Philosophy of Science and Information

Ioannis Votsis
New College of the Humanities / Heinrich-Heine University Duesseldorf
ioannis.votsis@nchum.org / votsis@phil.hhu.de

1. Introduction

Of all the sub-disciplines of philosophy, the philosophy of science has perhaps the most privileged relationship to information theory. This relationship has been forged through a common interest in themes like induction, probability, confirmation, simplicity, non-ad hoc-ness, unification and, more generally, ontology. It also has historical roots. One of the founders of algorithmic information theory (AIT), Ray Solomonoff, produced his seminal work on inductive inference as a direct result of grappling with problems first encountered as a student of the influential philosopher of science Rudolf Carnap. There are other such historical connections between the two fields. Alas, there is no space to explore them here. Instead this essay will restrict its attention to a broad and accessible overview of the aforementioned common themes, which, given their nature, mandate an (almost exclusive) emphasis on AIT as opposed to general information theory.

2. Induction, Probability and Confirmation

David Hume is widely known for having noted that there is something unsettling about the trust we put in inductive inferences. Roughly speaking, these are inferences where the truth of each and every premise does not guarantee, but nonetheless is meant to suggest, the truth of a conclusion. The most paradigmatic of such inferences, abundant in science and everyday life, project past observations into the future. But what underwrites their 'validity'? Hume reasoned that nothing can play that role as neither a deductive nor an inductive approach gets us anywhere. Take the latter. We may, for example, argue that inductive inferences have often produced true or at least largely accurate conclusions in the past hence they will continue to do so in the future. Alas, *that* inference is inductive thereby rendering this approach circular. Indeed, that's even putting aside additional difficulties, an evident one being that numerous *prima facie* reasonable inductive inferences yield false or highly inaccurate conclusions. Bertrand Russell's chicken story is instructive. A chicken, fed daily for a number of weeks, formulates the inductive inference that it will be fed every day. The day comes when its neck is wrung instead, thereby making that inference invalid.

Now consider the deductive approach. We may attempt to find deductive support for a principle of induction. Such a principle would presumably be a contingent and very general truth. Something to the effect that nature is uniform in such and such a way. Notice that deduction guarantees the truth of a conclusion so long as the premises employed are true and their content is *at least as general* as that of the conclusion. That's why deduction is sometimes characterized as 'content-preserving'. But we know, recall the poor chicken, that the future need not resemble the past. Hence, if the premises are made up of contingent truths about the past – as it seems they should if they are to be evidential in content – they would not be sufficiently general to guarantee the truth of a conclusion about the future.[1] The deductive approach to propping up induction then appears hopeless.

At least among philosophers, there is a consensus that the problem of induction is insoluble. One of the leading dissenting voices in modern times is Karl Popper. It's not so much that he

[1] Hume's own objections to the deductive approach differ from that just given. One of his objections is that deduction is inapplicable in such cases, for in his view deductive arguments involve only necessary truths as premises. Today this view is considered antiquated.

believes that the problem of induction can be solved as that he deems induction should be shunned. In its stead, Popper argues that scientific (and presumably everyday) reasoning proceeds, and ought to proceed, first conjecturally and then deductively. The conjectural stage, also known as the context of discovery, is not, according to him, guided by logic but by creativity and imagination. The stage that follows, also known as the context of justification, is guided by deductive logic. We deduce consequences from the conjectures and test them against observations. A contradiction spells the end of the conjecture. Agreement, Popper holds, merely postpones a conjecture's demise. Scientific conjectures cannot be verified but only falsified in such a framework. The problem of induction in Popper's so-called falsificationist framework simply fades away.

Despite Popper's hostility toward verification, his account of goings-on in the context of justification shares much with what we would today call the study of confirmation. This is the study of the conditions under which evidence supports, or ought to support, a hypothesis and sometimes even the level of that support. An influential figure in this study as well as its links to the studies of induction and probability is Carnap. Let us begin with his take on probability.

The axioms are the least controversial part of the theory of probability. What is highly controversial is the interpretation of the emerging notion of probability, which is meant to conform to those axioms. Carnap thought that two interpretations stand out and give rise to two distinct notions of probability. One interpretation conceives of probability in terms of relative frequencies (in the long run). This is what Carnap called 'probability$_2$'. It is the interpretation we most commonly encounter in statistics and applied science and can be expressed as follows: The probability of an event type *E* in a class of events types *C* is the relative frequency of instances of *E* in (the limit of) a long series of a random experiments. More simply, the probability is given by counting how many times *E* occurs, e.g. drawing spades, in a repeated test, e.g. always drawing from a randomly shuffled full-deck of cards, as opposed to alternatives in *C*, e.g. drawing clubs, hearts or diamonds. The other interpretation championed by Carnap conceives of probability in logical terms. This is what he calls 'probability$_1$'. The so-called 'logical' interpretation is an attempt to show that there exists an inductive relation between statements analogous to that of entailment in deductive logic. The analogy is facilitated by the notion of confirmation or support. We say that in cases where a statement *B* deductively follows from a statement *A* the latter statement fully supports the former. If we accept the claim that some support relations are only partial then it seems only reasonable to suppose that there are partial entailment relations. Hence, the need for a logic of induction.

To elucidate his notion of a partial entailment or inductive support relation, Carnap asks us to construct an artificial language that contains names for objects and properties as well as some basic logical machinery, e.g. the conjunction *&* and negation ~ operators. In such a language we can describe a possible state of the world in terms of combinations of conjunctions of atomic statements or their negated counterparts. Suppose there are two objects, denoted by letters *a* and *b*, and one property, denoted by letter *P*, we want to model with our language. A complete description of a *possible* state of this world is given by a statement that states for any given object whether or not it has that property. In our little example this gives rise to exactly four such complete descriptions, also known as 'state descriptions':

1. *Pa* & *Pb* 2. *~Pa* & *Pb* 3. *Pa* & *~Pb* 4. *~Pa* & *~Pb*

One upside of this approach is that the state descriptions can faithfully encode the content of statements about the world. For example, the statement that there exists at least one thing with

the property denoted by P, $(\exists x)Px$, is representable in terms of the disjunction of state descriptions 1, 2 and 3. Another upside is that we can use the state descriptions to determine inference relations. On the assumption that Pb, we can infer $(\exists x)Px$ by appealing to an inclusion relation between the corresponding representations. To be precise, Pb corresponds to the disjunction of state descriptions 1 and 2 and this is included in the disjunction of the state descriptions corresponding to $(\exists x)Px$. In short, we may model deductive inference, i.e. a relation of full support, in terms of this inclusion relation.

What about inductive inferences? Well, we may express such partial support via an overlap relation. Suppose we want to find out whether Pb partially supports the statement that every object has property P, i.e. $(\forall x)Px$. Whenever we ask of a given statement whether it supports another statement we assume that the first is true. On that assumption, certain state descriptions are ruled out, namely those where the statement does not hold. Thus, Pb rules out state descriptions 3 and 4 since they assert $\sim Pb$. Carnap suggests that the support Pb confers to $(\forall x)Px$ is given by the ratio of the overlap between the state description(s) corresponding to each statement to the number of state descriptions corresponding to Pb which is the evidence. Since $(\forall x)Px$ is represented by state description 1 and Pb by state descriptions 1 and 2, the overlap is state description 1. The support relation is thus determined at 1/2. Thus, assuming that each and every state description has an equal weight, the degree of confirmation conferred by Pb onto $(\forall x)Px$ is 0.5.

As it turns out that assumption leads to some undesirable consequences. Suppose we get Pa as evidence and we want to figure out how much support this statement lends to Pb. We know that Pb holds in two state descriptions: 1 and 2. Thus, its original confirmation level stands at 0.5 since it holds in two out of four possible state descriptions. Now we acquire evidence Pa, which also holds in two state descriptions: 1 and 3. The overlap between Pa and Pb is state description 1. But notice that the ratio of this overlap to the number of state descriptions corresponding to Pa is also 0.5. It seems that the confirmation level has remained unchanged even in light of new evidence. This contradicts the intuitive idea that the confirmation of a proposition should increase when we learn something new about, and hence have in some sense additional evidence for, that proposition.

Aware of this difficulty, Carnap proposed a novel way of assigning weights to descriptions. In his view, weight should be equally distributed between structural descriptions, not state descriptions. A structural description is a coarsening of the notion of a state description where what matters is, unsurprisingly, structure. In our example, state descriptions 2 and 3 share structure in that they both posit that one object possesses the property denoted by P and one object lacks it. State descriptions 1 and 4 each have a structure that's unlike no other, namely all objects possess the property denoted by P and none of them do, respectively. We thus end up with three different structural descriptions. Each is assigned a 1/3 weight. If we maintain indifference with respect to the weight of state descriptions *within* a given structural description, state descriptions 2 and 3 each gets assigned a weight of 1/6 whereas state descriptions 1 and 4 each gets 2/6.

The revised account of confirmation given by Carnap can be captured by the following function c:

Carnap's confirmation function: $c(h, e) = w(h \text{ overlap } e)/w(e)$ where h is a hypothesis, e is piece of evidence and w the structural weight function.

This then is Carnap's proposed logic of induction that serves also as a theory of confirmation. Clearly, it is not offered as a solution to the problem of induction but rather as a means through which we can make sense of the level of support premises grant to a conclusion.

Unfortunately, this is not the end of the story for Carnap's theory as it is afflicted by additional difficulties. One major difficulty concerns the largely arbitrary decision of weight assignment. Carnap was well aware that alternative assignments are possible, each giving rise to different confirmation functions. But the choice of confirmation function has dramatic consequences on which hypothesis, from a number of rivals, is best supported by a body of evidence. As will become obvious below this is a problem that appears in many guises and plagues various accounts of confirmation and simplicity.

Having glanced at the philosophical discussions surrounding induction, confirmation and probability, it is now time to turn to the corresponding information-theoretic ones. We begin with Solomonoff (1964a; 1964b), who picks up on the theme of induction. Solomonoff, like Carnap, is not interested in providing a solution to the problem of induction, as some have suggested.[2] As he clearly indicates: "In general, it is impossible to prove that any proposed inductive inference method is 'correct.'" (1964a, p. 4). Rather he is interested, again like Carnap, in the practical problem of figuring out which hypotheses are best equipped to handle future cases on the basis of some existing evidence. Unlike Carnap, he places this problem in an AIT framework. Indeed, Solomonoff thinks that all problems concerning inductive inferences can be restated in such a framework. Let us focus on those problems that concern inductive inferences from evidence to hypotheses. The restatement of such problems in an AIT framework rests on the largely undisputed assumption that languages, whether natural or artificial, encode information in terms of sequences of symbols. Since evidence and hypotheses carry information, it is not unreasonable to suggest that they can also be encoded thus. Now take a body of evidence formulated as a sequence of symbols. Any extension of this sequence can be thought of as a hypothesis that predicts how the existing sequence of symbols develops. Thus, the question 'What is the probability of a hypothesis given certain evidence?' now reduces to the question 'What is the probability that a given extension turns out true?'[3]

Solomoff's proposed answer to this question involves a theorem that is derivable from the axioms of probability:

Bayes theorem: $P(H/E) = P(E/H) * P(H) / P(E)$ where $P(H/E)$ is the posterior probability of a hypothesis H given a piece of evidence E, $P(H)$ is the prior probability of H, $P(E/H)$ is the likelihood of E given H and $P(E)$ is the prior probability of E.

The theorem is a central cog in an influential theory of confirmation that is known as Bayesian confirmation theory. Intuitively, the prior probabilities are the probabilities we take H and E to possess *before* any calculation is made and the likelihood of the evidence is how likely the evidence is made by the hypothesis. How do we determine these probabilities? Bayes theorem itself does not offer any guidance. Indeed, strongly subjective Bayesians deem

[2] Hutter (2007) claims that Solomonoff's theory effectively solves the problem of induction. Solomonoff himself is more guarded.

[3] Solomonoff makes these connections explicit when he asserts: "In the language of Carnap (1950), we want $c(a, T)$, the degree of confirmation of the hypothesis that a [the sequence extension] will follow, given the evidence that T [the original sequence] has just occurred. This corresponds to Carnap's probability$_1$." (1964a, p. 2).

that that's how things should be. Probabilities in their view should express degrees of belief, roughly, a subjective measure of confidence in a given proposition. This put them directly at odds with Carnap's account which deems probabilities to be purely objective.

Having said this, subjective Bayesians do not claim that we are entirely in the dark when attempting to determine initial probabilities. First of all, there are the trivial, least controversial assignments. Tautologies, for example, are thought of as certainly true and hence assigned probability 1. By contrast, contradictions are thought of as certainly false and therefore assigned probability 0. Similarly, when a hypothesis deductively entails the evidence, the likelihood is assigned probability 1. Other rules of thumb include the assignment of non-extreme values when there is no a-priori reason to have too little or too much confidence in a proposition and the assignment of low probabilities to surprising evidence and, conversely, high probabilities to evidence that is to be expected. A natural question to ask at this point is 'Why should we take this approach seriously if different subjects have the freedom to choose different priors?' An answer that goes a long way in allaying concerns, though admittedly not all the way, is that once the evidence begins to trickle in and Bayes theorem is repeatedly put to use, if certain rather reasonable conditions are met, any initial differences in the priors fade away as the values of the posterior probabilities converge. This effect is known as 'the washing out of the priors'.

Solomonoff contribution comes in the form of a method that removes subjectivity from the choice of priors. His method appeals, among other things, to the intuitions underlying Occam's razor.[4] Simpler hypotheses, Solomonoff reasons, are more likely to be predictively accurate than more complex ones. In AIT terms, the role of simpler hypotheses is played by shorter input strings, i.e. programs, in a Universal Turing Machine (UTM) whose output is the desired sequence extensions. That is to say, shorter input strings are claimed to be better predictors of a given output string than longer ones. A UTM is the highly abstracted notion of a machine that can emulate all other Turing machines, themselves abstractions, and therefore capable of implementing any computable function. On the assumption that simplicity is a virtue – more on this in the section below – simpler hypotheses are rewarded with higher prior probabilities.

In his bid to dodge the arbitrariness problem that afflicted Carnap's weight assignment, Solomonoff constructs what has since been called the 'universal distribution'. A distribution is a statistical notion that signifies the assignment of probabilities to each member of a set of alternative hypotheses about a given domain.[5] By a universal distribution, Solomonoff means an assignment of probabilities that concerns *all* the alternative hypotheses to *any* given domain. In AIT terms, it is the distribution that assigns probabilities to all output sequences in a UTM that has been fed with a random input sequence. Take a string α expressed as a binary sequence, e.g. 01011101... Various input strings, also expressed as binary sequences, produce α as an output. Suppose σ_i is one such input string. Let us denote its length with $L(\sigma_i)$. The

[4] It is widely thought that Solomonoff's approach is motivated solely or mostly by Occam's razor. Solomonoff certainly didn't think so. In his mind, the approach stands on much firmer ground via successful applications to a diverse number of specific problems where we have "strong intuitive ideas about what the solutions are" (1964a, p. 5).

[5] For example, suppose we want to find out the probability of which side a coin lands on when it is tossed. On the assumption that the coin and tossing mechanism are unbiased (and that the coin cannot land sideways) the distribution assigns equal probability to heads and to tails.

probability that α is produced by σ_i is given by $2^{-L(\sigma i)}$. This means that the shorter the input string the higher the probability.[6]

Let us take stock of what the notion of a universal distribution is meant to accomplish. By targeting such a distribution, Solomonoff mirrors Carnap's attempts to determine the probabilities of each and every hypothesis formulable in an artificial language. The only difference is that the language chosen is that of a UTM. That's supposed to be a key strength of the approach, for UTM, qua the most general-purpose type of machine, is uniquely positioned to arbitrate between competing hypotheses and hence, the intuition goes, the resulting distribution is likewise unique. To its supporters, it represents the most promising method of assigning prior probabilities to rival hypotheses.

Although it might not seem like it at first, the said universality also turns out to be a shortcoming of sorts. Under Solomonoff's proposal, the input strings, also known as 'descriptions', are meant to be fed into an ideal abstract machine. But clearly, our worldly dealings are with concrete machines. The push to develop a more practicable version of the aforementioned ideas has resulted in what is nowadays called the 'minimum description length' (MDL) approach. The approach dates back to Jorma Rissanen's pioneering work – see his (1978). This 'practical turn' facilitates the application of the central ideas of AIT to a number of fields, including data compression, machine learning, and model selection. Take the last field as an illustration. A model, crudely speaking, is a set of alternative hypotheses, functions or probability distributions that share the same form. For example, all polynomials of degree n, where $n > 0$, have the same form and therefore can be said to fall under the same model.[7] Model selection employs rules that determine which of a number of different models best accounts for a set of data. MDL's rules capitalize on the idea that the more a data set exhibits regularity, the shorter the formulable descriptions whose output is that data set. Following AIT norms, such rules then urge us to pick the model with the shorter description.

An interesting facet that is sometimes neglected in reconstructions of Solomonoff's arguments is that higher priors are not only reserved for shorter descriptions but also for multiple descriptions of the same sequence. The rationale behind this second condition is that "if an occurrence has many possible causes, then it is more likely" (1964a, p. 8). Note, however, that the calculation of priors now becomes more complicated as the two conditions sometimes pull in opposite directions. A natural solution to this problem is the assignment of weights, which Solomonoff duly proposes. Though such a move is clearly necessary, it makes philosophers of science twitchy for much the same reasons as those given above concerning Carnap's weight assignment. Unless a clear justification can be found for those weight assignments, it seems always possible to come up with an alternative assignment that inverts the rankings of what counts as the most simple and therefore in some sense most desirable hypothesis.

As already noted, Solomonoff was not alone in laying the foundations for AIT. Two other figures played an equally pivotal role: Andrey Kolmogorov (see, for example his 1965) and Gregory Chaitin (see, for example, his 1966). The central ideas found in AIT seem to have been independently produced by all three theorists. Interestingly, though Solomonoff seems to have got there first, the idea of measuring complexity in terms of the shortest program that

[6] Note that the shortest input string has length 1 and yields probability 0.5. The longer the input string, the closer the associated value gets to zero.

[7] Thus, second degree polynomials fall under the same model, third degree polynomials fall under another model, and so on.

can produce a certain output is now widely known as 'Kolmogorov complexity', also sometimes called 'Kolmogorov-Chaitin complexity'.

Earlier in this section we noted Solomonoff's interest in devising a practical solution to the problem of drawing reasonable inductive inferences. As it turns out, his solution is quite impractical. Not only because the machines at issue are abstract but also because the approach is un-computable, i.e. no Turing machine can compute some of its algorithms in a finite number of steps. Having said this, as with any ideal solution to a problem, its strength lies not in its practicability but rather in its ability to play a regulative role in our search to find solutions that approximate the ideal. That's where MDL and other AIT-inspired approaches come in handy.

3. Simplicity

William of Occam (also spelled 'Ockham') is one of few notable philosophers to have emerged in the middle ages. His name has become synonymous with simplicity as a virtue of hypotheses. This is sometimes understood as the claim that 'the simplest hypothesis is the most likely to be true'. Yet Occam's own pronouncements, typified in what has come to be known as 'Occam's razor', do not quite say this. For example, in *Summa Logicae*, he states: "Pluralitas non est ponenda sine necessitate" (1974, p. 185). This translates, roughly, as 'plurality is not to be posited without necessity'. The plurality at issue here seems to be an ontological one. That is to say, the emphasis is on reducing ontological complexity. There is no direct mention of simpler hypotheses in this or other quotations.[8] Having said this, it is natural to interpret the positing of fewer entities in terms of simpler, or as they are sometimes called 'more parsimonious', hypotheses, for, if anything, hypotheses are at least hotbeds of entity postulation.

One worry that philosophers express about AIT is that its formal treatment of simplicity is too hastily connected to intuitive formulations of simplicity principles, including Occam's razor. For example, it is not clear why a shorter input string *invariably* translates to a more frugal ontology and vice-versa. The philosopher of science Elliott Sober complains that syntactic approaches to simplicity like those deployed in AIT fall afoul of 'the problem of measurement': "Since a proposition can be encoded in many different ways, depending on the language one adopts, measuring simplicity in terms of code features will fail to be linguistically invariant" (2002, p. 16). Sober uses a *version* of the well-known grue paradox to demonstrate this problem. I here present the gist of his argument, omitting certain details. Compare the following hypotheses:

> H_1: All emeralds are green.
> H_2: All emeralds are green until a fixed future date d, thereupon they are blue.

If our simplicity judgments rely on the syntactic length of a hypothesis, then it appears that H_1 is simpler than H_2. Suppose, however, that we start out with a different language, one that contains the predicates grue and bleen instead of green and blue. An object is grue if and only if it is green prior to d or blue after d. An object is bleen if and only if it is blue prior to d or green after d. Utilizing these predicates, we can formulate the following two hypotheses:

[8] Here are some further quotations from Occam: "Si duae res sufficiunt ad eius veritatem, superfluum est ponere aliam (tertiam) rem" and "Frustra fit per plura, quod potest fieri per pauciora" (quoted in Charlesworth 1956, p. 105). The first translates roughly as 'If two entities are sufficient for truth, it is superfluous to posit a third' and the second roughly as 'It is in vain to attempt to do with more what can be done with fewer'.

H_1': All emeralds are grue until a fixed future date *d*, thereupon they are bleen.
H_2': All emeralds are grue.

In this language, H_2' comes out simpler than H_1'. Note, however, and this is the crucial point, that H_1 is logically equivalent to H_1' and H_2 is logically equivalent to H_2'. That, in effect, means that H_1 and H_1' express the same hypothesis. Ditto for H_2 and H_2'. So, depending on the language we start out with, we end up making inverse determinations of the simplicity of two hypotheses. That surely can't be right. Sober concludes: "Stipulating which language should be used resolves this ambiguity, but a further question needs to be answered. Why should we adopt one language, rather than another, as the representational system within which simplicity is measured?" (2002, p. 16).

Pertinent to this type of objection is an *invariance theorem* that Solomonoff and others proved. According to this theorem, for any two general-purpose machine languages and a sufficiently long output string, the length of the shortest description yielding that output in the one language will not exceed a constant *c* when compared to the length of the shortest description yielding the same output in the other. This is taken to mean that it does not matter anymore what language we choose. That doesn't seem right. The theorem suggests that the extent to which the lengths of shortest descriptions vary from language to language is limited. That definitely reduces the impact of objections such as the above but it does not eliminate them. After all, two general-purpose languages may still yield inverted simplicity judgments even though any differences in length will, following the invariance theorem, be comparatively small. It appears then that to banish grue-like objections altogether, AIT theorists need to prove a stronger theorem, e.g. one that establishes the existence of a uniquely privileged machine language.

Returning to Occam's dictum, what the clause 'without necessity' is meant to range over becomes a significant interpretational issue. Among the various candidates, two are worth mentioning and relate to a well-known distinction in the philosophy of science, namely explaining versus saving a class of phenomena. The phrase 'to save the phenomena' goes back to Andreas Osiander who wrote the preface to Copernicus' *De Revolutionibus Orbium Coelestium*. Probably eager to avoid the wrath of the church, Osiander argued that Copernicus' radical model of the universe with the Sun, not the Earth, at its center, was merely aimed at saving, i.e. accounting, for the phenomena in what we would nowadays call an *instrumentalist* manner. There is thus no question of the truth or even probable truth of this model. This contrasts with what we might call a *realist* view, according to which (adequate) explanations don't just save phenomena but also reveal the underlying structure of the world.

Such matters are not only important for scholarship or history. They are also matters about which disagreement can lead to radically distinct conceptions of how hypotheses ought to be chosen. Indeed, the said disagreement can be found both within philosophical discussions as well as AIT ones. As already hinted, there are philosophers who insist on instrumentalist, also known as pragmatic, readings of simplicity principles, e.g. Bas van Fraassen (1980), and those who plunge for more realist-oriented interpretations, e.g. Kevin T. Kelly (2008). Similarly, there are AIT theorists who claim that their fondness for simplicity has nothing to

do with truth, e.g. Peter Grünwald (2007), and those who unabashedly flirt with truth, e.g. Samuel Rathmanner and Marcus Hutter (2011).[9]

How exactly does this dispute matter? Well, the pragmatists are not perturbed much by the existence of conflicting simplicity judgments. Take two competing hypotheses that are expressible as programs in two separate computer languages, say C and PASCAL.[10] One program may be shorter in C and the other program shorter in PASCAL. Thus, depending on the language we start out with, each hypothesis is deemed simpler and hence to be preferred. For a pragmatist this is not as pressing a concern. That's because there is no overarching aim to find the one true hypothesis. The aim is rather to find hypotheses that make life easier for us by, for example, allowing us to make the same, or more or less the same, predictions faster, more efficiently, etc. This is especially true, if there is no compelling reason to prefer one language over the other. By contrast, it is plain that those whose goal is the truth have to reject the claim that one hypothesis is both closer to the truth and at the same time further away from it in relation to another.

In addition to MDL, there are other information-theoretic methods on offer. Sober, for example, is a fan of the Akaike Information Criterion (AIC). Named after its creator, the statistician Hirotugu Akaike, this is also a model selection method. AIC balances considerations of simplicity and goodness of fit to the data. The latter is calculated using well-known statistical methods like maximum-likelihood estimation. The former is measured via the number of free parameters, i.e. those whose values are adjustable, present in a model. The idea, roughly, is that models with fewer free parameters are simpler because they require less ad hoc intervention to produce a higher goodness of fit. In short, AIC rewards goodness of fit but penalizes complexity.[11] In doing so, it guards against the well-known problem of over-fitting, which can be explained as follows. Most data sets contain noise. This means that a model that fits the data perfectly is guaranteed to be false and an imperfect predictor of new data. One straightforward way to avoid this consequence is to opt for simpler models that do not hug the data as closely and hence have at least a chance of being true or perfect predictors. Thus, there seems to be a good rationale for penalizing complexity both through AIC but also through other approaches that counsel against over-fitting.

The last method to be briefly explored here is the Bayesian Information Criterion (BIC) due to Schwarz (1978). This is remarkably similar to AIC in that both approaches trade off simplicity and goodness of fit via an almost identical mathematical expression. In fact, the only difference between them seems to be that BIC takes into account the size of the data set in its estimation of the simplicity term. The consequence is that BIC tends to penalize complexity more than AIC, especially as the size of the data set increases.

These and other methods all vie to capture the idea that simplicity is a virtue and a powerful criterion in model selection. If there is no way to decide between the available methods, does this mean, as some philosophers have suggested, that simplicity is merely an aesthetic

[9] Grünwald, for example, says that "there is no place for a 'true distribution' or a 'true state of nature' in this view" (p. 27). Rathermanner and Hutter, by contrast, assert that "we are interested in finding the true governing process behind our entire reality" (p. 1089).

[10] Since no UTM is at hand, AIT theorists employ the next best thing, namely general-purpose computer languages like C and Pascal.

[11] For the mathematically inclined, AIC is computed through the following or a similar expression: $2k - 2 * log (P(D/L(M)))$ where D is a set of data, M is a model, $L(M)$ is the maximum likelihood estimate of M, $P(D/ L(M))$ is the likelihood of the data and k the number of free parameters in the model.

criterion? This question overlooks some important details in the debate over the right way to measure simplicity. First of all, it has already been made clear that, on account of the presence of noise in data sets, some (at least minimal) bias toward simpler models is required. Secondly, the aforementioned and other methods have been shown, either by mathematical proof or by simulation, that they are, under certain conditions, quite good at finding and even finding fast the (by stipulation) true hypothesis or at least the one best predicts the data (Gerda Claeskens and Nils Lid Hjort 2008). Thirdly, though these methods do not produce identical judgments, they do, once again under certain conditions, exhibit strong convergences (see, for example, Jun Shao 1997). And, fourthly, even if the world is in fact rather complex and hence demands rather complex models to faithfully describe it, it is obviously not maximally complex and therefore imparts upon us the tenet that complexity should sometimes be penalized. These four considerations transform the original question from one where the virtue of simplicity as a non-merely-aesthetic criterion is doubted to one where what is being doubted is only how much of an objective role simplicity should play in determining our selections of models and hypotheses.

Two notions intimately related to simplicity are non-ad hoc-ness and unification. Both are considered virtues and are employed by practicing scientists as informal criteria in deciding between competing models and hypotheses. Although we do not have the space to explore them at length here, it is important to at least make some cursory remarks regarding the role they play in AIT and the philosophy of science. In both disciplines, simplicity and non-ad hoc-ness are often mentioned in the same breath (e.g. Grünwald 2007 and Kelly 2008). The semantic proximity of the two notions becomes obvious when one considers, for instance, that the request to reduce the number of free parameters in a model also has the direct effect of suppressing ad hoc-ness. After all, the fewer parameters we can adjust the less of an opportunity on offer to fit the data in a quick-gains, short-sighted, fashion.

The unifying power of a hypothesis is perhaps not as easy to connect to its level of simplicity and non-ad hoc-ness. Alas, no serious attempts to articulate the notion of unification seem to exist within the AIT literature, though this author suspects that this will be an area of growth in the future. The same is not true of the philosophical literature. The *locus classicus* here is Friedman (1974), who argues that the fewer independently acceptable law-like premises required in the derivation of an explanation the more unified that explanation.[12] Though Friedman does not specifically address the connection between simplicity and unification or ad hoc-ness, it doesn't take much cognitive ability to identify the common emphasis on fewer postulates. Other philosophers have taken a more direct approach to connecting the aforementioned themes, e.g. Kelly (2008) and Votsis (2015). The latter account builds on Friedman's insights to argue that the more confirmationally connected the content parts of a hypothesis, the higher that hypothesis' degree of unification. Highly unified hypotheses in this sense are invariably non-ad hoc, and hence in some sense quite simple, in that they are not composed of confirmationally unrelated parts that are forcibly contrived to fit together.

4. Scientific Realism

An inveterate debate in philosophy is that between the realists and the anti-realists. The realists advance an ontological claim that some category of things is real and, moreover, an epistemological claim that we have knowledge of this category. Anti-realists deny at least the

[12] Friedman's view is flawed but fruitful as has been repeatedly pointed out in the literature. Interestingly, it chimes well with a passage found in Aristotle, who asserts in *Posterior Analytics*: "Let one demonstration be better than another if, other things beings equal, it depends on fewer postulates or suppositions or propositions" (2002, p. 39).

second claim, sometimes also the first. The debate manifests itself in distinct ways depending on the sub-field of philosophy within which it is conducted. That includes meta-ethics, the philosophy of language and the philosophy of mathematics. At present we are interested in the philosophy of science manifestation, widely known as 'the scientific realism debate'. Scientific realists argue that our best scientific theories, i.e. those that enjoy substantial predictive and explanatory success, reveal real objects and their properties to us, e.g. that DNA molecules are helical in structure or that neutrinos possess a half-integer spin. Moreover, they argue that historically consecutive theories become increasingly successful and, in so doing, move closer to a true description of the world. Scientific anti-realists deny that any such knowledge can be had or progress towards it can be made. To be exact, nowadays the central point of contention is whether unobservables, i.e. objects and properties that are not verifiable via our unaided senses, are knowable. While scientific anti-realists are at most willing to concede that observables are knowable, scientific realists admit the knowability of both observables and unobservables.

There are various connections between the scientific realism debate and information theory. The first of these has already been touched upon in our discussion of simplicity. Scientific realists typically cite simplicity as a truth-apt criterion for choosing between rival hypotheses. By contrast, their anti-realist counterparts claim that it is at best a pragmatic consideration in such matters, at worst a merely aesthetic one. That the scientific realists are keen on recruiting simplicity and other so-called 'theoretical virtues' like unifying power becomes all the more evident when the hypotheses in question are empirically equivalent, i.e. they possess identical empirical consequences. In such cases, deciding between rival hypotheses on purely empirical grounds becomes impossible. Faced with such an impasse, scientific realists employ simplicity as a tiebreaker criterion in the hope that it is indeed capable of leading us to the truth.

The second connection we also already touched upon. Whether or not a scientific realist or anti-realist view of science is more warranted presumably depends on whether or not the claims about which we can be realists or anti-realists can be, and indeed are, confirmable. For example, were it to turn out, as some have argued, that the support a piece of evidence provides can spread to different parts of hypothesis and indeed to parts that make claims about unobservable entities, then scientific realism would gain the upper hand in the debate. Note that the issue of how far support spreads is central to the study of confirmation. Otherwise put, it is the issue of which of a competing number of inductive inferences (that take a piece of evidence as input and yield one or more parts of a hypothesis as output) is most warranted. And that's precisely a topic that AIT theorists also obsess about, one of their counsels being that we should choose those inferences that maintain a certain kind of balance between simplicity and goodness of fit.

The third and final connection is one that we are to freshly address in the remainder of this section. It concerns the source of our ontology. According to the majority of scientific realists, that ontology is best sourced from the wells of successful science and finds its most paradigmatic form in the entities and properties posited by physics. Indeed, some scientific realists advocate an even stronger claim, namely that the only things that can truly be said to exist are those posited by fundamental physics. Thus, pianos, proteins, governments and pulsar stars are nothing more than a bunch of fundamental particles that behave in accord with the laws of physics. How does information theory bear on this issue? Well, a view has recently been put forth that the fundamental ontology of the world is informational. Seeing as this view builds on another, namely structural realism, it is sensible to consider the latter first.

Structural realism is undoubtedly the most influential realist view in the last fifteen years. Its central tenet is that we should be realists only about structure.[13] Though controversy clouds how exactly structure is to be understood, every party in the debate agrees that some abstract logico-mathematical notion is required. One such notion, for example, is set-theoretical. A structure *S* in this sense is denoted by a pair <*U*, *R*>, where *U* is a non-empty set of objects and *R* a non-empty set of relations, i.e. ordered *n*-tuples, defined over those objects. What's so special about such structures? Well, they allow us to obviate the specific character or nature of the objects and relations under consideration and focus instead on their structure (see chapter seven on Levels of Abstraction). For example, the objects may be human beings, particles or mountains and the relations may be *x* is meaner than *y*, *x* is in a higher energy state than *y*, and *x* has a higher elevation than *y*. For a number of reasons, one of which being that the history of science seems to show a continuity only of structures across scientific revolutions, structural realists argue that the *posited* specific character or nature of the objects and relations becomes irrelevant. All that seems to matter are general logico-mathematical properties of relations, e.g. that a relation between the target objects is irreflexive, anti-symmetric and transitive. That's why structural realists find the set-theoretical and/or other such notions of structure valuable.

There have traditionally been two kinds of structural realism. Crudely put, epistemic structural realism (ESR) holds that we cannot know more about unobservable objects other than the logico-mathematical properties of the relations they instantiate. Equally crudely, ontic structural realism (OSR) holds that such objects are at best weak relatives of traditionally conceived objects and at worst fictions that need to conceptualized away. In the last few years, Luciano Floridi (2011) has developed a brand of information-theoretic realism that is a close relative of structural realism, especially OSR, which he calls 'informational structural realism' (ISR). *Qua* realism, ISR is ontically committed to the existence of a mind-independent reality and epistemically committed to some knowledge of that reality in both its observable and potentially its unobservable guises. *Qua* structural, ISR is committed to a structural conception of reality. *Qua* informational, ISR is committed to an understanding of reality that is purely informational. In more detail, Floridi defines ISR as follows:

> Explanatorily, instrumentally and predictively successful models (especially, but not only, those propounded by scientific theories) at a given LoA can be, in the best circumstances, increasingly informative about the relations that obtain between the (possibly sub-observable) informational objects that constitute the system under investigation (through the observable phenomena) (2011, p. 361).

Four parts are worth highlighting. The first concerns the reference to successful models. This reference is in step with the realist idea that *success* is a key motivator for the view that we may potentially possess some knowledge of unobservables, or sub-observables as Floridi calls them. The second part concerns the LoA, i.e. level of abstraction, concept. In rough terms, a *level of abstraction* is that component of a theory that "make[s] explicit and clarify[ies] its ontological commitment… [by] determin[ing] the range of available observables" (p. 348). This is required to provide an analysis of the system under study through a model that identifies the system's structure. Naturally, different levels of abstraction are possible. Floridi argues that the levels of abstraction required by his version of structuralism concurrently entail a first-order ontological commitment to the structural properties of the system and a second-order ontological commitment to the structural character of the system's *relata*. The

[13] For an in-depth critical survey of the varieties of structural realism, readers may consult Frigg and Votsis (2011).

third part, i.e. the part about such models being increasingly informative, is also borrowed from realism. It relays the idea that there is *progress* in getting to know the systems under study. The fourth part concerns the ontology, which, as already noted, consists not of garden-variety physical things but of *informational* things that are structurally conceived.

Why would we want to replace a physical with an informational ontology? Floridi's argument is, in effect, that it would offer a much more general and unified ontological platform. To better understand this argument, we need to make a small detour into three interdependent notions from computer science, namely *portability*, *interoperability* and *scalability*. Roughly speaking, we say of a piece of software that it is portable when, for example, it can be run on more than one type of microprocessor. Equally roughly, we say of a piece of software or hardware that it is interoperable when, for example, it can communicate and interact with other pieces of software and hardware of different types. Finally, we say that a software or hardware solution is scalable when, for example, it remains a solution to a problem even if the size of the problem varies. Floridi's suggestion is that an informational ontology is much more portable, interoperable and scalable than a traditionally conceived physical ontology. As an illustration take portability. OSR, Floridi notes, is already quite portable in that it conceives of its ontology in such a way that it is exchangeable between physical and mathematical theories. This is a consequence of the fact that the ontology in OSR is described through highly abstract mathematical structures. ISR takes a step further, he then argues, by making its ontology "portable to computer science as well" (p. 359). Similar remarks are made in relation to the virtues of interoperability and especially scalability.

Although an interesting idea, it is quite difficult to fathom how the world itself is somehow informational. We certainly represent the world through information. No contention there. But to call the ontology of this world informational is something that opponents of this view would perhaps deem to be a category mistake. Not unless, of course, what is meant by an informational ontology is something much more akin to what is generally understood by a physical ontology. Such an interpretation would run the risk of turning the dispute into a terminological squabble. Interestingly, Floridi's characterization of the operative notion of information in terms of "differences *de re*" seems to have that effect as an unintended byproduct (p. 356). For he appears to be telling us that all that matters are the differences in and between ontological units.[14] But if that's the case, asserting that these units are informational or indeed physical adds nothing of essence to the story. This is not an invitation to conflate informational and physical ontologies but, rather, a reminder that the notion of a physical ontology can be construed in a minimalist way, i.e. without making strong metaphysical assumptions about what it takes for something to be physical.

Let us end this section by saying that the jury is still out on whether ISR will, and more importantly whether it justifiably ought to, develop into a major force within the scientific realism debate. But even if ISR fails on both accounts, it should be clear that there may yet be space in that debate for an informational ontology. Perhaps not in terms of replacing a physical ontology but rather in a pluralistic framework where informational and physical 'entities' live side-by-side.

5. Some concluding remarks

[14] In information-theoretic terms, we might express such differences as distinct symbols, e.g. 0s as opposed to 1s. In physical-theory terms, we might express such differences as distinct states, e.g. spin up or spin down.

In spite of the numerous connections between the fields of information theory and philosophy of science, the interaction between the fields' practitioners remains disappointingly slender. At least part of the reason why philosophers of science have not engaged with the literature on information, and particularly, on AIT (as much as this author and others would have hoped for) seems to be that, more often than desirable, the formal results in that literature are hastily and haphazardly linked to existing philosophical problems.[15] On the other side of the divide, at least part of the reason why information theorists and in particular AIT theorists have not engaged as much with the philosophical literature seems to be that philosophers tend to pursue more 'arcane' aspects of the foregoing themes and certainly less clearly practicable ones. As a result, philosophers sometimes lose touch with reality, in spite of their best intentions. These obstacles notwithstanding, I would like to end on a more positive and constructive note. It is this author's hope that entries like the present will assist in fostering greater interaction between philosophers and information theorists. After all, both groups are keen on making progress towards solving some of the world's most daunting problems.

References

Barnes, J. (2002) *Aristotle*: *Posterior Analytics*, reprinted second edition, translated with commentary, Oxford: Oxford University Press.

Akaike, H. (1973) 'Theory and an Extension of the Maximum Likelihood Principle' in B. N. Petrov and F. Csaki (Eds.), *Second International Symposium on Information Theory*, Budapest: Akademiai Kiado, pp. 267‑81.

Carnap, R. (1950) *Logical Foundations of Probability*, Chicago: The University of Chicago Press.

Chaitin, G. (1966) 'On the Length of Programs for Computing Finite Binary Sequences', *Journal of the ACM*, vol. 13: 547–569.

Charlesworth, M. J. (1956) 'Aristotle's Razor', *Philosophical Studies*, vol. 6: 105-112.

Claeskens, G. and N. L. Hjort (2008) *Model Selection and Model Averaging*, Cambridge: Cambridge University Press.

Floridi, L. (2011) *The Philosophy of Information*, Oxford: Oxford University Press.

Frigg, R. and I. Votsis (2011) 'Everything You Always Wanted to Know about Structural Realism but Were Afraid to Ask', *European Journal for the Philosophy of Science*, vol. 1(2): 227–276.

Friedman, M. (1974) 'Explanation and Scientific Understanding', *Journal of Philosophy* vol. 71(1): 5-19.

Grünwald, P. (2007) *The Minimum Description Length Principle*, Cambridge, MA: MIT Press.

Hume, D. ([1739] 1975) *A Treatise of Human Nature*, L. A. Selby-Bigge & P. H. Nidditch (Eds.), Oxford: Clarendon Press.

Hutter, M. (2007) 'On Universal Prediction and Bayesian Confirmation', Theoretical Computer Science, vol. 384: 33–48.

Kelly, K. (2008) 'Ockham's Razor, Truth, and Information' in *Handbook of the Philosophy of Information*, J. van Behthem and P. Adriaans (Eds.), Dordrecht: Elsevier.

Kolmogorov, A. (1965) 'Three Approaches to the Quantitative Definition of Information', *Problems of Information Transmission*, vol. 1(1): 1–7.

Li, M. and P. Vitányi (1997) *An Introduction to Kolmogorov Complexity and its Applications*, second edition, New York: Springer-Verlag.

Ockham, W. (1974) *Summa Logicae*, P. Boehner, G. Gál & S. Brown (Eds.), New York: The Franciscan Institute.

[15] Solomonoff stands out in having valiantly toiled to make his formal work intuitively comprehensible.

Popper, K. R. (1959) *The Logic of Scientific Discovery*, New York: Basic Books.
Rathmanner, S. and M. Hutter (2011) 'A Philosophical Treatise of Universal Induction', *Entropy*, vol. 13: 1076-1136.
Rissanen, J. J. (1978) 'Modeling by Shortest Data Description', *Automatica*, vol. 14(5): 465–471.
Shao, J. (1997) 'An Asymptotic Theory for Linear Model Selection', *Statistica Sinica*, vol. 7: 221-264.
Solomonoff, R. J. (1964a) 'A Formal Theory of Inductive Inference. Part I', *Information and Control*, vol. 7(1): 1-22.
Solomonoff, R. J. (1964b) 'A Formal Theory of Inductive Inference. Part II', *Information and Control*, vol. 7(2): 224-254.
Sober, E. (2002) 'The Problem of Simplicity', in *Simplicity*, *Inference and Modelling*: *Keeping it Sophisticatedly Simple*, A. Zellner, H. A. Keuzenkamp and M. McAleer (Eds.), Cambridge: Cambridge University Press.
Van Fraassen, B. (1980) *The Scientific Image*, Oxford: Clarendon Press.
Votsis, I. (2015) 'Unification: Not Just a Thing of Beauty', *Theoria*, vol. 30(1): 97-114.