Information Epistemology

Abstract

Information is a more basic epistemological concept than probability. It can be used to create the foundations for probability, and it delivers substantial epistemological results that cannot be obtained by using probability theory without the use of information theory. The paper shows in what ways information theory and probability theory are equivalent, and in what ways information theory is epistemologically prior to probability theory. A measure-theoretic proof is provided that Bayesian updating and the principle of minimal discrimination (using the Kullback-Leibler Divergence) are compatible. One section is devoted to Kolmogorov Complexity, Chaitin's incompleteness theorem, and their implications for an epistemology based on information theory.

1. Information and Probability

Epistemology in general and belief revision in particular, or conditionalization of belief on evidence, can have different starting points. Timothy Williamson, for example, has made a good case for using knowledge itself as an epistemologically primary and indivisible concept [48]. For Bayesians, probability is the central concept. Given mutually exclusive and exhaustive hypotheses $\{H_1, H_2, \ldots, H_n\}$ and a new piece of evidence E, it is 'rational' to revise the probabilistic weight given to the hypotheses by the following formula or otherwise be liable to a Dutch book (Teller [43]):

$$Q(H_i) = P(H_i|E) = \frac{P(E|H_i)P(H_i)}{\sum_{k=1}^{n} P(E|H_k)} \tag{1}$$

Q is the updated probability assignment after evidence E comes in (the 'posterior probability distribution'), P is the 'prior probability distribution,' before we know about the evidence.

Information is a serious alternative to probability as a fundamental notion of epistemology. There is a tight mathematical connection between information and probability, so that the question which of them is primary is in some ways a war of words because they are sides of the same coin. In other ways, however, where belief revision is not quantifiable, considerations of information may be more helpful in determining rational, epistemically justifiable, or just pragmatically workable belief. There is some evidence that our brains are successful this way: compression of data may be related to intelligence (Friston [15]), whereas it is unclear how good we are with our probability assessments.

Probability theory gives us relatively little guidance about the prior probability distribution. For a finite, discrete set of hypotheses $(H_i)_{i=1,\ldots,n}$ an application of Laplace's principle of indifference will often do the job: $P(H_i) = n^{-1}$. It is less certain what to do in the infinite case, in the continuous case, or when Bertrand's paradox strikes [3].[1] The principle of maximum entropy, however, developed by E. T. Jaynes, generalizes Laplace's principle of indifference to the continuous case, where the normal distribution is the one that is maximally noncommittal with regard to missing information [25, 622f]. Bertrand's paradox can be addressed by investigating parameter transformations. The constraints we pose on prior probability distributions generally originate in information theory. Entropy is just a way of measuring the reciprocal of information. For a detailed and concise proof that the normal distribution contains the largest amount of uncertainty see Kampé de Fériet [27], in English Guiaşu [20, 298ff]; for a proof of similar significance for the Poisson distribution with respect to information theory see Ingarden and Kossakowski [23].

Most of information theory is based on probability. Aleksandr Khinchin calls information theory an "important and interesting chapter of the general theory of probability" [29, 1]. One of the stock formulas of information theory is Shannon's Information Entropy and defines the entropy of a probability distribution where the x_i are the elements of a finite event space with cardinality n:

$$H(X) = -\sum_{i=1}^{n} p(x_i) \log p(x_i) \tag{2}$$

Discontent about this dependence of information theory on probability theory rises early, especially voiced by Andrey Kolmogorov. We will have a brief look at Ingarden and Urbanik [24], Kolmogorov [31], and Kampé de Fériet and Forte [28] for formal attempts to define information theory without recourse to probability theory that entails probability. First, here is a more intuitive approach using counterfactuals.

If we had a measure on the set of worlds (we can come up with a stylized set of worlds where we have such a measure), it would be intuitive to say that the more a proposition constrains which possible worlds are consistent with the actual world, the more information it gives us. Let X be a proposition.

Then

$$I^*(X) = \frac{\mu(\text{set of all possible worlds})}{\mu(\text{set of all possible worlds in which } X \text{ is true})} \tag{3}$$

gives us a first approximation how we might measure information.

We rewrite this definition for notational convenience:

$$I^*(X) = \frac{\mu(\Omega)}{\mu(\Omega|X)}$$

I^* is on an awkward scale, does not accord with the bit (binary digit) of information theory, and needs to be normalized. Call $I(X)$ the information density of the proposition X:

$$I(X) = \log_2(I^*(X)) \tag{4}$$

One eighth of possible worlds is consistent with the actual world if it is constrained by the result of three independent tosses of a fair coin. If X is this result, $I(X) = 3$, which corresponds nicely to the fact that X contains 3 bits of information.

In a stylized set of a finite number of possible worlds, it seems intuitive to say that the information in X and Y is unrelated if

$$\mu(\Omega|X \wedge Y) = \frac{\mu(\Omega|X) \cdot \mu(\Omega|Y)}{\mu(\Omega)}$$

Consequently,

$$I^*(X \wedge Y) = I^*(X) \cdot I^*(Y)$$

and so, using (4),

$$I(X \wedge Y) = I(X) + I(Y)$$

The joint information density of unrelated information is additive, and to whatever degree the information is redundant it is less additive.

Information density is intuitively related to probability. The relation must be intuitive, as we have no measure of possible worlds except in stylized mathematical forms. The first of many analogies to probability is that probability inclines to a similar stylization of possible worlds. In fact, the same intuitions as above give us:

$$P(X) = \frac{\mu(\Omega|X)}{\mu(\Omega)}$$

which translates into:

$$P(X) = 2^{-I(X)}$$

Let us consider information density our primary epistemological notion so that information density is numerically defined in mathematically stylized scenarios (as probabilities are often numerically defined in mathematically stylized scenarios) and otherwise defined (let's say, 'epistemologically defined') if no mathematical stylization is accessible.

Rationality constraints on the epistemological definition of information density are analogous to rationality constraints on subjective probabilities, following Kolmogorov's axioms of probability theory. The information density of a proposition X should in analogy to Kolmogorov's axioms (i) be between 0 and ∞ ($0 \leq I(X) \leq \infty$), (ii) be 0 if X yields no information, ∞ if X constrains the set of possible worlds to a set measuring 0, whatever that may mean in context, and (iii) be countably additive with other propositions $X', X'', \ldots$ if all those propositions together with X are pairwise unrelated to each other in terms of yielding information ($I(X \wedge X' \wedge X'' \wedge \ldots) = \sum_m I(X^m)$). For mostly equivalent axiomatizations of information theory in the literature see Shannon [40]; Fadeev [12]; Khinchin [29].

Once information density is axiomatized in this way, our first strike is to define probability in terms of information density[2] (in accordance with our intuitions):

$$P(X) = 2^{-I(X)} \tag{5}$$

It follows immediately that P is a probability measure and two propositions X and Y are independent if and only if $P(X \wedge Y) = P(X)P(Y)$.

Strike number two: knowledge and evidence are both co-extensive with the set of propositions whose information density is 0. $E = K$, as in Timothy Williamson's *The Limits of Knowledge* [48]. Strike number three, inspired by Jaynes's principle of maximum entropy, but now also applied to posterior distributions: if choosing from an alternative hypothesis set $\{H_1, H_2, \ldots\}$, the hypothesis H^* commanding the highest amount of epistemic warrant is the one (not necessarily unique) for which

$$I(\mathcal{E} \wedge H^*) = \min(\{I(\mathcal{E} \wedge H_1), I(\mathcal{E} \wedge H_2), \ldots\}) \tag{6}$$

where $\mathcal{E}$ is one's body of evidence.

The rest of the paper is a mopping up operation thinking about what (6) might mean as a definition and in scientific practice. It states that the hypothesis, explanation, or model is the best which is minimally informative. Knowledge is those propositions which provide no new information in context, as in the rejoinder, 'I already knew that.' If the proposition is false or a Gettier proposition [16], the rejoinder is inappropriate (although the speaker may not recognize it). Thus knowledge depends on both internal and external factors and the epistmological debate can continue as we are used to it.

Probabilistic language provides us with a rich intuitive inventory that is further backed up by rich mathematical notions where (in rare cases, such as coin-flipping and dice-rolling) the probabilities are quantitatively precise. The same is true for information: where the assignment of quantitative information is not available we still have good intuitions about what is informative and what is not. Sometimes, probability assignments are infeasible, maybe even on principle, while information density assignments are not.[3] Operating in the background and supportive of our intuitions, there is a substantial body of mathematics where the information density is numerically defined. This body of mathematics, information theory, differs from probability theory in interesting ways and yields substantial results not achievable by probability theory, although, as we shall see, there are important equivalence relations between information theory and probability theory. In particular, I am providing a measure-theoretic proof in the next section that Bayes' theorem follows from a purely information-theoretic approach: the probability distribution given by Bayesian updating is the unique minimally informative posterior probability distribution.

There is some recognition in the literature (the chief contender is Kolmogorov) that information is a concept more basic than probability.[4] There is no debate in information theory analogous to the debate about subjective, objective, and frequentist views of probability. Instead there are well-established theorems about what information theory *cannot* do for us, for example provide an algorithm for computing the complexity of information. Using the proof that Turing's halting problem cannot be solved (Turing [44]), which bears a resemblance to Gödel's incompleteness theorem, we know that there is no such algorithm (Chaitin [7]). There is then no algorithm giving us a solution for (6) other than the continued labour of human thought and experience. In information theory, this is known as the full employment theorem.

In conclusion of this section, let me summarize my argument in point form:

- The use of probabilistic concepts among epistemologists is ubiquitous, whereas few of them seem to concern themselves with information theory. I will show that there is a good case that information is a more basic concept than probability, that it can be used to create the foundations for probability, and that it delivers substantial epistemological results that cannot be obtained by using probability theory without the use of information theory.

- Using information theory as our basis we can (a) define probabilities using (5), (b) then use the Kullback-Leibler Divergence as our most general starting point for a mathematical investigation of information (the Kullback-Leibler Divergence is a generalization of the better-known Shannon Entropy for the continuous case), which will further show that (c) Bayesian rationality constraints also follow from the principle of maximum entropy.

- Information theory has the added advantage that it will give us solid results about continuous probability distributions. As mentioned before, the normal distribution and the Poisson distribution have unique characteristics in terms of being minimally informative (or maximally entropic). Without the axioms of information theory, the use of these distributions is subjective (adherents of subjective probabilities, following Frank P. Ramsey and Bruno de Finetti, would consider this a virtue.)

- Information theory, however, soon diverges from applications known from probability theory to go into considerations of complexity. There is an intuitive correspondence between information and complexity. Complexity (mathematically formalized as Kolmogorov Complexity) cannot be algorithmically assessed, according to Chaitin's incompleteness theorem (Chaitin [7]). There are consequently heavy methodological limits on what information theory can do for us in scientific practice. Solomonoff's theory of inductive inference [42] and Wallace's theory of minimum message length [46; 47] are only of limited use to practicioners. There is a strand of non-Bayesian inference methods, however, which appears to have methodological import, especially dealing with simplicity, for example Hirotsugu Akaike's information criterion [1; 4] or Jorma Rissanen's theory of minimum description length [39; 19]. For a Bayesian response see Dowe et al. [11].

2. Information and Divergence

The Kullback-Leibler divergence, in its most general measure-theoretic terms, is

$$D_{\mathrm{KL}}(P\|Q)=\int_X p\log\frac{p}{q}d\mu \tag{7}$$

Let not the math deter you. (Although I will say for those who are interested that μ is any measure on the event space X for which p and q are the Radon-Nikodym derivatives $\frac{dP}{d\mu}$ and $\frac{dQ}{d\mu}$ respectively.) The Kullback-Leibler Divergence was introduced specifically for expressing what we have called (after Jaynes) the principle of maximum entropy, which Solomon Kullback calls the principle of minimum discrimination: given new facts, a new distribution Q should be chosen which is as hard to discriminate from the original distribution P as possible so that the new data produces as small an information gain ($D_{\mathrm{KL}}(P\|Q)$) as possible. The Kullback-Leibler Divergence also provides a link between Shannon's Entropy in the discrete case and Boltzmann's continuous entropy: a concise and illuminating proof for this is in section 2.2 of Guiaşu [20]. For practical applications of the Kullback-Leibler Divergence see Clarke and Barron [9].[5]

Here is a simple example. Imagine a city with 100 taxis. 30 are green, 70 are blue. Your accuracy in recognizing the correct colour of a taxi is 90%.

	Green	Blue	Total
Accurate	27	63	90
Inaccurate	3	7	10
Total	30	70	100

Assume you see a green taxi (sG). How much probability should you assign to the event that the taxi you just saw was in fact green (G)? Bayes' formula gives us a quick answer:

$$Q(G)=\frac{P(sG|G)P(G)}{P(sG)}=\frac{\frac{90}{100}\cdot\frac{30}{100}}{\frac{34}{100}}\approx 0.79$$

You may (erroneously) think that because you are 90% accurate your chances of seeing a taxi that is in fact green are 90%, so $Q'(G)=0.9$. The Kullback-Leibler divergence $D_{\mathrm{KL}}(Q\|Q')$ measures the additional information needed for transmission of a message (in bits, if the base of the logarithm is 2, which we will assume throughout this paper) because your probabilities (Q') are off. Due to your faulty probability assessment you are inefficiently coding your messages. $D_{\mathrm{KL}}(Q\|Q')$ measures the inefficiency. Because the event space is discrete we can substitute the sum for the integral in (7):

$$D_{\text{KL}}(Q\|Q') = Q(G)\log\frac{Q(G)}{Q'(G)} + Q(B)\log\frac{Q(B)}{Q'(B)} =$$

$$0.79 \cdot \log\frac{0.79}{0.90} + 0.21 \cdot \log\frac{0.21}{0.10} \approx 0.08$$

Because your code is based on Q' rather than Q you need 0.08 more code (per bit).

The Kullback-Leibler Divergence, which is the unique divergence to minimize for minimum discrimination in Kullback's sense, will serve as our formal definition of information density with respect to probability distributions. A few remarks are in order:

1. We are using probabilities to define information. There are three approaches to circumvent this order of priority: (a) Define probability by using the axioms of information theory. We will see how this works later in this section. (b) Abandon the probabilistic approach altogether and define information in terms of complexity. The next section will address this approach. (c) Use the notion of probability at hand by definition (5), basing it on the intuitive understanding of information given in (3).

2. The Kullback-Leibler divergence is not a true metric measuring the 'distance' between two probability distributions. There are topological considerations, but they are seriously hampered by the fact that the Kullback-Leibler divergence is neither symmetric nor does it obey the triangle inequality.

3. There are other ways of introducing the notion of information with respect to probability distributions, notably Shannon's Entropy. This would have been more intuitive, as Shannon's Entropy simply bitwise quantifies the information in a random variable:

   $$H(X) = -\sum_{i=1}^{n} p(x_i)\log p(x_i)$$

   Take one toss of a fair coin for example. As I have nothing to lean on to predict it, I will simply need one bit to communicate the result:

   $$H(X) = -\frac{1}{2}\log\frac{1}{2} - \frac{1}{2}\log\frac{1}{2} = 1$$

> If the coin is not fair, heads or tails will be privileged, which I can use to encode the result more efficiently. Shannon's Entropy will be accordingly smaller. If I roll a die instead of a coin, Shannon's Entropy will be accordingly larger as I need more bits to encode the result (if it is a fair eight-sided die, 3 bits). For a proof that Shannon's Entropy is unique in quantifying this information, subject to highly intuitive axioms, see Khinchin [29, 9]. The Shannon Entropy, however, is much less capable of generalization than the Kullback-Leibler divergence which remains well-defined for continuous distributions and is invariant under parameter transformations. Shannon's Entropy, on the other hand, is easily defined using the Kullback-Leibler Divergence (n is the number of possible outcomes, P_{u} is the uniform distribution for these outcomes). For a proof see Guiaşu [20, 27]:
>
> $$H(X) = \log n - D_{\mathrm{KL}}(P \| P_{\mathrm{u}})$$

Now we want to show that Bayesian updating reflects the principle of maximum entropy. In other words, Bayesian conditionalization gives us the posterior probability distribution that is least informative given the evidence. We arrive at this result not by using axioms of probability but rather by using the axioms of information theory, in our case the Kullback-Leibler Divergence. Let P be the prior probability distribution, Ω the event space, $E \subset \Omega$ our evidence with $P(E) \neq 0$. Using the Radon-Nikodym theorem, there is a non-negative, real-valued, measurable function p such that

$$P(X) = \int_X p(\omega) dP(\omega) \text{ for all } X \subset \Omega$$

(We will from now on abbreviate $dP(\omega)$ by dP.) The evidence restricts the event space from Ω to E. The successor probability distribution of P, given evidence E, is

$$P_E(X) = \frac{P(X)}{P(E)} \text{ for } X \subset E$$

It is easy to show that P_E is a probability measures for E. Let

$$p_E(\omega) = \frac{p(\omega)}{P(E)}$$

p_E is the Radon-Nikodym derivative of P_E with respect to P:

$$P_E(X) = \frac{P(X)}{P(E)} = \frac{1}{P(E)} \int_X p(\omega) dP = \int_X p_E dP \text{ for } X \subset E$$

Now let us define the function:

$$q_E(\omega) = \frac{\chi_E(\omega)p(\omega)}{P(E)}$$

χ_E is the characteristic function of E for which $\chi_E(\omega) = 1$ if $\omega \in E$ and $\chi_E(\omega) = 0$ otherwise. q_E renders the following expression minimal (zero, in fact):

$$\int_E p_E(\omega) \log \frac{p_E(\omega)}{q_E(\omega)} dP \tag{8}$$

Now define $Q(X) = \int_X q_E(\omega)dP$ for all $X \subset \Omega$, which is easily shown to be a probability distribution on Ω, and $Q_E(X) = \int_X q_E(\omega)dP$ for all $X \subset E$, which is easily shown to be a probability distribution on E, for example

$$Q_E(E) = \int_E \frac{\chi_E(\omega)p(\omega)}{P(E)} dP = 1$$

$$Q(\Omega) = \int_\Omega \frac{\chi_E(\omega)p(\omega)}{P(E)} dP = 1$$

$Q_E = Q$ on E and, because of (8), $D_{\mathrm{KL}}(P_E \| Q_E)$ is minimal. It remains to show that $Q(X)P(E) = P(X \wedge E)$ for $X \subset \Omega$, which is just Bayes' formula (multiply both sides by $P(X)(P(X)P(E))^{-1}$ and compare to formula (1)):

$$Q(X)P(E) = P(E) \int_X \frac{\chi_E(\omega)p(\omega)}{P(E)} dP =$$

$$\int_{X \wedge E} \chi_E(\omega)p(\omega)dP + \int_{X \wedge E^c} \chi_E(\omega)p(\omega)dP = P(X \wedge E)$$

I wonder, you may say, didn't we just show that Bayesian conditionalization means what we already knew about Bayesian conditionalization? Yes, but not quite. Our base assumptions were assumptions of information theory, not probability theory. However, we did use what we know from probability theory, i.e. the particular form of the Bayesian posterior probability distribution. It turns out that this particular form is the only one that fulfills the principle of maximum entropy. This type of proof is analogous to two independent attempts of basing information theory on probability theory rather than the reverse, one by Ingarden and Urbanik [24], the other one by Kampé de Fériet and Forte [28]. The above proof was inspired by them but is original to this paper as it specifically addresses Bayesian conditionalization.

A completely different proof using Lagrange multipliers rather than the Radon-Nikodym theorem is given in Caticha and Giffin [5] and Giffin [17]. Ariel Caticha and Adom Giffin show that "Bayes updating is a special case of ME [maximum relative entropy method] updating" [5, 11]. In his PhD thesis [17], Giffin shows that "ME is capable of producing every aspect of orthodox Bayesian inference and proves the complete compatibility of Bayesian and entropy methods" [17, 62]. He furthermore shows that ME can handle updating with data and constraints simultaneously, which neither Bayesian updating nor MaxEnt can do on their own, but only in cooperation.

In the following section, I will provide a brief glimpse of Ingarden's and Kampé de Fériet's project, heavily leaning on Silviu Guiaşu in the latter case [20, 37ff].

(i) Ingarden and Urbanik

Ingarden and Urbanik motivate their project as follows:

> information seems intuitively a much simpler and more elementary notion than that of probability. It gives more a cruder [sic] and global description of some situations physical or other than probability does. Therefore, information represents a more primary step of knowledge than that of cognition of probabilities (just as probability description is cruder and more global than deterministic description). Furthermore, a prinicipal separation of notions of probability and information seems convenient and useful from the point of view of statistical physics. In physics there prevail situations where information is known (e.g. entropy of some macroscopic systems) and may be measured with a high degree of accuracy, whereas probability distribution is unknown and practically cannot be measured at all ... Finally, it may be remarked that a new axiomatic definition of information, free of the inessential connection with probability, clears the way for future generalizations of this notion. [24, 136]

Ingarden and Urbanik define information for non-trivial Boolean rings $A, B, C, \ldots$, whose elements will be denoted by $a, b, c, \ldots$ Let $\mathcal{H}$ be a class of Boolean rings for which all the subrings of its elements are elements and all elements have proper supersets that are elements. Let F be a real-valued function defined on $\mathcal{H}$. We say that two rings A and B are F-equivalent if there exists an isomorphism φ of A onto B such that $F(C) = F(\varphi(C))$ for any subring C of A. $\mathcal{H}$ is decomposed into disjoint sets of F-equivalent rings.

For any pair of isomorphic rings A and B we define

$$\delta_F(A, B) = \min_{\varphi} \max_{C} |F(C) - F(\varphi(C))|$$

where C is running over all subrings of A and φ over all isomorphisms of A onto B. The function ϱ_F defined for any A and B makes $\mathcal{H}$ a pseudo-metric space:

$$\varrho_F(A, B) = \begin{cases} 1 & \text{if } A \text{ and } B \text{ are non-isomorphic} \\ \frac{\delta_F(A,B)}{1+\delta_F(A,B)} & \text{if } A \text{ and } B \text{ are isomorphic} \end{cases}$$

$\varrho_F(A, B) = 0$ if and only if A and B are F-equivalent. A ring A is said to be F-homogeneous if for every automorphism ψ of A and for every subring B of A we have the equality $F(B) = F(\psi(B))$.

A real-valued regular function H (we have skipped a step showing what it means for a function to be regular, for details see [24, 138]) defined on $\mathcal{H}$ is called an information if it has the following properties:

1. The information of rings and their subrings is properly connected. (For details, see [24, 138f].)
2. Local character of information. (For details, see [24, 139].)
3. Monotoneity: if B is a proper subring of A, then $H(B) < H(A)$.
4. Indistinguishability: Isomorphic H-homogeneous rings are H-equivalent.
5. Normalization: if A has two elements and is H-homogeneous, then $H(A) = 1$.

The fundamental theorem about the connection between information theory and probability theory follows:

> Let H be an information on $\mathcal{H}$. Then for every $A \in \mathcal{H}$ there exists one and only one strictly positive probability measure p_A defined on A such that
>
> $$p_B(b) = \frac{p_A(b)}{p_A(1_B)} \text{ for all } b \in B$$
>
> for every subring B of A and
>
> $$H(A) = -\sum_{j=1}^{n} p_A(a_j) \log_2 p_A(a_j)$$
>
> where the a_j are the atoms of A.

The proof of this theorem is substantial and constitutes the heart of Ingarden and Urbanik's paper. You can see the strategy, however: first define what information density is (a real-valued function on the sets for which we want to define information) and then notice how probability theory as we know it falls into place. Kampé de Fériet's approach is more intuitive and accords well with our counterfactual intuitions expressed in formula (3).

(ii) Kampé de Fériet and Forte

Consider a measurable space of events $\{\Omega, \mathcal{K}\}$. The σ-algebra $\mathcal{K}$ is the set of events, and Ω is the total event. The real-valued function I defined on the σ-algebra $\mathcal{K}$ is called a local information on $\mathcal{K}$ if the following axioms are satisfied:

axiom of extreme values The certain event does not contain any information, while the impossible event contains infinite information: $I(\Omega) = 0, I(\phi) = +\infty$.

axiom of monotony For any two events A and B belonging to $\mathcal{K}$ such that $B \subset A$ we have $I(A) < I(B)$.

axiom of union Let f be a topological strictly decreasing convex function. For any disjoint events A and B we have $I(A \cup B) = f(f^{-1}(I(A)) + f^{-1}(I(B)))$.

axiom of continuity The function I is continuous from below at every set in $\mathcal{K}$, i.e. for every increasing sequence of events $(A_i)_{i \in M \subset \mathbb{N}}$ for which $A_1 \subset A_2 \subset \cdots$ we have

$$\lim_{n\to\infty} I(A_n) = I\left(\lim_{n\to\infty} A_n\right) = I\left(\bigcup_{n=1}^{\infty} A_n\right)$$

The theorem analogous to Ingarden and Urbanik's theorem embedding probability theory in information theory is:

> Let $I(A)$ be a local information of the σ-algebra $\mathcal{K}$. If the function F is topological and satisfies the conditions $f^{-1} : \mathbb{R}^+ \to [0,1], f^{-1}(0) = 1, f^{-1}(\infty) = 0$ then the set function μ_f, defined by the equality $\mu_f(A) = f^{-1}(I(A))$ for every event $A \in \mathcal{K}$ is a probability measure on $\mathcal{K}$.

Kampé de Fériet and Forte differ from Ingarden and Urbanik in as far as Shannon's Entropy is only a special case of connection between information theory and probability theory in which information-independence coincides with the probability-independence of events (Guiaşu [20, 41]).

3. Information and Complexity

Kolmogorov's frustration with the priority of probability theory comes from a different place than Ingarden's or Kampé de Fériet's. He wants an information density measure that applies to individual sequences of symbols rather than to the probability distributions behind the sequences of symbols.[6] This approach leads us away from probability to a different intuition connected with information. The series $0, 1, 2, 3, 4, 5, 6, 7, 8, 9$ contains less information than the series $3, 1, 4, 1, 5, 9, 2, 6, 5, 3$. If one of the elements of a series is missing the epistemically justified belief is that we should supply the one that is least informative, i.e. the one which coheres with the pattern. At the same time, we need to beware of overfitting. Formalizations of these intuitions can be found in Solomonoff's theory of inductive inference [42], Wallace's theory of minimum message length (Wallace and Boulton [46]; Wallace and Dowe [47]), Hirotsugu Akaike's information criterion (Akaike [1]; Bozdogan [4]), and Jorma Rissanen's theory of minimum description length (Rissanen [39]; Grunwald [19]).

For our purposes, Kolmogorov's point supplies another reason why information is an epistemologically significant notion, perhaps more so than probability. There is no equivalent of Chaitin's incompleteness theorem in probability theory, and all the projects quoted in the last paragraph show why that is epistemologically relevant. As the name suggests, in *Three Approaches to the Quantitative Definition of Information* [32] Kolmogorov suggests that there are a couple of alternatives to the probabilistic quantification of information: one combinatorial, the other algorithmic. He admits that the combinatorial approach is outstripped by the probabilistic approach,[7] but contends that this is not so for the algorithmic approach. Even a "superficial discourse" [32, 663] reveals the truth of

> two general theses: (1) Basic information theory concepts must and can be founded without recourse to the probability theory in such a manner that 'entropy' and 'mutual information' concepts are applicable to individual values. (2) Thus introduced, information theory concepts can form the basis of the term random, which naturally suggests that random is the absence of periodicity. [32, 663]

The superficial discourse states that algorithmic information in a sequence of symbols is the length of its shortest description. For descriptions, we will use Turing machines. A description for a string s is a program that will output the string s as input to a given Turing machine. This principle was independently articulated in 1964 by Solomonoff, in 1965 by Kolmogorov, and in 1965 by Chaitin. It is called Kolmogorov Complexity (a notorious example of the Matthew Effect, as it should be called Solomonoff Complexity).

For formalization we use universal Turing machines (UTMs). A Turing machine U is universal if it can simulate an arbitrary Turing machine on arbitrary input. For a proof that UTMs exist see Chaitin [6, 14]. Solomonoff proved the invariance theorem of information theory which states that a UTM provides an optimal means of description, up to a constant. Formally,

$$C_U(s) \leq C_M(s) + c$$

The proof is relatively simple using the universality of U (Li and Vitanyi [34, 104]). This is good news for information theory: complexity does not depend on the particular Turing machine we use.

A string is random if its complexity is approximately the same as its length. Both Kolmogorov [31, 663] and Solomonoff [42, 7] appear to equate lack of randomness with regularity or periodicity. In view of fractal geometry, this is questionable. There is a very short computer program that will generate an image of the Mandelbrot set whose boundary does not simplify at any given magnification. It could be argued, although I am not in a position to do this formally, that despite its very low Kolmogorov Complexity the Mandelbrot set is irregular.

More importantly, however, there are two questions that arise as an immediate consequence of this definition of complexity: are most sequences of symbols complex or simple, and is there an algorithm which will tells us how complex a given sequence of symbols is? Gregory Chaitin gives a detailed answer to the first question in [6] (the upshot is that nearly all sequences of symbols are random) and supplies a proof for Chaitin's incompleteness theorem: whether a specific string is complex cannot be formally proved, if the string's complexity is above a certain threshold. The most accessible proofs for Chaitin's incompleteness theorem are online, e.g. `http://wapedia.mobi/en/Kolmogorov_complexity`.

As Kolmogorov recognizes, algebraically the Kolmogorov Complexity lacks the power of Shannon's Entropy.[8] Chaitin's incompleteness theorem assures us that there are no general algorithms or formulas that can deliver an assessment of complexity, i.e. the information contained in a string. The next section will address the question if there is any epistemological merit to the complexity approach to information, or more widely to the information approach to epistemology.

4. Information and Epistemology

Considering the results of the last section, information epistemology has the virtue of being more an epistemology of ignorance than an epistemology of knowledge. There are efforts to use Kullback's principle of minimum relative entropy in certain fields of Artificial Intelligence, but there is also the sober recognition that while there is a "very strong and solid mathematical foundation, the problem is that it is often very hard or even impossible to compute" (van de Ven and Schouten [45, 1]). As we saw in the last section, there are principled obstacles to assessing it from a general perspective. Computer programs that compress data must do so on a case by case basis, and there is no optimal algorithm.

Yet it is intuitively right that given the data we should not maintain beliefs that add more information than necessary. As Jaynes notes, "there is nothing in the general laws of motion that can provide us with any additional information about the state of a system beyond what we have obtained from measurement" [25, 624]. We take this for granted about our laws of motion, so it may be considered reasonable to have similar expectations for epistemology in general.[9] It is not enough to say that our beliefs should rest on evidence: they must rest on evidence in the right way, which appears to be most concisely summarized by saying that our beliefs should not add information where there is no need to do so.

Consider an example from experimental design. The principle of maximum entropy suggests that we should design experiments that make our evidence information-rich, followed up by a hypothesis choice that is information-poor. Jose M. Bernardo [2] has devised an algorithm for prior probability distributions that measures and maximizes missing information: the Reference Posterior Distribution (based on a measure of information provided by experiments in Lindley [35]). Let $p(\theta)$ be our prior density. Then the expected information is:

$$\int p(x) \int p(\theta|x) \log \frac{p(\theta|x)}{p(\theta)} d\theta \, dx$$

where $p(x) = \int p(x|\theta)p(\theta)d\theta$ and $p(\theta|x) = p(x|\theta)p(\theta)/p(x)$. Maximizing this information yields a prior probability distribution that agrees with what we have learned about information theory so far, giving us the uniform distribution $p_k = n^{-1}$ for the discrete case ($\sum_{k\in\{1,\ldots,n\}} p_k = 1$) and the normal distribution for the continuous case. It is largely invariant to parameter transformations (Yang [49]) and ensures that the expected information of the experiment will be maximal. This is no longer a principle of 'indifference': it is pragmatic with a view to Jeffreys' tempering condition, "assign to each seriously proposed hypothesis sufficiently high prior probability to allow the possibility that it will achieve a higher posterior probability than any rival hypothesis as a result of the envisaged observations" [41, 267].[10]

If all is indifferent, there is no information. All there is is entropy, the end of time. But it is not so. We might say, paraphrasing Leibniz's 'cur aliquid potius existit quam nihil?,' why is there information and not rather total entropy? If, on the other hand, all we have ever believed turns out to be false, there is an infinity of information. This lack of entropy altogether must be the closest approximation of Dante's Hell, of Leibniz's Demon. We are, hopefully, somewhere in the middle between total entropy and total information, somewhere between determinants and possibilities, living a life where things are one way and not another, but not fixed yet either, a life full of both specificity and unknowns.

Beginning in the forties and petering out in the seventies, there was great academic interest in information theory, fueled no doubt by the British and American mathematicians who cracked the Japanese and the German codes. Then the interest waned, or moved over to the engineers, when Turing's Halting Problem arrested algorithmization of information theory as Gödel's incompleteness theorem arrested Hilbert's project. Nothing in so dramatic a fashion ever happened to probability theory, and so it still holds epistemological attention, and so there is little we know yet about the limits of its language.

Information theory may have epistemological primacy over probability precisely because there is a mathematical definition of its limits, as there is for logic in Gödel's incompleteness theorem. (For a critique of this analogy, see Raatikainen [38] and Franzén [14], which bears the tell-all title *Gödel's Theorem: An Incomplete Guide to Its Use and Abuse.*) The principle of maximum entropy has as little practicality to it as Bayesian epistemology. Scientists are not known to hunker down with their calculators, plugging in priors and likelihoods to figure out posteriors. What Bayesian epistemology does is give us a pattern of thought, belief revision, and some intelligent ideas about experimental design and hypothesis choice. What this paper tries to achieve is to understand a similar pattern inspired by information theory. A theorem showed that Bayesian epistemology is a special application of this information-inspired reasoning pattern.

There is a Wittgensteinian point here: there is not *more* to be explained about the world other than that things are one way and not another. Things, however, could also go the other way. As important as information is in our Age of Information, "there is no theory, nor even definition, of information that is both broad and precise enough to make such a [notion] meaningful" (Goguen [18, 27]). There may be a justified shift away from an objective interpretation of information as a substance to a subjective interpretation of information as a sign, as in Hjørland [21].[11] Instead of converging on physics (as in the analogy between Boltzmann's Entropy and Shannon's Entropy), information theory may give way to semiotic theories, which in turn may give us results that sound oddly familiar (Noam Chomsky: "we find that no finite-state Markov process that produces symbols with transition from state to state can serve as an English grammar [8, 113]"). Whatever the direction may be, recognizing the significance of information for epistemology will be, well, informative at least.

Endnotes

[1] For an interesting case where flat priors do not coincide with non-informative priors see Zhu and Lu [50]. "The lesson from this discussion is extremely interesting; it tells us that flat priors (such as the uniform prior) are not always the same thing as non-informative priors. A seemingly informative prior can actually be quite weak in the sense that it does not influence the posterior opinion very much. It is clear in our example that the MLE is the result of using a weak prior, whereas the most intuitive non-informative prior (the uniform prior) is not as weak or non-informative as one would have thought" [50, 6]. Jeffreys' priors also are not flat priors as they replace the uniform probability distribution $P(\text{parameter lies between } a \text{ and } b) = \int_a^b dt = b - a$ by $P(\text{parameter lies between } a \text{ and } b) = \int_a^b (t(1-t))^{-.5} dt$ (the argument of the integral is the determinant of the Fisher Information matrix). They pose an interesting problem as it can be shown that their invariance to parameter transformation depends on a variance with respect to experimental design inherent in the Fisher Information. Jeffreys priors therefore violate the likelihood principle that inferences about the parameter should only rest on observed data. We cannot have both invariance to parametrization, for which using the Fisher information turns out be a necessary condition, and invariance to our choice of experimental design. For further information see de Cristofaro [10].

[2] "Though the usual order, according to which information is defined by means of probability, can be reversed, and one can introduce information first, without using probabilities, probabilities inevitably come in at a later stage. The fact that a theory which starts with the aim of defining information without probability leads to the proof of the

existence of probability supports the view that the notion of information cannot be separated from that of probability. To each event A there correspond two numbers: its probability $p(A)$ and its information content $I(A)$ which are connected by the formulas $I(A) = \log_e \frac{1}{p(A)}, p(A) = e^{-I(A)}$" (Guiaşu [20, 37]).

[3] "In physics there prevail situations where information is known (e.g. entropy of some macroscopic systems) and may be measured with a high degree of accuracy, whereas the probability distribution is unknown and practically cannot be measured at all" (Ingarden and Urbanik [24, 136]).

[4] "As a matter of fact, the information seems to represent a more primary step of knowledge than that of cognition of probabilities" (Guiaşu [20, 29]).

[5] "It is seen that $D(P_\theta^n \| M_n)$ is (a) the cumulative risk of Bayes' estimators of the density function, (b) the redundancy of a source code based on M_n, (c) the exponent of error probability for Bayes' tests of a simple versus composite hypothesis, and (d) a bound on the financial loss in a stock-market portfolio selection problem" (Clarke and Barron [9, 455]).

[6] "The need for attaching definite meaning to the expressions $H(x|y)$ and $I(x|y)$, in the case of individual values x and y that are not viewed as a result of random tests with a definite law of distribution, was realized long ago by many who dealt with information theory" (Kolmogorov [31, 662]).

[7] "If we make the variable x and y 'random variables' with given joint probability distributions, we can obtain a considerably richer system of concepts and relationships" (Kolmogorov [32, 161]).

[8] "The meaning of the new definition is very simple. Entropy $H(x|y)$ is the minimal length of the recorded sequence of zeros and ones of a 'program' P that permits construction of the value of x, the value of y being known ... Although Martin-Löf and I realized the importance of the new concept, the development was hindered because the simplest formulas that can be produced as a result of simple algebraic transposition of (1) [Shannon's Entropy] could not be derived from the new definitions" (Kolmogorov [31, 662]).

[9] Such generalizations are dangerous, as seen in this rather far-fetched remark by Solomonoff: "Suppose that all of the sensory observations of a human being since his birth were coded in some sort of uniform digital notation and written down as a long sequence of symbols. Then, a model that accounts in an optimum manner for the creation of this string, including the interaction of the man with his environment, can be formed by supposing that the string was created as the output of a universal machine of random input" [42, 13].

[10] Compare this also to Jaynes' comment that "the maximum-entropy distribution may be asserted for the positive reason that it is uniquely determined as the one which is maximally noncommittal with regard to missing information, instead of the negative one that there was no reason to think otherwise" ([25, 623]).

[11] "The problem is also about whether problems of information science are best served with theories like Shannon and Weaver's information theory or with theories more related to semiotics. In the history of information science, the tendency has been a development

from information theory toward more semiotic theories" (Hjørland [21, 1455]).

References

[1] Akaike, H., 1974. A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

[2] Bernardo, J. M., 1979. Reference Posterior Distributions for Bayesian Inference. *Journal of the Royal Statistical Society. Series B.*, 41(2):113–147.

[3] Bertrand, J. L. F., 1888. *Calcul des probabilités.* Gauthiers-Villars, Paris.

[4] Bozdogan, H., 2000. Akaike's Information Criterion and Recent Developments in Information Complexity. *Journal of Mathematical Psychology*, 44(1):62–91.

[5] Caticha, A. and Giffin, A., 2006. Updating Probabilities. In *MaxEnt 2006, the 26th International Workshop on Bayesian Inference and Maximum Entropy Methods.*

[6] Chaitin, G., 1974. Information-theoretic computation complexity. *Information Theory, IEEE Transactions on*, 20(1):10–15.

[7] Chaitin, G. J., 1966. On the Length of Programs for Computing Finite Binary Sequences. *J. ACM*, 13(4):547–569.

[8] Chomsky, N., 1956. Three Models for the Description of Language. *Information Theory, IRE Transactions on*, 2(3):113–124.

[9] Clarke, B. S. and Barron, A. R., 1990. Information-theoretic asymptotics of Bayes methods. *Information Theory, IEEE Transactions on*, 36(3):453–471.

[10] de Cristofaro, R., 2004. On the foundations of likelihood principle. *Journal of Statistical Planning and Inference*, 126(2):401 – 411.

[11] Dowe, D. L.; Gardner, S.; and Oppy, G., 2007. Bayes not Bust! Why Simplicity is no Problem for Bayesians. *The British Journal for the Philosophy of Science*, 58(4):709–754.

[12] Fadeev, D., 1957. *Zum Begriff der Entropie eines endlichen Wahrscheinlichkeitsschemas.* Deutscher Verlag der Wissenschaften, Berlin.

[13] Feller, W., 1957. *An Introduction to Probability Theory and Its Applications.* Wiley, New York.

[14] Franzén, T., 2005. *Gödel's Theorem: An Incomplete Guide to Its Use and Abuse.* A K Peters, Ltd.

[15] Friston, K., 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

[16] Gettier, E. L., 1963. Is Justified True Belief Knowledge? *Analysis*, 23(6):121–123.

[17] Giffin, A., 2008. *Maximum Entropy: The Universal Method for Inference.* PhD dissertation, University at Albany, State University of New York, Department of Physics.

[18] Goguen, J. A., 1997. Towards a Social, Ethical Theory of Information. In G. Bowker, editor, *Social Science Research, Technical Systems, and Cooperative Work: Beyond the Great Divide*, pages 27–56. Erlbaum.

[19] Grunwald, P., 2000. Model Selection Based on Minimum Description Length. *Journal of Mathematical Psychology*, 44:133–152.

[20] Guiaşu, S., 1977. *Information Theory with Application.* McGraw-Hill, New York.

[21] Hjørland, B., 2007. Information: Objective or subjective/situational? *Journal of the American Society for Information Science and Technology*, 58(10):1448–1456.

[22] Ingarden, R. S., 2006. Information Theory and Thermodynamics of Light Part I Foundations of Information Theory. *Fortschritte der Physik*, 12:567–594.

[23] Ingarden, R. S. and Kossakowski, A., 1971. Poisson Probability Distribution and Information Thermodynamics. *Bulletin of the Polish Academy of Sciences Mathematics Astronomy Physics*, 19:83–86.

[24] Ingarden, R. S. and Urbanik, K., 1962. Information Without Probability. *Colloquium Mathematicum*, 9:131–150.

[25] Jaynes, E. T., 1957. Information Theory and Statistical Mechanics. *Physical Review Online Archive (Prola)*, 106(4):620–630.

[26] Jones, D. S., 1979. *Elementary Information Theory.* Clarendon, Oxford.

[27] Kampé de Fériet, J., 1963. *Théorie de l'Information. Principe du Maximum de l'Entropie et ses Applications à la Statistique et à la Mécanique.* Publications du Laborataire de Calcul de la Faculté des Sciences de l'Université de Lille, Lille.

[28] Kampé de Fériet, J. and Forte, B., 1967. Information et probabilité. *Comptes rendus de l'Académie des sciences*, A 265:110–114.

[29] Khinchin, A. I., 1957. *Mathematical Foundations of Information Theory.* Dover Publications, New York.

[30] Kolmogorov, A. N., 1950. *Foundations of the Theory of Probability.* Chelsea Publishing, New York.

[31] Kolmogorov, A. N., 1968. Logical Basis for Information Theory and Probability Theory. *Information Theory, IEEE Transactions on*, 14(5):662–664.

[32] Kolmogorov, A. N., 1968. Three Approaches to the Quantitative Definition of Information. *International Journal of Computer Mathematics*, 2(1):157–168.

[33] Kullback, S., 1959. *Information Theory and Statistics.* Dover Publications.

[34] Li, M. and Vitanyi, P., 1997. *An Introduction to Kolmogorov Complexity and Its Applications (Texts in Computer Science).* Springer.

[35] Lindley, D. V., 1956. On a Measure of the Information Provided by an Experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005.

[36] Loève, M., 1955. *Probability Theory.* Van Nostrand, Princeton, NJ.

[37] MacKay, D. J. C., 2002. *Information Theory, Inference, and Learning Algorithms.* Cambridge University, Cambridge.

[38] Raatikainen, P., 1998. On Interpreting Chaitin's Incompleteness Theorem. *Journal of Philosophical Logic*, 27:569–586.

[39] Rissanen, J., 1968. Modeling by Shortest Data Description. *Automatica*, 14:465–471.

[40] Shannon, C. E., 1948. A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27(1):379–423, 623–656.

[41] Shimony, A., 1993. *The Search for a Naturalistic World View.* Cambridge University Press.

[42] Solomonoff, R., 1964. A Formal Theory of Inductive Inference. *Information and Control*, 7(1):1–22.

[43] Teller, P., 1976. Conditionalization, Observation, and Change of Preference. In *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science.* D. Reidel, Dordrecht.

[44] Turing, A. M., 1937. On Computable Numbers, with an Application to the Entscheidungsproblem. *Proc. London Math. Soc.*, s2-42(1):230–265.

[45] van de Ven, A. and Schouten, B. A., 2010. A Minimum Relative Entropy Principle for AGI. In *Advances in Intelligent Systems Research.* Atlantis Press, Amsterdam.

[46] Wallace, C. S. and Boulton, D. M., 1968. An Information Measure for Classification. *Computer Journal*, 11(2):185–194.

[47] Wallace, C. S. and Dowe, D. L., 1999. Minimum Message Length and Kolmogorov Complexity. *The Computer Journal*, 42(4):270–283.

[48] Williamson, T., 2000. *Knowledge and Its Limits.* Oxford University Press, Oxford.

[49] Yang, Y., 1995. Invariance of the Reference Prior under Reparametrization. *Test*, 4(1):83–94.

[50] Zhu, M. and Lu, A. Y., 2004. The Counter-Intuitive Non-Informative Prior for the Bernoulli Family. *Journal of Statistics Education*, 12(2):1–10.