it). Formally, let g, h be two labelled hypergraphs, and let h be the lower-level description of one element l_i of the vocabulary $\mathcal{T}_g = T_{V_g} \cup T_{E_g}$ of g , then

$$h \odot g \equiv h \odot l_i \in \mathcal{T}_g \quad (72)$$

In the example above, the entire graph may be written as $(DpA)BrC$, $(12qA) \odot A$, or more compactly, $Dp[(12qA) \odot A]BrC = (DpA)BrC \otimes (12pA) \odot A$, where we use the symbols of the tensor product $\otimes$ to represent the independence and additive nature of the information contained in the two graphs (see section 6.2).

We can “de-nest” the description by reversing the nesting, which correspondinds to joining the two graphs. Mathematically, $Dp[(12qA) \odot A]BrC = Dp12qABrC$, or

An equivalent, representation that we may use at times to save space is

in which the explanans is pictured as being “folded inside” the higher-level element and connected to it. The nesting element can be “described” at the lower level as any element of a τ , not just the explanandum. In the graph above, for example, we could describe D as the input of A via p , i.e. $(DpA) \circlearrowleft D$, obtaining

We can also nest descriptions of relations. For example, in the graph above we might nest in r the entire sub-graph between p and q and call it p , obtaining $DpABrC$, $p12q \circlearrowleft p$, and the graphical representation is

where we represent the container as a rectangle because it would be impractical to use a romboid.

Note that the entire descriptor is a self-contained graph, whose composing sets of labels, vertices are distinct from those of the higher-level graph.

Crucially, therefore,

because the latter represents a graph with six vertices and two disconnected components, in which two distinct vertices have the same label, whereas the former is a graph with three vertices and one vertex described by a nested, separate, graph. Therefore whenever a graph is a description, to avoid confusion we shall always represent it as nested within another element - or adopt other clarifying convention.

6.1.5 Nesting of entire τ

As noted above, we need not assume that a higher-level entity and its nested description are connected, although if they are not, then the τ is fragmented. For example, there is nothing incoherent in $D((ApB) \circlearrowleft C)rE$, but if we de-nest it we obtain a graph with two disconnected components: $DCrE$, ApB .

Whenever we “give a label” to a graph (or any other object) to identify it (e.g. “graph g”, “graph h”, etc.), we are implicitly nesting the graph in an arbitrary symbol. Indeed, and in light of equation 13, we shall frequently give a highest level of description to the τ in terms of a nesting in the τ symbol itself:

If however we want to *conceptualize* the τ itself, in order to connect it to other τ 's, we need to create a connection as before. We can do so by expanding the arity of one of the relations, in order to generate the element that will connect the graph to the higher-level concept.

For example, the graph

could be nested in $\boxed{X}$, $\boxed{Y}$, or $\boxed{r}$, but to nest the concept of the *whole graph* in a descriptor symbol, say “ XrY ”, we operate as follows:

1. Expand the arity of the relation to include a new vertex, labelled as the nesting symbol:

2. nest the graph in a descriptive symbol

Retrospectively, we may understand the previous examples of nesting, in which nested descriptions are connected to their higher-level entity, as instances of this conceptualization.

6.1.6 Size of T and description length of τ

Since they describe all and every τ , sections 6.1.3 and 6.1.4 define what the elements of T are. This allows us to calculate the size of the T from which any given τ is derived, and therefore to derive a universal and optimal encoding $-\log p(\tau)$ of it. The minimum size of the set T must be sufficient to allow for the encoding of τ itself, which entails that T must contain *at least* all of the hypergraphs of equivalent complexity, i.e. the set of all hypergraphs that could be constructed with an equivalent number of vertices, hyperedges and labels.

Calculating the size of such a set is a relatively simple combinatorial problem, once the hypergraph is represented by what we may call a “labelled adjacency tensor”. By the latter we mean a multi-dimensional array in which each non-empty cell represents a hyperarc and contains a symbol corresponding to the label of that hyperarc. Adjacency matrices and tensors are ordinarily used to represent and manipulate graphs, but they are typically numerical in nature, where the numbers represent the weight of edges.

The adjacency tensor we shall use to describe a τ will have as many dimensions as the largest arity (number of vertices in hyperarc) in the graph and it will have each non-null element associated to a label (a distinctive symbol), which in turn can represent other tensors, due to the nesting property described above. Note that a number is also a symbol, so ordinary numerical adjacency tensors are just labelled tensors like others, where the meaning of the label comes from the lower-level description of the symbol in terms of a number (see section 6.2.1).

To quantify the size of $|T|$ based on the adjacency tensor of the corresponding τ , we separate the problem between calculating the total number of possible combinations of graph structures with the same adjacency tensor and multiplying it by the total number of combinations of labels for vertices and for hyperarcs.

Given a hypergraph $g = \langle V, E, T_E, T_V, l_V, l_E \rangle$ defined as above, with $|V| = d$ vertices and r hyperedges e_i of maximum arity a the total number of graphs of equivalent description length is given by

$$|T| = \frac{d^a! \times d! \times r!}{r!(d^a - r)! \times \prod_{j=1}^p m_{vj}! \times \prod_{i=1}^q m_{ei}! \times \prod_{i=1}^q s(e_i)} \quad (73)$$

in which the m terms are the multiplicities of the labels, and $s(e_i)$ is the hyperarcs’ symmetry.

If the graph contains relations of different arity, i.e. arity smaller than a , its description can be accommodated by including an additional vertex with a distinct “null” symbol, leading to $d = |V| + 1$.

The corresponding amount of information needed to identify a particular element $\tau \in T$ is the logarithm of this quantity, which can be written in terms of binomial or multinomial distributions as:

$$\log |T| = \log \binom{d^a}{r} + \log \binom{d}{m_{v1}, m_{v2} \dots m_{vp}} + \log \binom{r}{m_{e1}, m_{e2} \dots m_{eq}} - \sum_{e_i=1}^q \log s(e_i) \quad (74)$$

in which the elements on the right-hand side quantify, respectively, the (description length of the) structure of the graph, the labelling of vertices, the labelling of relations, and the symmetry within hyperedges. The quantity 5 can be calculated exactly for small graphs or approximately for larger ones, using Stirling's approximation:

$$\log_2 n! \sim \frac{1}{2} \log n + n \log \frac{n}{e} + c \quad (75)$$

where $c = \log_2 \sqrt{2\pi}$ (see Appendix A1 7).

If the graph is nested within some other entity which conceptualizes it (see section 6.1.5), then a relation is established between the the nesting element and one of the elements in the τ . Mathematically, this requires adding the information needed to identify the element, leading to:

$$\log |T_{nested}| = \log |T| + \log(d + r) \quad (76)$$

if the graph is a nested description of a concept.

6.2 D function

In section 6.1.6 we defined the size of τ in terms of a single set T , corresponding to a single level of description of the τ . The last step to complete the quantification of τ consists in extending this measure recursively to any number of nested descriptions, effectively allowing us to expand the set T *ad libitum*. We do so by defining a recursive function we will indicate as D , for “description”.

Let τ be a subset of a set T and let $\theta \subseteq T$ be the set of elements that are equivalent according to some equivalence relation, with $\tau \subseteq \theta \subseteq T$. Indicating any subgraph nested in τ as $\tau_i \supset \tau$, the complete description length of τ is given by the recursive function:

$$D(\tau) = 0 \text{ if } T = \emptyset \quad (77)$$

$$D(\tau) = \log \frac{1}{p(\tau)} + \sum_{\tau_i \supset \tau} D(\tau_i) \text{ otherwise} \quad (78)$$

where, unless otherwise specified,

$$p(\tau) = \frac{|\theta|}{|T|} \quad (79)$$

with $|T|$ and $|\theta|$ calculated as in equation 73. Allowing for other specifications of $p(\tau)$ permits us to accommodate probability distributions that are derived by other, empirical or theoretical, means (e.g. Bayesian updating, qualitative elicitation of beliefs). But this is merely a formal move to avoid binding our definition too strictly, because any such alternative probability distribution is ultimately still characterizable as resulting from an equivalence relation, as indeed all probabilities are (see section 6.2.1). In the most general case, i.e. when there is no prior information, the probability is uniform and solely determined by the sizes of sets T and θ .

6.2.1 D() as a universal descriptor of τ

Applied to the graphical representation defined previously, the D function in 7, gives a consistent and coherent measure of the description length of all fundamental mathematical objects of interest to us:

- **A symbol** is a special case of a graph with only one node, one edge of arity one, and a single label:

$$\boxed{s}.$$

Its universe set has $|T| = 1$ and it will therefore have

$$D(s) = 0 \text{ if } s \text{ is a symbol} \quad (80)$$

This would be expected theoretically, since a symbol taken alone has no meaning and carries no intrinsic information. A symbol only carries information in relation to other symbols, i.e. its sole purpose is to represent its difference relative to a set of alternatives.

- **A set** s is a collection of $|s| > 0$ symbols, all different from each other. It contains an amount of information equivalent to that needed to identify any element within it, i.e.

$$D(s) = \log |s| \text{ if } s \text{ is a set} \quad (81)$$

- **A labelled set** is a set whose elements have a nesting description with non-zero structure. For example, in a set s of $|s|$ elements, each element $s_i \in s$ might be described as explananda of two τ 's

which are distinguishable by their structure. The result is an alphabet set with two distinguishable states, $\{A_1, A_2\}$ that “label” or “categorize” each element.

Any element s_i of this labelled set i could be described with the same amount of information, either following a “bottom up” or a “top-down” description. Indicating with $A(i)$ the labelling of s_i :

$$D(s_i \in s) = \log |s| + D(A(i)) = \log \frac{|s|}{|A(i) \subset s|} + \log |A(i) \subset s| + D(A(i)) \quad (82)$$

where the second calculation is “bottom up”, in the sense that it counts each element in s , whereas the third is “top-down”, as it first identifies the categories A and then identifies each element within them.

- **A multiset** is thus a labelled set where the lower-level information about each instantiation is ignored, such that any element is described solely in terms of the partial information carried by its category:

$$D(s_i \in s) \rightarrow \log \frac{|s|}{|A(i) \subset s|} + D(A(i)) \equiv \log \frac{1}{p(A(i))} + D(A(i)) \quad (83)$$

This “lossy” description of a set is key to the generation of knowledge (see 6.4).

- **A random variable** is a draw from a multiset. Mathematically, it could be described as a graph with a single vertex, no relations, and an alphabet defined by the multiset. Graphically, we can represent a random variable as a symbol split into two or more elements which represent the alphabet.

A classic random variable carries information $-\log p(A_i)$, where $p(A_i)$ is calculated as in 79).

Understood as a τ , however, the actual, complete description length of a random variable is given by equation 83, which includes any nested “description” of the event A_i .

- **A joint random variable** seen as a τ results from a multiset with multiple nesting labellings. For example, if $Y_1 = srY_1$ is a second labelling of the set above, an element labelled A_1, Y_1 may be represented as:

and, indicating with $AY(i), A(i), Y(i)$ the joint and separate labelings of element s_i we have

$$\begin{aligned}
 D(s_i \in s) &= \log \frac{|s|}{|AY(i) \subset s|} + D(AY(i)) = \\
 &= \log \frac{1}{p(AY(i))} + D((s_i p A(i)) \otimes A(i) r Y(i)) = \\
 &= \log \frac{1}{p(Y(i)) p(A(i)|Y(i))} + [3 + D(Y(i)) + D(r)] + [3 + D(A(i)) + D(p) + D(s_i)] \rightarrow \\
 &= \log \frac{1}{p(Y(i)) p(A(i)|Y(i))} + [3 + D(Y(i)) + D(r) + 0] + [3 + D(A(i)) + D(p) + 0] = \\
 &= \log \frac{1}{p(Y(i))} + \log \frac{1}{p(A(i)|Y(i))} + D(Y(i)) + D(A(i)) \quad (84)
 \end{aligned}$$

because $3 + D(Y) + D(r) + 0$ is the description of the higher-level Y without the description of the individual instantiation s (the labels might generate some confusion, but must be remembered that labels in higher and lower-level descriptions are not/not necessarily indicating the same structure. They are just a device to indicate that there is a correspondence between the higher-level element and its nested description see section 6.1.4).

- **Independent random variables**, which we may indicate as variables joined by the tensor product $\otimes$ have, by the ordinary definition of independence, τ description:

$$D(\bigoplus_i^n X_i) = \sum_i^n \log \frac{1}{p(X = X(i))} + \sum D(X(i)) \quad (85)$$

where the lower-level description $D(X(i))$ of each label or category is taken only once. If independent instantiations of the same random variable are considered, then:

$$D(A \otimes A) = \log \frac{1}{p(A)p(A)} + D(A) = 2 \log \frac{1}{p(A)} + D(A) \quad (86)$$

- **A string of symbols** is a sequence of independent random variables, and the description of its τ is the sum of the individual probabilities and their descriptions. We can represent such an object as a multi-set of symbols nested into a symbol, e.g. τ

where the symbols are same or different based on features in their description, and the nesting symbol τ qualifies them as all pertaining to the same category.

The total information contained in this τ is the sum of the logarithm of the probabilities plus the sum of the different descriptors, taken only once. The probabilities are normally defined based on a multiset (or a theory about it), but they could also be derived from the string itself.

Applying 79 to a string of n symbols, from a multiset $s : \{m_1 s_1, m_2 s_2 \dots m_p s_p\}$ of size n , with $\sum m_i = n$, the total information is:

$$D(\tau) = 0 + \sum m_i \log \frac{n}{m_i} + \sum D(S_i) = n \sum \frac{m_i}{n} \log \frac{n}{m_i} + \sum D(S_i) = nH(S) + \sum D(S_i) \quad (87)$$

where $H(S)$ is the Shannon entropy. Thus, the explanandum and input shown in the original version of the K function 1 are descriptions of strings. Mathematically, the asymptotic probabilities underlying a description are derived from the string itself, as it approaches infinity, as implied in the Kolmogorov complexity shown in equation 3.

Therefore, our definition of knowledge does not require the Kolmogorov complexity of the string. Further, the object that we are describing is not an ordinary string, due to the addition of the nested descriptors of each symbol. This is why we shall posit that this object results from a particular operation called *cumulation*, defined by a cumulation operator (see section 6.3.1).

- **Trivially simple graphs**, consisting of multiple symbols that are either not related to each other or that are fully related to each other, and in which the labels of vertices and edges are either all identical, have description length zero:

$$D(g) = 0 \text{ if } g \text{ is fully connected or fully disconnected and unlabelled} \quad (88)$$

This is theoretically expected, since both types of graphs are intrinsically simple to describe, and they represent a τ that, at that level of description, carries no information.

- **Maximally complex graphs**, conversely, occur when the number of relations is exactly half the total number of possible relations, i.e. $r = d^a/2$ and when the labels of vertices and relations are all different, i.e. when $|T_V| = d, |T_E| = r$ (see Figure 2.3)
- **Sparse graphs**, i.e. graphs in which the number of vertices d^a is much larger than that of relations r , will be simpler to describe in proportion to their sparseness. The sparseness increases as either the number of vertices d or the maximum arity of the relations a increase, relative to the number of relations r . In the limit of $d^a \gg r$, and using Stirling's approximation, it can be shown that:

$$D \sim \left(\frac{1}{2} + d + ar\right) \log d \quad (89)$$

which indicates that, all else equal, $D()$ grows approximately linearly with an increase in either the arity a or the number of relations r (see Appendix 7), and supra-linearly with an increase in the vertices (concepts) d . Figure 2.3 illustrates the various properties discussed above.

6.3 Operations on τ

6.3.1 The cumulation operator

Information is generated by the encoding of differences. For differences to be encoded, the entities being compared must be classified as equivalent at one level and potentially different at another level. For example, results of repeated trials in which coloured balls are drawn from a jar are defined as equivalent at one level (e.g. all these balls are implicitly undistinguishable in shape, and they are all picked the same way from the same jar), but different at another (they vary in colour, which is a property that “nests” the ball, or is nested within it). The concept of a random variable (and generally that of probability, information and knowledge) emerges as the interplay of this assumption of identity at one level and difference at another.

In order to formalize this process, we defined the cumulation operator $\oplus$, which joins homologous τ 's in strings (see section 6.2.1), recording any new nested description that it didn't encounter before

in a vocabulary. For example, let us have a series of n objects, all described as the same τ at one level, but containing nested descriptions $\tau_1, \tau_2 \dots \tau_n$ that may or may not be similar, which entails that they may be grouped in equivalence categories w with multiplicities m_w , such that $\sum m_w = n$, then

$$D(\tau_1 \oplus \tau_2 \dots \tau_n) \equiv D(\bigoplus^n \tau) = \sum m_w [\log \frac{n}{m_w} + D(\tau_w)] = nH(W) + \sum D(\tau_w) \quad (90)$$

For example, let us have three τ , two with lower-level description “A” and one with “B”, then

where the latter is our graphical representation of a string, and it has total description length:

$$D(A \oplus A \oplus B) = 2 \times \log \frac{3}{2} + D(A) + 1 \times \log \frac{3}{1} + D(B) = 3 \times 0.918 + D(A) + D(B) \quad (91)$$

where $0.918 = H(W)$. Note that if and until the τ have identical nested description, then $H(W) = 0$ and so $D(\bigoplus^n \tau) = D(\tau)$.

This same operation can be extended to τ structures of any complexity that are nested into any number of levels, by applying the $\oplus$ operation element-by-element (tensor cell by tensor cell), level by level. The identity or difference between any two elements at a level of description is given by the identity or difference of the structure of the tensor describing that level.

Let $\tau_1, \tau_2 \dots \tau_n$ be τ with nested descriptions, then

$$D(\bigoplus^n \tau) = \log \frac{1}{p(\tau_1)} + \sum_{\tau_a \succ \tau_1} D(\bigoplus^n \tau_a) \text{ if } T_1 = T_2 \dots = T_n \neq \emptyset \quad (92)$$

$$D(\bigoplus^n \tau) = nH(W) + \sum D(\tau_w) \text{ otherwise} \quad (93)$$

where $T_i = T_j \iff d_i = d_j, r_i = r_j$ etc. means that the τ at that level of description come from the same universe set, which entails that they have the same description length and homologous structure, i.e. identical adjacency matrix and *analogous* distribution of labels (i.e. the labels may be formally different, but have the same distribution on the same graph structure).

For example let us cumulate two τ , ($Ar([BpD](\succ B)$) and $Ar([BCqD](\succ B)$), which are homologous at higher levels, but not in their description of B

then

$$\begin{aligned} D(Ar([BpD](\succ B) \oplus As([BCqD](\succ B))) &= \\ 3 + D(A \oplus A) + D(r \oplus s) + D(B \oplus B) &= \\ 3 + [0 + D(A)] + [2 \times 1 + D(r) + D(s)] + [2 \times 1 + D(BpD) + D(BCqD)] &= \\ 3 + 2 \times (1 + 1) + D(r) + D(s) + D(BpD) + D(BCqD) &= \\ 3 + 2 \times (1 + 1) + D(A) + D(r) + D(s) + [3 + 2D(B) + D(p) + 2D(D)] + [7.34 + D(C) + D(q)] &= \\ 17.34 + \sum_i D(i) & \quad (94) \end{aligned}$$

In general, it will be preferable to re-describe the two τ 's in order to maximize their similarity at higher levels, because this makes the total description more compact (increasing the K) and pin-points divergences more accurately. This may be achieved by nesting some components of the more complex structure, or, conversely, by making the simpler structure more complex. In particular, we can re-describe any two τ 's via a common adjacency tensor that has arity corresponding to the largest arity of the two, a number of vertices corresponding to the largest number of vertices plus the null vertex, such that $a = \max(a_1, a_2)$ and $d = \max(|V_1|, |V_2|) + 1$. Taking the example above: we can re-scale the nested description of B to a common graph with $a = 3, d = 4$, and get:

where $\emptyset$ represents a null vertex, which leads to:

$$\begin{aligned}
D(\tau) &\equiv D(Ar([BpD](\succ B) \oplus As([BCqD](\succ B))) = \\
&\quad 3 + D(A \oplus A) + D(r \oplus s) + D(B \oplus B) = \\
3 + [0 + D(A)] + [2 + D(r) + D(s)] + \{7.34 + [0 + D(B)] + [2 + D(\emptyset) + D(C)] + [2 + D(p) + D(q)] + [0 + D(D)]\} = \\
&\quad 3 + 2 \times (1 + 1 + 1) + 7.34 + \sum_i D(i) = D(Ar([BpD](\succ B) \oplus As([BCqD](\succ B))) + 2 \times 1 \quad (95)
\end{aligned}$$

which is longer by $n \times 1$ bits, where the 1 corresponds to the additional difference that the finer-grained description was able to pin-point. This difference corresponds to the divergence between the two τ 's, as we explain below.

6.3.2 Divergence between τ

Since for a string of two symbols $H(W)$ is either 1 or 0, for any element that is either different or identical, we generally have

$$D(\tau_1 \oplus \tau_2) = \sum_{\tau_{1i}=\tau_{2i}} \log \frac{1}{p(\tau_{1i})} + 2 \times d + \sum_{a(\succ)\tau_1 \oplus \tau_2} D(\tau_a) \quad (96)$$

where the first term on the right side captures the nested structures common to both τ 's, the third captures all differing lower-level descriptions, and the variable d quantifies the number of divergences, i.e. the number of vertices or relations that have different lower-level descriptions.

The calculation above presumes that $n = 2$, but we can define a normalized version of any cumulated τ by setting $n = 1$, regardless of the number of cumulated τ . We shall symbolize this process of normalization with a lower bar, i.e. $\underline{\oplus}^n$.

$$D(\tau_1 \underline{\oplus} \tau_2) = \sum_{\tau_{1i}=\tau_{2i}} \log \frac{1}{p(\tau_{1i})} + d + \sum_{a(\succ)\tau_1 \underline{\oplus} \tau_2} D(\tau_a) \quad (97)$$

From 97 follows that

$$D(\tau_1 \underline{\oplus} \tau_2) = D(\tau_1) + d + \sum_{a(\neq)\tau_1} D(\tau_a) = D(\tau_2) + d + \sum_{a(\neq)\tau_2} D(\tau_a) \quad (98)$$

and therefore we can define a metric of divergence between any two τ 's as

$$D(\tau_i | \tau_j) = D(\tau_1 \underline{\oplus} \tau_2) - D(\tau_j) = d + \sum_{a(\neq)\tau_j} D(\tau_a) \quad (99)$$

It can be shown that

- $D(g|h) \geq 0$
- $D(g|h) = 0 \iff g = h$

if we further assume that the term $\sum_{a \in \tau_j} D(\tau_a)$ is either 0 or a constant identical for all τ , then

- $d(g|h) = d(h|g)$
- $D(g|h) \leq D(g|e) + D(e|h)$

which makes this divergence a metric, unlike the well-known Kullback-Leibler divergence. The general case with m, n cumulated τ will be:

$$D(\bigoplus^n \tau | \bigoplus^m \tau) = D(\bigoplus^{n+m}) - D(\bigoplus^m \tau) = \sum_{\tau_i \in \tau} [H(W_{i,m+n}) - H(W_{i,m})] + \sum_{a \in \bigoplus^m \tau} D(\tau_a) \quad (100)$$

where $H(W_i)$ is the entropy of the string generated for each element i , i.e. each vertex and relation at each homologous level. Since when elements i are all identical then $H(W) = 0$, the summation adds up all and only the elements that are different between two or more of the cumulated τ , element by element, level by level.

6.3.3 Conditioning of τ

The concept of a conditioning of a τ is already implicitly defined, because all τ are understood as conditions required for some knowledge claim $K > 0$ [1]. Nonetheless, this is an important notion that is worth making explicit. In particular, we may want to consider the conditioning of τ itself or, equivalently, of its input or explanans. For example, a particular theory might only apply to a subclass of phenomena, relative to another theory that applies to another subclass.

The process is analogous to a selection of cases from a sequence of events. Let n be the total number of events. We can picture the process as a string of length n with $\tau_1, \tau_2, \dots, \tau_w$ distinct symbols, each to be conditioned to a number w subsets of events, $m_1, m_2, \dots, m_w$. Using equation 7 and 13 we get

$$K = \frac{n(H(W) + E[D(\tau_i)] - (H(W) + E[D(\tau'_i)]))}{n(H(W) + E[D(\tau_i)])} = \frac{E[D(\tau_i) - D(\tau'_i)]}{E[D(\tau_i)] + H(W)} \quad (101)$$

where $H(W) = \sum (m_i/n) \log(n/m_i)$ is the entropy generated by the different τ 's and $E[\]$ is the expected value. The K value of a τ conditioned on a particular τ_i would be

$$K = \frac{D(\tau_i) - D(\tau'_i)}{D(\tau_i) + c} \quad (102)$$

in which $c = \log \frac{n}{m_i}$ is a conditioning factor that “narrows down” the universe to a subset of all events (instances, repetitions, observations, phenomena, etc.) that could be made. Therefore, any conditioning can be defined as the information required to “select” relevant instances from a broader set of instances, which is the same logic by which the K function was derived in the first place (see [1]). A practical example of conditioning is given in section 4.2.1.

6.3.4 Conceptualization

Section 6.1.5 describes a process by which a τ can be associated and nested in a symbol *about* that τ . This operation, which we may call conceptualization, expands the dimensionality of whatever τ is being conceptualized, by incrementing its arity and adding a new label. The resulting description length increases by an amount related to the complexity of the τ being conceptualized.

For example, the τ :

has $D(\tau) = 3 + D(X) + D(r) + D(Y)$. The corresponding conceptualization entails

which has $D(XrY) = 7.34 + D(X) + D(Y) + D(r) + \log 4 = 6.34 + D(\tau)$ bits

This increase in description length is again expected theoretically, since the mere description length of an object should be smaller than the description of the object plus the *concept* of that same object.

6.4 Knowledge as information compression

It can be shown that knowledge occurs by encoding relations amongst higher-level of descriptions, discarding lower-level information.

For example, let us have, in the set s described in section 6.2.1 a second nesting description, with labels Y, Z . A pattern results from the coincidence of the two labels. For each individual element i , for example, we calculate the case where i has both A and Y , which we may indicate as YA_i . Applying equation 13 and including the information to distinguish each element i , the K value would be derived as:

$$K(Y_i; A_i\tau) = \frac{\log \frac{|s|}{|Y|} + \log |Y| - \left[\log \frac{|A|}{|YA|} + \log |YA| \right]}{\log \frac{|s|}{|Y|} + \log |Y| + D(Y) + \log \frac{|s|}{|A|} + D(A) + \log |A| - \log p(\tau)} = \frac{\log \frac{|s|}{|A|}}{\log \frac{|s|}{|Y|} + \log |Y| + D(Y) + \log \frac{|s|}{|A|} + \log |A| + D(A) - \log p(\tau)} \quad (103)$$

whereas if we omit from the description of each individual element i we have

$$K(Y; A\tau) = \frac{\log \frac{|s|}{|Y|} - \left[\log \frac{|A|}{|YA|} \right]}{\log \frac{|s|}{|Y|} + D(Y) + \log \frac{|s|}{|A|} + D(A) - \log p(\tau)} = \frac{\log \frac{|s|}{|A|} - \log \frac{|Y|}{|YA|}}{\log \frac{|s|}{|Y|} + D(Y) + \log \frac{|s|}{|A|} + D(A) - \log p(\tau)} \quad (104)$$

Even though the numerator has lost information, we can show that $K(Y; A\tau) > K(Y_i; A_i\tau)$, i.e. knowledge has increased because, using the recursive property of K in equation 18, we obtain the condition

$$K(Y; A\tau) > K(Y_i; A_i\tau) \iff K(Y; A\tau)(\log |Y| + \log |A|) > \log \frac{|Y|}{|YA|} \quad (105)$$

which is always satisfied.

Proof:

$$\begin{aligned}
K(Y; A\tau)(\log |Y| + \log |A|) &> \log \frac{|Y|}{|YA|} \Rightarrow \\
\frac{\log \frac{|s|}{|A|} - \log \frac{|Y|}{|YA|}}{\log \frac{|s|}{|Y|} + \log \frac{|s|}{|A|} + D(Y) + D(A) - \log p(\tau)} &> \frac{\log |Y| - \log |YA|}{\log |Y| + \log |A|} \Rightarrow \\
\frac{\log \frac{|s|}{|A|} - \log \frac{|Y|}{|YA|}}{\log \frac{|s|}{|Y|} + \log \frac{|s|}{|A|}} &> \frac{\log |Y| - \log |YA|}{\log |Y| + \log |A|} \Rightarrow \\
\frac{-[\log \frac{|s|}{|A|} - \log \frac{|Y|}{|YA|}]}{-[\log \frac{|s|}{|Y|} + \log \frac{|s|}{|A|}]} &> \frac{\log |Y| - \log |YA|}{\log |Y| + \log |A|} \Rightarrow \\
\frac{\log \frac{|A|}{|s|} + \log \frac{|Y|}{|YA|}}{\log \frac{|Y|}{|s|} + \log \frac{|A|}{|s|}} &> \frac{\log |Y| - \log |YA|}{\log |Y| + \log |A|} \Rightarrow \\
\frac{\log |Y| - \log |YA| + \log |A| - \log |s|}{\log |Y| + \log |A| - 2 \log |s|} &> \frac{\log |Y| - \log |YA|}{\log |Y| + \log |A|} \Rightarrow \\
\log |A| - \log |s| &> -2 \log |s| \Rightarrow \log |A| > -\log |s| \quad (106)
\end{aligned}$$

which is always true since $\log |A| \geq 0$ and $\log |s| > 0$.

Thus, the K value is maximised whenever, relative to a pattern described at a certain level of abstraction, all lower-level information is discarded. If K is a measure of knowledge, then this results states that we have more knowledge by appropriately discarding lower-level details.

For example, in the process of drawing coloured balls from a jar, we have more knowledge if we ignore the identity of each ball instead of retaining that information. This is a non-trivial result because it counters what might superficially seem more intuitive (that it is better, in principle, to have complete information).

We take this result as a corroboration of the second postulate of K theory, which states that knowledge is information compression. It also further confirms the first postulate, i.e. that information is finite, by showing that the compression attained is necessarily lossy (see also a proof in [1], section 2.3.6).

Knowledge is information compression because it presupposes the abstraction of properties from phenomena and the encoding of patterns amongst these. From a potentially infinite amount of information available, we acquire knowledge by “filtering out” coarse-grained categories, shedding all lower-level or higher-level distinctions amongst them, and encoding patterns amongst what remains. In order to “know” anything, we have to “condition out” the greatest part of a potentially infinite τ limiting our understanding to a small part of it.

6.5 The total K of a τ

Combined with the recursive quantification of τ (section 6.2, the recursive interpretation of K allows us to define what the total knowledge contained in a structure is, giving rigorous proof to the intuitive approach used in [1] to quantify the “total knowledge of a system” and produce a “multi-dimensional K ”.

We start by noticing that any K can be described by a potentially infinite series. For example, if we let $\Delta\tau = n_Y H(Y) - n_Y H(Y|X, \tau)$, $D(\tau) = n_Y H(Y) + n_X H(X) + \log p(\tau)$, such that $K = \Delta\tau/D(\tau)$ then:

$$\begin{aligned}
&\frac{\Delta\tau + 4 \frac{\Delta\tau + 3 \frac{\Delta\tau + 2 \frac{\Delta\tau + \frac{\Delta\tau}{D(\tau)}}{D(\tau)+1}}{D(\tau)+2}}{D(\tau)+3}}{D(\tau)+4} = K \quad (107)
\end{aligned}$$

Now, the same process can be applied to a set of $\Delta\tau$ and $D(\tau)$ that are not constant. In particular, let $\bar{\tau} = \tau_1 \textcircled{\curvearrowright} \tau_2 \textcircled{\curvearrowright} \tau_3 \dots$ be a nested τ of arbitrary depth and complexity, and let $\underline{\tau} = \tau \textcircled{\curvearrowleft} \dots \textcircled{\curvearrowleft} \tau_2 \textcircled{\curvearrowleft} \tau_1$ be the de-nested version, then we can calculate the total K of this τ as

$$K(\bar{\tau}) = \dots \frac{\Delta\tau_4 + \sum_1^3 D(\tau) \frac{\Delta\tau_2 + D(\tau_1) \frac{\Delta\tau_1}{D(\tau_1)}}{\sum_1^2 D(\tau_i)}}{\sum_1^4 D(\tau_i)} = \frac{\sum \Delta\tau_i}{\sum D(\tau_i)} = \frac{\sum \Delta\tau}{D(\bar{\tau})} \quad (108)$$

which shows that the total K results from the sum, at the numerator, of the number of bits yielded by all statements (all sub- τ 's) contained in the nested τ and, at the denominator, the sum of the descriptions of the individual nested τ 's, which by definition 7 is the description of the total τ .

Further, if and to the extent that the nested τ are different from each other, such that $D(\underline{\tau}) \geq D(\bar{\tau})$ then,

$$K(\bar{\tau}) \geq \frac{\sum \Delta\tau}{D(\underline{\tau})} = K(\underline{\tau}) \quad (109)$$

showing that the K calculated from the nested version of the τ is an upper bound of the one calculated with the de-nested version. The latter, if calculated completely, would constitute the true, exact calculation of the total knowledge expressed by a given τ .

7 APPENDIX A1

We apply Stirling's approximation:

$$\log_2 n! \sim \log_2 \sqrt{2\pi n} \left(\frac{n}{e}\right)^n = \log_2 \sqrt{2\pi} + \frac{1}{2} \log_2 n + n \log_2 \frac{n}{e} \equiv c + \frac{1}{2} \log n + n \log \frac{n}{e} \quad (110)$$

to equation 73 as follows:

$$\begin{aligned} \log \frac{1}{p(g)} &= \log \frac{d^a! \times d! \times r!}{r!(d^a - r)! \times \prod_{j=1}^p m_{vj}! \times \prod_{i=1}^q m_{ei}! \times \prod_{i=1}^q r(e_i)!} \\ &= \log d^a! + \log d! + \log r! - \log r! - \log(d^a - r)! - \sum_{j=1}^p \log m_{vj}! - \sum_{i=1}^q \log m_{ei}! - \sum_{i=1}^q \log r(e_i)! \\ &= \log d^a! + \log d! - \log(d^a - r)! - \sum_{j=1}^p \log m_{vj}! - \sum_{i=1}^q \log m_{ei}! - \sum_{i=1}^q \log r(e_i)! \quad (111) \end{aligned}$$

Momentarily omitting the elements representing symmetries in labelling, we get

$$\begin{aligned} \log d^a! + \log d! - \log(d^a - r)! &\sim \\ c + \frac{1}{2} \log d^a + d^a \log \frac{d^a}{e} + c + \frac{1}{2} \log d + d \log \frac{d}{e} - c - \frac{1}{2} \log(d^a - r) - (d^a - r) \log \frac{(d^a - r)}{e} \\ &= c + \frac{1}{2} \log d^a + d^a \log \frac{d^a}{e} + \frac{1}{2} \log d + d \log \frac{d}{e} - \frac{1}{2} \log(d^a - r) - d^a \log \frac{(d^a - r)}{e} + r \log \frac{(d^a - r)}{e} \\ &= c + \frac{1}{2} \log \frac{d^{a+1}}{d^a - r} + d^a \log \frac{d^a}{d^a - r} + d \log d + r \log \frac{d^a - r}{e} \quad (112) \end{aligned}$$

which for sparse graphs, i.e. $d^a \gg r$, approaches the limit

$$c + 1 + (d + \frac{1}{2}) \log d + r \log \frac{d^a}{e} \approx c' + (\frac{1}{2} + d + ar) \log d \quad (113)$$

Therefore, for a given number of nodes (concepts) d , the complexity grows linearly with *either* an increase in the number of relations of their maximal arity, in proportion to the magnitude of the other. Each additional relation increases the complexity by $a \log d$, and each additional dimension (arity) increases it by $r \log d$.

The labeling of either vertices or edges has a marginally modulating effect on this result, since it will have longer description length when the multiplicities are equally distributed, in other words where $m_{vj} = d/p, \forall j$ and $m_{ei} = r/q, \forall i$, with a maximum when $p = 1, q = 1$. Using again Stirling's approximation:

$$\sum \log m_{vj}! \approx p \left(c + \frac{1}{2} \log \frac{d}{p} + \frac{d}{p} \log \frac{d}{pe} \right) \quad (114)$$

which has maximal value $(c + 1/2 + d) \log d$. Analogous calculations for edges give $\sum \log m_{ei}! \approx (c + 1/2 + r) \log r$. Therefore, in the maximal complexity scenario of all distinctive labels, equation 113 reduces to:

$$c' + (-c + ar) \log d - (c + \frac{1}{2} + r) \log r \quad (115)$$

which still grows with either an increase in a or r , albeit more slowly in the latter case.

References

- [1] Daniele Fanelli. A theory and methodology to quantify knowledge. *Royal Society Open Science*, 6(4), 2019.
- [2] M Li and P Vitányi. *An Introduction to Kolmogorov Complexity and Its Applications*. Springer New York, 2009.
- [3] Hector Zenil, Narsis A. Kiani, and Jesper Tegnér. A review of graph and network complexity from an algorithmic information perspective, 2018.
- [4] M. Gell-Mann and S. Lloyd. Information measures, effective complexity, and total information. *Complexity*, 2:44–52, 1996.
- [5] Fridolin Gross. Between mechanical clocks and emergent flocks: complexities in biology. *Synthese*, 2021.
- [6] John F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks/Cole, 2000, 4th edition, 2000.
- [7] Marie-Laure Chein, Michel; Mugnier. *Graph-based knowledge representation*. Springer-Verlag, London, 2009.
- [8] Michel Chein, Marie Laure Mugnier, and Madalina Croitoru. Visual reasoning with graph-based mechanisms: The good, the better and the best. *Knowledge Engineering Review*, 28(3):249–271, 2013.
- [9] Peter Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.
- [10] Peter Grünwald and Teemu Roos. Minimum description length revisited. *International Journal of Mathematics for Industry*, 11(1):1930001–1–1930001–29, 12 2019.
- [11] Luis Enrique Sucar. *Probabilistic Graphical Models: Principles and Applications*. Springer Publishing Company, Incorporated, 2015.
- [12] Salem Benferhat, Philippe Leray, and Karim Tabia. Belief Graphical Models for Uncertainty Representation and Reasoning. In Henri Prade Pierre Marquis, Odile Papini, editor, *A Guided Tour of Artificial Intelligence Research*, pages 209–246. Springer, Cham, 2020.

- [13] J Pearl. *Causality*. Cambridge University Press, 2009.
- [14] Aly El Gamal. Learning graph grammars. Technical report, 2015.
- [15] Wilfried; Klavžar Sandi Hammack, Richard; Imrich. *Handbook of Product Graphs*. Routledge, 2nd edition, 2011.
- [16] Alain Bretto. *Hypergraph Theory*. Springer, 2013.
- [17] Tasha Fairfield and Andrew E. Charman. Explicit bayesian analysis for process tracing: Guidelines, opportunities, and caveats. *Political Analysis*, 25(3):363–380, 2017.
- [18] H Akaike. Information theory and an extension of the maximum likelihood principle. *2nd international symposium on information theory. Abstracts of papers*, page 276, 1971.
- [19] Kenneth P. Burnham and David R. Anderson. *Model selection and multimodel inference - 2nd edition*. Springer Science + Business Media New York, New York, second edition, 2002.
- [20] Wes Bonifay and Li Cai. On the complexity of item response theory models. *Multivariate behavioral research*, 52:465–484, 2017.
- [21] Andrew Gelman and Eric Loken. The statistical Crisis in science. *American Scientist*, 102(6):460–465, 2014.
- [22] Ronald L. Wasserstein and Nicole A. Lazar. The asa statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2):129–133, 2016.
- [23] Jouni Kuha. AIC and BIC: Comparisons of assumptions and performance, 2004.
- [24] Marco Del Giudice and Steven W. Gangestad. A traveler’s guide to the multiverse: Promises, pitfalls, and a framework for the evaluation of analytic decisions. *Advances in Methods and Practices in Psychological Science*, 4(1):2515245920954925, 2021.
- [25] Mark Rubin. Does preregistration improve the credibility of research findings?, 2020.
- [26] P Thagard. *Computational Philosophy of Science*. A Bradford, 1993.
- [27] Roland M. Mueller. A meta-model for inferring inter-theory relationships of causal theories. In *Proceedings of the Annual Hawaii International Conference on System Sciences*, volume 2015-March, pages 4908–4917, 2015.
- [28] Thomas S Kuhn. *The structure of scientific revolutions*. The University of chicago Press, Chicago, second edition, 1970.
- [29] T M Cover and J A Thomas. *Elements of Information Theory*. Wiley, 2012.
- [30] Daniele Fanelli. Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences of the United States of America*, 115(11):2628–2631, 2018.
- [31] R A Klein, K A Ratliff, M Vianello, R B Adams, S Bahník, M J Bernstein, K Bocian, M J Brandt, B Brooks, C C Brumbaugh, Z Cemalcilar, J Chandler, W Cheong, W E Davis, T Devos, M Eisner, N Frankowska, D Furrow, E M Galliani, F Hasselman, J A Hicks, J F Hovermale, S J Hunt, J R Huntsinger, H Ijzerman, M S John, J A Joy-Gaba, H B Kappes, L E Krueger, J Kurtz, C A Levitan, R K Mallett, W L Morris, A J Nelson, J A Nier, G Packard, R Pilati, A M Rutchick, K Schmidt, J L Skorinko, R Smith, T G Steiner, J Storbeck, L M Van Swol, D Thompson, A E van ’t Veer, L A Vaughn, M Vranka, A L Wichman, J A Woodzicka, and B A Nosek. Investigating Variation in Replicability A ”Many Labs” Replication Project. *Social Psychology*, 45(3):142–152, 2014.

- [32] A A Aarts, J E Anderson, C J Anderson, P R Attridge, A Attwood, J Axt, M Babel, S Bahník, E Baranski, M Barnett-Cowan, E Bartmess, J Beer, R Bell, H Bentley, L Beyan, G Binion, D Borsboom, A Bosch, F A Bosco, S D Bowman, M J Brandt, E Braswell, H Brohmer, B T Brown, K Brown, J Bruning, A Calhoun-Sauls, S P Callahan, E Chagnon, J Chandler, C R Chartier, F Cheung, C D Christopherson, L Cillessen, R Clay, H Cleary, M D Cloud, M Cohn, J Cohoon, S Columbus, A Cordes, G Costantini, L D C Alvarez, E Cremata, J Crusius, J DeCoster, M A DeGaetano, N Della Penna, B den Bezemer, M K Deserno, O Devitt, L Dewitte, D G Dobolyi, G T Dodson, M B Donnellan, R Donohue, R A Dore, A Dorrough, A Dreber, M Dugas, E W Dunn, K Easey, S Eboigbe, C Eggleston, J Embley, S Epskamp, T M Errington, V Estel, F J Farach, J Feather, A Fedor, B Fernandez-Castilla, S Fiedler, J G Field, S A Fitneva, T Flagan, A L Forest, E Forsell, J D Foster, M C Frank, R S Frazier, H Fuchs, P Gable, J Galak, E M Galliani, A Gampa, S Garcia, D Gazarian, E Gilbert, R Giner-Sorolla, A Glockner, L Goellner, J X Goh, R Goldberg, P T Goodbourn, S Gordon-McKeon, B Gorges, J Gorges, J Goss, J Graham, J A Grange, J Gray, C Hartgerink, J Hartshorne, F Hasselman, T Hayes, E Heikensten, F Henninger, J Hodsoll, T Holubar, G Hoogendoorn, D J Humphries, C O Y Hung, N Immelman, V C Irsik, G Jahn, F Jakel, M Jekel, M Johannesson, L G Johnson, D J Johnson, K M Johnson, W J Johnston, K Jonas, J A Joy-Gaba, H B Kappes, K Kelso, M C Kidwell, S K Kim, M Kirkhart, B Kleinberg, G Knezevic, F M Kolorz, J J Kossakowski, R W Krause, J Krijnen, T Kuhlmann, Y K Kunkels, M M Kyc, C K Lai, A Laique, D Lakens, K A Lane, B Lassetter, L B Lazarevic, E P LeBel, K J Lee, M Lee, K Lemm, C A Levitan, M Lewis, L Lin, S Lin, M Lippold, D Loureiro, I Luteijn, S Mackinnon, H N Mainard, D C Marigold, D P Martin, T Martinez, E J Masicampo, J Matacotta, M Mathur, M May, N Mechin, P Mehta, J Meixner, A Melinger, J K Miller, M Miller, K Moore, M Moschl, M Motyl, S M Muller, M Munafo, K I Neijenhuijs, T Nervi, G Nicolas, G Nilsonne, B A Nosek, M B Nuijten, C Olsson, C Osborne, L Ostkamp, M Pavel, I S Penton-Voak, O Perna, C Pernet, M Perugini, R N Pipitone, M Pitts, F Plessow, J M Prenoveau, R M Rahal, K A Ratliff, D Reinhard, F Renkewitz, A A Ricker, A Rigney, A M Rivers, M Roebke, A M Rutchick, R S Ryan, O Sahin, A Saide, G M Sandstrom, D Santos, R Saxe, R Schlegelmilch, K Schmidt, S Scholz, L Seibel, D F Selterman, S Shaki, W B Simpson, H C Sinclair, J L M Skorinko, A Slowik, J S Snyder, C Soderberg, C Sonnleitner, N Spencer, J R Spies, S Steegen, S Stieger, N Strohminger, G B Sullivan, T Talhelm, M Tapia, A te Dorsthorst, M Thomae, S L Thomas, P Tio, F Traets, S Tsang, F Tuerlinckx, P Turchan, M Valasek, A E van 't Veer, R Van Aert, M van Assen, R van Bork, M van de Ven, D van den Bergh, M van der Hulst, R van Dooren, J van Doorn, D R van Renswoude, H van Rijn, W Vanpaemel, A V Echeverria, M Vazquez, N Velez, M Vermue, M Verschoor, M Vianello, M Voracek, G Vuu, E J Wagenmakers, J Weerdmeester, A Welsh, E C Westgate, J Wissink, M Wood, A Woods, E Wright, S Wu, M Zeelenberg, K Zuni, Collaboration Open Sci, and Open Science Collaboration. Estimating the reproducibility of psychological science. *SCIENCE*, 349(6251):8, 2015.
- [33] C F Camerer, A Dreber, E Forsell, T H Ho, J Huber, M Johannesson, M Kirchler, J Almenberg, A Altmejd, T Z Chan, E Heikensten, F Holzmeister, T Imai, S Isaksson, G Nave, T Pfeiffer, M Razen, and H Wu. Evaluating replicability of laboratory experiments in economics. *SCIENCE*, 351(6280):1433–1436, 2016.
- [34] Colin F Camerer, Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Gideon Nave, Brian A Nosek, Thomas Pfeiffer, Adam Altmejd, Nick Buttrick, Taizan Chan, Yiling Chen, Eskil Forsell, Anup Gampa, Emma Heikensten, Lily Hummer, Taisuke Imai, Siri Isaksson, Dylan Manfredi, Julia Rose, Eric-Jan Wagenmakers, and Hang Wu. Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2018.
- [35] Florian Cova, Brent Strickland, Angela Abatista, Aurélien Allard, James Andow, Mario Attie, James Beebe, Renatas Berniunas, Jordane Boudesseul, Matteo Colombo, Fiery Cushman, Rodrigo Diaz, Noah N'Djaye Nikolai van Dongen, Vilius Dranseika, Brian D. Earp, Antonio Gaitán Torres, Ivar Hannikainen, José V. Hernández-Conde, Wenjia Hu, François Jaquet, Kareem Khalifa, Hanna Kim, Markus Kneer, Joshua Knobe, Miklos Kurthy, Anthony Lantian, Shen-yi Liao,

Edouard Machery, Tania Moerenhout, Christian Mott, Mark Phelan, Jonathan Phillips, Navin Rambharose, Kevin Reuter, Felipe Romero, Paulo Sousa, Jan Sprenger, Emile Thalabard, Kevin Tobia, Hugo Viciano, Daniel Wilkenfeld, and Xiang Zhou. Estimating the Reproducibility of Experimental Philosophy. *Review of Philosophy and Psychology*, 2018.

- [36] Brian A. Nosek and Timothy M. Errington. Making sense of replications. *eLife*, 6, 2017.
- [37] Richard A. Klein, Michelangelo Vianello, Fred Hasselman, Byron G. Adams, Reginald B. Adams, Sinan Alper, Mark Aveyard, Jordan R. Axt, Mayowa T. Babalola, Štěpán Bahník, Rishtee Batra, Mihály Berkics, Michael J. Bernstein, Daniel R. Berry, Olga Bialobrzeska, Evans Dami Binan, Konrad Bocian, Mark J. Brandt, Robert Busching, Anna Cabak Rédei, Huajian Cai, Fanny Cambier, Katarzyna Cantarero, Cheryl L. Carmichael, Francisco Ceric, Jesse Chandler, Jen-Ho Chang, Armand Chatard, Eva E. Chen, Winnee Cheong, David C. Cicero, Sharon Coen, Jennifer A. Coleman, Brian Collisson, Morgan A. Conway, Katherine S. Corker, Paul G. Curran, Fiery Cushman, Zubairu K. Dagona, Ilker Dalgar, Anna Dalla Rosa, William E. Davis, Maaike de Bruijn, Leander De Schutter, Thierry Devos, Marieke de Vries, Canay Dogulu, Nerisa Dozo, Kristin Nicole Dukes, Yarrow Dunham, Kevin Durrheim, Charles R. Ebersole, John E. Edlund, Anja Eller, Alexander Scott English, Carolyn Finck, Natalia Frankowska, Miguel-Ángel Freyre, Mike Friedman, Elisa Maria Galliani, Joshua C. Gandi, Tanuka Ghoshal, Steffen R. Giessner, Tripat Gill, Timo Gnambs, Ángel Gómez, Roberto González, Jesse Graham, Jon E. Grahe, Ivan Grahek, Eva G. T. Green, Kakul Hai, Matthew Haigh, Elizabeth L. Haines, Michael P. Hall, Marie E. Heffernan, Joshua A. Hicks, Petr Houdek, Jeffrey R. Huntsinger, Ho Phi Huynh, Hans IJzerman, Yoel Inbar, Åse H. Innes-Ker, William Jiménez-Leal, Melissa-Sue John, Jennifer A. Joy-Gaba, Roza G. Kamilolu, Heather Barry Kappes, Serdar Karabati, Haruna Karick, Victor N. Keller, Anna Kende, Nicolas Kervyn, Goran Knežević, Carrie Kovacs, Lacy E. Krueger, German Kurapov, Jamie Kurtz, Daniël Lakens, Ljiljana B. Lazarević, Carmel A. Levitan, Neil A. Lewis, Samuel Lins, Nikolette P. Lipsey, Joy E. Losee, Esther Maassen, Angela T. Maitner, Winfrida Malingumu, Robyn K. Mallett, Satia A. Marotta, Janko Mededović, Fernando Mena-Pacheco, Taciano L. Milfont, Wendy L. Morris, Sean C. Murphy, Andriy Myachykov, Nick Neave, Koen Neijenhuijs, Anthony J. Nelson, Félix Neto, Austin Lee Nichols, Aaron Ocampo, Susan L. O'Donnell, Haruka Oikawa, Masanori Oikawa, Elsie Ong, Gábor Orosz, Malgorzata Osowiecka, Grant Packard, Rolando Pérez-Sánchez, Boban Petrović, Ronaldo Pilati, Brad Pinter, Lysandra Podesta, Gabrielle Pogge, Monique M. H. Pollmann, Abraham M. Rutchick, Patricio Saavedra, Alexander K. Saeri, Erika Salomon, Kathleen Schmidt, Felix D. Schönbrodt, Maciej B. Sekerdej, David Sirlopú, Jeanine L. M. Skorinko, Michael A. Smith, Vanessa Smith-Castro, Karin C. H. J. Smolders, Agata Sobkow, Walter Sowden, Philipp Spachtholz, Manini Srivastava, Troy G. Steiner, Jeroen Stouten, Chris N. H. Street, Oskar K. Sundfelt, Stephanie Szeto, Ewa Szumowska, Andrew C. W. Tang, Norbert Tanzer, Morgan J. Tear, Jordan Theriault, Manuela Thomae, David Torres, Jakub Traczyk, Joshua M. Tybur, Adrienn Ujhelyi, Robbie C. M. van Aert, Marcel A. L. M. van Assen, Marije van der Hulst, Paul A. M. van Lange, Anna Elisabeth van 't Veer, Alejandro Vásquez- Echeverría, Leigh Ann Vaughn, Alexandra Vázquez, Luis Diego Vega, Catherine Verniers, Mark Verschoor, Ingrid P. J. Voermans, Marek A. Vranka, Cheryl Welch, Aaron L. Wichman, Lisa A. Williams, Michael Wood, Julie A. Woodzicka, Marta K. Wronska, Liane Young, John M. Zelenski, Zeng Zhijia, and Brian A. Nosek. Many Labs 2: Investigating Variation in Replicability Across Samples and Settings. *Advances in Methods and Practices in Psychological Science*, 1(4):443–490, 2018.
- [38] Andrew Gelman. The Connection Between Varying Treatment Effects and the Crisis of Unreplicable Research: A Bayesian Perspective. *Journal of Management*, 41(2):632–643, 2015.
- [39] Andrew Gelman. Don't characterize replications as successes or failures, 2018.
- [40] Brent M. Wilson, Christine R. Harris, and John T. Wixted. Science is not a signal detection problem, 2020.

- [41] J J Van Bavel, P Mende-Siedlecki, W J Brady, and D A Reinero. Contextual sensitivity in scientific reproducibility. *Proceedings of the National Academy of Sciences of the United States of America*, 113(23):6454–6459, 2016.
- [42] B Voelkl and H Wurzel. Reproducibility Crisis: Are We Ignoring Reaction Norms? *Trends in Pharmacological Sciences*, 37(7):509–510, 2016.
- [43] Michael Ramscar. Learning and the replicability of priming effects. *Current Opinion in Psychology*, 12(12):5, 2016.
- [44] Jeremy Freese and David Peterson. Replication in Social Science. *Annual Review of Sociology*, 43(1):147–165, 2017.
- [45] Bart Penders, J. Britt Holbrook, and Sarah de Rijcke. Rinse and repeat: Understanding the value of replication across different ways of knowing. *Publications*, 7(3), 2019.
- [46] Prakash Gorroochurn, Susan E. Hodge, Gary A. Heiman, Martina Durner, and David A. Greenberg. Non-replication of association studies: "Pseudo-failures" to replicate?, 2007.
- [47] Jeff Miller. What is the probability of replicating a statistically significant effect?, 2009.
- [48] David J. Stanley and Jeffrey R. Spence. Expectations for Replications: Are Yours Realistic? *Perspectives on Psychological Science*, 9(3):305–318, 2014.
- [49] Prasad Patil, Roger D. D Peng, and Jeffrey T. T Leek. What Should Researchers Expect When They Replicate Studies? A Statistical View of Replicability in Psychological Science. *Perspectives on Psychological Science*, 11(4):539–544, 2016.
- [50] A Etz and J Vandekerckhove. A Bayesian Perspective on the Reproducibility Project: Psychology. *PLoS ONE*, 11(2):12, 2016.
- [51] Ylona Chun Tie, Melanie Birks, and Karen Francis. Grounded theory research: A design framework for novice researchers. *SAGE Open Medicine*, 7, 1 2019.
- [52] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3:77–101, 2006.
- [53] David Waldner. Process Tracing and Qualitative Causal Inference. *Security Studies*, 24(2):239–250, 2015.
- [54] Jennifer J. Pokorny, Alex Norman, Anthony P. Zanesco, Susan Bauer-Wu, Baljinder K. Sahdra, and Clifford D. Saron. Network analysis for the visualization and analysis of qualitative data. *Psychological Methods*, 23(1):169–183, 2018.
- [55] Damien Contandriopoulos, Catherine Larouche, Mylaine Breton, and Astrid Brousselle. A sociogram is worth a thousand words: proposing a method for the visual analysis of narrative data. *Qualitative Research*, 18:70–87, 2018.
- [56] Richard Swedberg. Can you visualize theory? on the use of visual thinking in theory pictures, theorizing diagrams, and visual sketches. *Sociological Theory*, 34:250–275, 2016.
- [57] David Collier. Understanding process tracing. *Political Science and Politics*, 44(4):823–830, 2011.
- [58] Yaneer Bar-Yam. Multiscale complexity/entropy. *Advances in Complex Systems*, 7:47–63, 2004.
- [59] Karl Friston. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11:127 EP –, 01 2010.
- [60] Juergen Schmidhuber. Formal Theory of Creativity, Fun, and Intrinsic Motivation (1990-2010). *IEEE Transactions on Autonomous mental Development*, 2(3):230–247, 2010.

- [61] M.M. Hurley, D.C. Dennett, and R.B. Adams. *Inside Jokes: Using Humor to Reverse-engineer the Mind*. MIT Press, Cambridge, MA, 2011.
- [62] Phil Maguire, Rebecca Maguire, and Philippe Moser. Understanding consciousness as data compression. *Journal of Cognitive Science*, 17:63–94, 03 2016.
- [63] Giulio Ruffini. An algorithmic information theory of consciousness. *Neuroscience of Consciousness*, 2017(1), 10 2017. nix019.