Towards an Informational Pragmatic Realism*

Ariel Caticha
Department of Physics, University at Albany-SUNY,
Albany, NY 12222, USA.

Abstract

I discuss the design of the method of entropic inference as a general framework for reasoning under conditions of uncertainty. The main contribution of this discussion is to emphasize the pragmatic elements in the derivation. More specifically: (1) Probability theory is designed as the uniquely natural tool for representing states of incomplete information. (2) An epistemic notion of information is defined in terms of its relation to the Bayesian beliefs of ideally rational agents. (3) The method of updating from a prior to a posterior probability distribution is designed through an eliminative induction process that singles out the logarithmic relative entropy as the unique tool for inference. The resulting framework includes as special cases both MaxEnt and Bayes' rule. It therefore unifies entropic and Bayesian methods into a single general inference scheme. I find that similar pragmatic elements are an integral part of Putnam's *internal realism*, of Floridi's *informational structural realism*, and also of van Fraasen's *empiricist structuralism*. I conclude with the conjecture that their valuable insights can be incorporated into a single coherent doctrine — an *informational pragmatic realism*.

1 Introduction: an informational realism

The awareness that the concepts of truth and reality are central to science is quite old. In contrast, the recognition that the notion of information is central too is still in its infancy which makes Floridi's *Philosophy of Information* particularly timely [1]. I find much to agree with his the development of a method of levels of abstraction, the adoption of the semantic model of theories, the valuable guidance provided by singling out 18 open problems, including the defence of an *informational structural realism* (see also [2]). Most importantly, Floridi's whole argument displays a certain pragmatic attitude that I find most appealing:[1]

*Invited paper published in the Special Issue on "Luciano Floridi and the Philosophy of Information", Mind and Machines **24**, 37-70 (2014).

[1] The Perceian slant of Floridi's pragmatism is perhaps more explicit in [3] and [4].

> "The mind does not wish to acquire information for its own sake. It needs information to defend itself from reality and survive. So information is not about representing the world: it is rather a means to model it in such a way to make sense of it and withstand its impact." ([1], p. xiv)

However, Floridi and I approach our subject – information – from different directions. Floridi is motivated by developments in computer science and artificial intelligence while I am mostly concerned with physics. One might therefore expect important differences and, indeed, there are many. For example, Floridi offers no discussion of Bayesian and entropic methods of inference which are the main theme of this paper. Nevertheless, rather than contradictory, I find that our approaches are sufficiently different that, in the end, they may actually complement each other.

The connection between information and physics is natural: throughout we will adopt the view that science consists in using information about the world for the purpose of predicting, modeling, explaining and/or controlling phenomena of interest to us. If this image turns out to be even remotely accurate then we might expect that the laws of science would reflect, at least to some extent, the methods for manipulating information. We can, moreover, entertain the radical hypothesis that the relation between physics and nature is considerably more indirect than usually assumed: *the laws of physics are nothing but schemes for processing information about nature.* The evidence supporting this latter notion is already quite considerable: most of the formal structures of thermodynamics and statistical mechanics [5], and also of quantum theory (see e.g. [6][7][8] and references therein) have been derived as examples of the methods of inference discussed later in this paper.

The basic difficulty is that inferences must be made on the basis of information that is usually incomplete and one must learn to handle uncertainty. Indeed, to the extent that "reality" is ultimately unknowable the incompleteness of information turns out to be the norm rather than the exception — which explains why statistical descriptions such as we find in quantum mechanics are unavoidable. The gods do play dice. And this is not surprising: from the informational perspective indeterminism is natural and demands no further explanation. Instead, what needs to be explained is why in some special and peculiar circumstances one obtains the illusion of determinism.

The objective of the present paper is to describe the method of entropic inference as a general framework for reasoning under conditions of uncertainty. The main contribution of the discussion is to emphasize the pragmatic elements in the derivation. The application of entropic methods to derive laws of physics will not be pursued here. Instead, we will tackle three problems. First, we describe how a scheme is designed to represent one's state of partial knowledge[2] as a web of interconnected beliefs with no internal inconsistencies; the tools to do it are probabilities [9][10]. Since the source of all difficulties is the lack

[2] For our purposes we do not need to be particularly precise about the meaning of the term 'knowledge'. Note however that under a pragmatic conception of truth there is no real difference between a 'justified belief' and the more explicit but redundant 'justified true belief'.

of complete information two other issues must inevitably be addressed. One concerns information itself: what, after all, is information? The other issue is: what if we are fortunate and some information does come our way, what do we do with it? We discuss the design of a scheme for updating the web of beliefs. It turns out that the instrument for updating is uniquely singled out to be entropy[3] — from which these entropic methods derive their name. [8] The resulting framework includes as special cases both Jaynes' MaxEnt and Bayes' rule. It therefore unifies entropic and Bayesian methods into a single general and manifestly consistent inference scheme.[4]

In the final discussion I point out that closely related pragmatic elements can be found in Putnam's *internal realism* [11], in Floridi's *informational structural realism* [2] and also in van Fraasen's more recent *empiricist structuralism* [12]. It might therefore be possible to pursue the systematic development of a position – an *informational pragmatic realism* – that takes advantage of the valuable insights achieved in those three doctrines.

2 Background: the tensions within realism

Scientific realism means different things to different people but most scientific realists would probably agree that (1) there exists a real world out there that is largely independent of our thoughts, language and point of view; and (2) and that one goal of science is to provide descriptions of what the world is really like. van Fraasen, who is not himself a realist, describes this form of realism as follows:

> "Science aims to give us, in its theories, a literally true story of what the world is like; and acceptance of a scientific theory involves the belief that it is true." [13]

There is a tension between theses (1) and (2): Can we ever know that our scientific descriptions are true, that is, that they match or correspond to reality? To put it bluntly, is science at all possible?

One solution to this skeptical challenge is an empiricism that accepts thesis (1) of an independent external reality, but denies thesis (2) that we can actually get to know and describe it. This is a type of anti-realism. In such a philosophy true descriptions are not likely and science, if it is to succeed, must strive towards a more modest goal. According to van Fraasen's *constructive empiricism*

[3] Strictly the tool for updating is *relative entropy*. However, as we shall later see, all entropies are relative to some prior and therefore the qualifier relative is redundant and can be dropped. This is somewhat analogous to the situation with energy: it is implicitly understood that all energies are relative to some reference frame but there is no need to constantly refer to a *relative energy*.

[4] I make no attempt to provide a review of the literature on entropic inference. The following incomplete list reflects only some contributions that are directly related to the particular approach described in this paper: [8], [10], [14], [15], [16], [17], and [18].

> "Science aims to give us theories which are empirically adequate; and acceptance of a theory involves as belief only that it is empirically adequate." [13]

The skeptical challenge also leads to a different kind of tension. On one hand all past scientific theories have turned out to be false and therefore it is natural to infer that our best current theories will turn out to be false too — this is the *pessimistic meta-induction argument*. And if all theories are ultimately false, why, then, do we even bother to do science?

On the other hand our current theories are extremely successful. How can this be? It seems that the only possible explanation is that our theories do indeed capture something right about reality — this is the *no-miracles argument*. Indeed, as Putnam puts it:

> "The positive argument for realism is that it is the only philosophy that doesn't make the success of science a miracle." ([19], p.73)

The tension can be resolved through a compromise: our theories are right in some respects, which explains their present success, and wrong in others, which will explain their future failures.

In constructing new theories it would be very helpful if we could know ahead of time which are the features in our current theories that ought to be preserved and which discarded. The position known as *structural realism* claims to have such knowledge. It asserts that what science somehow manages to track faithfully is the mathematical *structure* that describes relations among entities while the intrinsic nature of these entities remains largely unknown. Thus, structural realism (belief in a true description of structure) is intermediate between a full-blown realism (belief in a true description of both structure and ontology) and a skeptical empiricism (belief only in empirical adequacy). Unfortunately, the question of drawing a clear distinction between what constitutes intrinsic nature and what constitutes structure is tricky; it is not even clear that such a distinction can be drawn at all.

There are many other attempted answers — none fully successful yet — and the literature on these topics is enormous. (See e.g. [20], [21].) There are two possible directions of research that are relevant to the theme of this paper. They arise from recognizing the central roles played by the notion of 'truth' (and its related cousin 'reality') and of 'information'. Clearly, different notions of 'true' and 'real' and of 'information' will affect how we decide what is 'true' and what is 'real' and of how 'information' will help us make those decisions.

3 Pragmatic realism

The notion of truth adopted in scientific realism, in structural realism, and in constructive empiricism is a 'correspondence' theory of truth; true statements are claimed to enjoy a special relation to reality: they match it or correspond to it. Unfortunately the precise nature of such correspondence is not clear.

For, in order to say that a description matches reality, we need to compare that description with something, and *for us* this something can only be another description. Reality-in-itself remains forever inaccessible.

Pragmatism represents a break with this notion. There are several notions of pragmatic truth – and they are not free from controversy either. The central element is a concern with the practical effectiveness of ideas as tools for achieving our purposes, a concern with whether we ought to believe in them or not, with their justification. I think there is much here that can be appropriated and suitably modified to do science:

> Truth is a useful idealization. True statements are those that would be fully accepted – that is, believed – by ideally rational agents under ideal epistemic conditions.

And, just as for other conceptual idealizations in science – such as inertial frames, thermal equilibrium, rigid bodies and frictionless planes – the fact that they are never attainable in practice does not in the least detract from their usefulness. In fact, the notion of truth is most useful when left a bit vague. 'True' is the compliment we pay to an idea when we are fairly certain that it is working adequately. 'True' is an idea we can trust. And there need be nothing too permanent about truth either; if we discover that the idea is no longer working so well, then we just withdraw the compliment and say "I was wrong; I thought it was true but it wasn't." We do this all the time. There is no reason why all those features of truth that are embodied in the correspondence model should be preserved; in a truly pragmatic account we just need to preserve what is useful.

In a pragmatic approach all pragmatic virtues, and not just empirical adequacy, contribute to the assessment of truth:[5]

> Science aims to give us theories that are useful for the purposes of description, explanation, prediction, control, etc.; the acceptance of a theory involves the conviction that the theory is indeed useful.

To put it bluntly science does indeed aim to give us theories that are true — but the aim is for *pragmatic* truth. Such truth is objective in the sense that a statement that fails to be empirically adequate — and therefore not useful for predictions, control, etc. — is objectively false. Moreover, being objective does no mean independent of us. The presumably true propositions derive their meaning from the theory or conceptual framework in which they are embedded; a framework that is designed by us for our purposes.

The notion of truth is related to that of reality. Putnam, who rejects the label 'pragmatist' for himself but is nevertheless called a 'neo-pragmatist' by others, calls his position an *internal* realism because the question

> "*...what objects does the world consist of?* is a question that only makes sense to ask *within* a theory or description." [23]

[5] [22] gives a similar pragmatic position which emphasizes explanatory power.

He further explains that

> "In an internalist view ... signs do not intrinsically correspond to objects, independently of how those objects are employed and by whom. But a sign that is actually employed in a particular way by a particular community of users can correspond to particular objects *within the conceptual scheme of those users*. 'Objects' do not exist independently of conceptual schemes. *We* cut up the world into objects when we introduce one or another scheme of description. Since the objects *and* the signs are alike *internal* to the scheme of description, it is possible to say what matches what." [11]

Ironically, realists who favor a correspondence theory of truth would rightfully classify internal realism and any other pragmatic realism as forms of anti-realism.

We can be more explicit: it is we who supply the concepts of chairs and tables but, once the concepts are in place, whether the object in front of me is a table or a chair is a matter of objective fact about which my (subjective) judgements can be (objectively) wrong. It is in this sense that this doctrine is a realism — there *really* is a chair in front of me — but it is pragmatic realism in that the concept of chair is invented by us for our purposes and designs. And atoms are as real as chairs. In the middle ages the question might have been whether atoms were real but today we have a different perspective — the concept of atom has proved to be so undeniably useful that we can safely assert: *'real' is what atoms are.*

Naturally, a pragmatic account that denies that a detailed theory of truth is possible or even necessary will certainly fail to satisfy someone whose interests lie precisely in the development of such a detailed theory. But for those of us whose interests lie in most other fields, such as science, the pragmatic account of truth as a value judgement, as a compliment, may be quite satisfactory — it is all we need. Indeed, elements of pragmatism are common in 20th century physics but this is not always sufficiently emphasized. This is the case, for example, for both Einstein and Bohr despite their otherwise deep and well known disagreements about the nature of physical laws. Einstein's realism is remarkably pragmatic. He deliberately distanced himself from a correspondence notion of truth and was more concerned with the role of truth for the purpose of inference and reasoning:

> "Truth is a quality we attribute to propositions. When we attribute this label to a proposition we accept it for deduction. Deduction and generally the process of reasoning is our tool to bring cohesion to a world of perceptions. The label 'true' is used in such a way that this purpose is served best." (quoted in [24], p.90)

Likewise, there is a close affinity between Bohr and the pragmatism of William James (see [25]):

> "... in our description of nature the purpose is not to disclose the real essence of the phenomena but only to track down, so far as it is possible, relations between the multiple aspects of our experience." ([26], p.18)

and also

> "Owing to the very character of such mathematical abstractions, the formalism [of quantum mechanics] does not allow pictorial representation on accustomed lines, but aims directly at establishing relations between observations obtained under well-defined conditions." ([27], p.71)

a position that is clearly pragmatic and supports structuralism.

4 The pragmatic design of probability theory

Science requires a framework for inference on the basis of incomplete information. Our first task is to show that the quantitative measures of *plausibility* or *degrees of belief* that are the tools for reasoning should be manipulated and calculated using the ordinary rules of the calculus of probabilities — and *therefore* probabilities *can* be interpreted as degrees of belief.

The procedure we follow differs in one remarkable way from the traditional way of setting up physical theories. Normally one starts with the mathematical formalism, and then one proceeds to try to figure out what the formalism might possibly mean; one tries to append an interpretation to it. This is a very difficult problem which has affected statistical physics — what is the meaning of probability and of entropy? — and also quantum theory — what is the meaning of the wave function? Here we proceed in the opposite order. First we decide what we are talking about, degrees of belief or degrees of plausibility (we use the two expressions interchangeably) and then we *design* rules to manipulate them; we design the formalism, we construct it to suit our purposes. The advantage of this pragmatic approach is that the issue of meaning, of interpretation, is settled from the start.

4.1 Rational beliefs

The terms rational and rationality are loaded with preconceived notions. We shall use them with a very limited and technical meaning to be explained below. To start we emphasize that the degrees of belief discussed here are those that would be held by an idealized "rational" agent who would not be subject to the practical limitations under which we humans operate. Humans may hold different beliefs and it is certainly important to figure out what those beliefs might be — perhaps by observing their gambling behavior — but this is not our present concern. Our objective is neither to assess nor to describe the subjective beliefs of any particular individual. Instead we deal with the altogether different but very common problem that arises when we are confused and we want some guidance about what we are *supposed* to believe. Our concern here is not so much with beliefs as they actually are, but rather, with beliefs as they *ought* to be — at least as they ought to be and still deserve to be called *rational*. We are concerned with an ideal standard of rationality that we humans ought to attain at least when discussing scientific matters.

The challenge is that the concept of rationality is notoriously difficult to pin down. One thing we can say is that rational beliefs are constrained beliefs. The essence of rationality lies precisely in the existence of some constraints — not everything goes. We need to identify some *normative criteria of rationality* and the difficulty is to find criteria that are sufficiently general to include all instances of rationally justified belief. Here is our first criterion of rationality:

> *The inference framework must be based on assumptions that have wide appeal and universal applicability.*

Whatever guidelines we pick they must be of general applicability — otherwise they would fail when most needed, namely, when not much is known about a problem. Different rational agents can reason about different topics, or about the same subject but on the basis of different information, and therefore they could hold different beliefs, but they must agree to follow the same rules — and thus wide appeal is necessary. What we seek here are not the specific rules of inference that will apply to this or that specific instance; what we seek is to identify some few features that all instances of rational inference might have in common.

The second criterion is that

> *The inference framework must not be self-refuting.*

It may not be easy to identify criteria of rationality that are sufficiently general and precise. Perhaps we can settle for the more manageable goal of avoiding irrationality in those glaring cases where it is easily recognizable. And this is the approach we take: rather than providing a precise criterion of rationality to be carefully followed, we design a framework with the more modest goal of avoiding some forms of irrationality that are sufficiently obvious to command general agreement. The basic desire is that the web of rational beliefs must avoid inconsistencies. If a conclusion can be reached in two different ways the two ways must agree. As we shall see this requirement turns out to be extremely restrictive.

Finally,

> *The inference framework must be useful in practice — it must allow quantitative analysis.*

Otherwise, why bother? Incidentally, nature might very well transcend a description in terms of a closed set of mathematical formulas. It is not nature that demands a mathematical description; it is the pragmatic demand that our inference schemes — our models — be useful that imposes such a description.

We conclude this brief excursion into rationality with two remarks. First, we have adopted a technical and very limited but pragmatically useful notion of rationality which defines it in terms of avoiding certain obvious irrationalities. Real humans are seldom rational even in this very limited sense. And second, whatever specific design criteria are chosen, one thing must be clear: they are justified on purely pragmatic grounds and therefore they are meant to be only provisional. The design criteria themselves are not immune to change and improvement. Better rational criteria will lead to better scientific theories which will themselves lead to improved criteria and so on. Thus, the method of science is not independent from the contents of science.

4.2 Quantifying rational belief

In order to be useful we require an inference framework that allows quantitative reasoning.[6] The first obvious question concerns the type of quantity that will represent the intensity of beliefs. Discrete categorical variables are not adequate for a theory of general applicability; we need a much more refined scheme.

Do we believe proposition a more or less than proposition b? Are we even justified in comparing propositions a and b? The problem with propositions is not that they cannot be compared but rather that the comparison can be carried out in too many different ways. We can classify propositions according to the degree we believe they are true, their plausibility; or according to the degree that we desire them to be true, their utility; or according to the degree that they happen to bear on a particular issue at hand, their relevance. We can even compare propositions with respect to the minimal number of bits that are required to state them, their description length. The detailed nature of our relations to propositions is too complex to be captured by a single real number. What we claim is that a single real number is sufficient to measure one specific feature, the sheer intensity of rational belief. This should not be too controversial because it amounts to a tautology: an "intensity" is precisely the type of quantity that admits no more qualifications than that of being more intense or less intense; it is captured by a single real number.

However, some preconception about our subject is unavoidable; we need some rough notion that a belief is not the same thing as a desire. But how can we know that we have captured pure belief and not belief contaminated with some hidden desire or something else? Strictly we can't. We hope that our mathematical description captures a sufficiently purified notion of rational belief, and we can claim success only to the extent — again, a pragmatic criterion — that the formalism proves to be useful.

The inference framework will capture two intuitions about rational beliefs. First, we take it to be a defining feature of the intensity of *rational* beliefs that if a is more believable than b, and b more than c, then a is more believable than c. Such transitive rankings can be implemented using real numbers and therefore we are led to claim that

> *Degrees of rational belief (or, as we shall later call them, probabilities) are represented by real numbers.*

Before we proceed further we need to establish some notation. The following choice is standard.

[6] The argument below follows [28]. It is an elaboration of the pioneering work of Cox [9] (see also [10]).

Notation

For every proposition a there exists its negation not-a, which will be denoted $\tilde{a}$. If a is true, then $\tilde{a}$ is false and vice versa.

Given any two propositions a and b the conjunction "a AND b" is denoted ab or $a \wedge b$. The conjunction is true if and only if both a and b are true.

Given a and b the disjunction "a OR b" is denoted by $a \vee b$ or (less often) by $a + b$. The disjunction is true when either a or b or both are true; it is false when both a and b are false.

Typically we want to quantify the degrees of belief in a, $a \vee b$, and ab in the context of some background information expressed in terms of some proposition c in the same universe of discourse as a and b. Such propositions we will write as $a|c$, $a \vee b|c$ and $ab|c$.

The real number that represents the degree of belief in $a|b$ will initially be denoted $[a|b]$ and later in its more standard form $p(a|b)$ and all its variations.

Degrees of rational belief will range from the extreme value v_F that represents certainty that the proposition is false (for example, for any a, $[\tilde{a}|a] = v_F$), to the opposite extreme v_T that represents certainty that the proposition is true (for example, for any a, $[a|a] = v_T$). The transitivity of the ranking scheme implies that there is a single value v_F and a single v_T.

The representation of OR and AND

The inference framework is designed to include a second intuition concerning rational beliefs:

> *In order to be rational our beliefs in $a \vee b$ and ab must be somehow related to our separate beliefs in a and in b.*

Since the goal is to design a quantitative theory, we require that these relations be represented by some functions F and G,

$$[a \vee b|c] = F([a|c], [b|c], [a|bc], [b|ac]) \tag{1}$$

and

$$[ab|c] = G([a|c], [b|c], [a|bc], [b|ac]) \; . \tag{2}$$

Note the *qualitative* nature of this assumption: what is being asserted is the existence of some unspecified functions F and G and not their specific functional forms. The same F and G are meant to apply to all propositions; what is being *designed* is a single inductive scheme of universal applicability. Note further that the arguments of F and G include all four possible degrees of belief in a and b in the context of c and not any potentially questionable subset.[7]

The functions F and G provide a representation of the Boolean operations OR and AND. The requirement that F and G reflect the appropriate associative and distributive properties of the Boolean AND and OR turns out to be

[7] In contrast, [9] sought a representation of AND, $[ab|c] = f([a|c], [b|ac])$, and negation, $[\tilde{a}|c] = g([a|c])$.

extremely constraining. Indeed, we will show that all allowed representations are equivalent to each other and that they are equivalent to probability theory: the associativity of OR requires F to be equivalent to the sum rule for probabilities and the distributivity of AND over OR requires G to be equivalent to the product rule for probabilities.

Our method will be *design by eliminative induction*: now that we have identified a sufficiently broad class of theories — quantitative theories of universal applicability, with degrees of belief represented by real numbers and the operations of conjunction and disjunction represented by functions — we can start weeding the unacceptable ones out.

4.3 The sum rule

Our first goal is to determine the function F that represents OR. The space of functions of four arguments is very large. Without loss of generality we can narrow down the field to propositions a and b that are mutually exclusive in the context of some other proposition d. Thus,

$$[a \vee b|d] = F([a|d],[b|d],v_F,v_F) \ , \tag{3}$$

which effectively restricts F to a function of only two arguments,

$$[a \vee b|d] = F([a|d],[b|d]) \ . \tag{4}$$

The restriction to mutually exclusive propositions does not represent a loss of generality because any two arbitrary propositions can be written as the disjunction of three mutually exclusive ones,

$$a \vee b = [(ab) \vee (a\tilde{b})] \vee [(ab) \vee (\tilde{a}b)] = (ab) \vee (a\tilde{b}) \vee (\tilde{a}b) \ .$$

Therefore, the general rule for the disjunction $a \vee b$ of two arbitrary propositions can be obtained by successive applications of the special rule for mutually exclusive propositions.

4.3.1 The associativity constraint

As a minimum requirement of rationality we demand that the assignment of degrees of belief be consistent: if a degree of belief can be computed in two different ways the two ways must agree. How else could we claim to be rational? All functions F that fail to satisfy this constraint must be discarded.

Consider any three statements a, b, and c that are mutually exclusive in the context of a fourth d. The consistency constraint that follows from the associativity of the Boolean OR,

$$(a \vee b) \vee c = a \vee (b \vee c) \ , \tag{5}$$

is remarkably constraining. It essentially determines the function F. Start from

$$[a \vee b \vee c|d] = F\left([a \vee b|d],[c|d]\right) = F\left([a|d],[b \vee c|d]\right) \ . \tag{6}$$

Use F again for $[a \vee b|d]$ and also for $[b \vee c|d]$, to get

$$F\{F([a|d],[b|d]),[c|d]\} = F\{[a|d],F([b|d],[c|d])\} \ . \tag{7}$$

If we call $[a|d] = x$, $[b|d] = y$, and $[c|d] = z$, then

$$F\{F(x,y),z\} = F\{x,F(y,z)\} \ . \tag{8}$$

Since this must hold for arbitrary choices of the propositions a, b, c, and d, we conclude that *in order to be of universal applicability* the function F must satisfy (8) for arbitrary values of the real numbers (x, y, z). Therefore the function F must be associative.

Remark: The requirement of universality is crucial. Indeed, in a universe of discourse with a discrete and finite set of propositions it is conceivable that the triples (x, y, z) in (8) do not form a dense set and therefore one cannot conclude that the function F must be associative for arbitrary values of x, y, and z. For each specific finite universe of discourse one could design a tailor-made, single-purpose model of inference that could be consistent, i.e. it would satisfy (8), without being equivalent to probability theory. However, we are concerned with designing a theory of inference of universal applicability, a single scheme applicable to *all universes of discourse* whether discrete and finite or otherwise. And the scheme is meant to be used by *all rational agents* irrespective of their state of belief — which need not be discrete. Thus, a framework designed for broad applicability requires that the values of x form a dense set.

4.3.2 The general solution and its regraduation

Equation (8) is a functional equation for F. It is easy to see that there exist an infinite number of solutions. Indeed, by direct substitution one can easily check that eq.(8) is satisfied by any function of the form

$$F(x,y) = \phi^{-1}(\phi(x) + \phi(y) + \beta) \ , \tag{9}$$

where ϕ is an arbitrary invertible function and β is an arbitrary constant. What is not so easy to to show is this is also the *general* solution, that is, given ϕ one can calculate F and, conversely, given any associative F one can calculate the corresponding ϕ. The proof of this result is given in [8][9].

The significance of eq.(9) becomes apparent once it is rewritten as

$$\phi(F(x,y)) = \phi(x) + \phi(y) + \beta \quad \text{or} \quad \phi([a \vee b|d]) = \phi([a|d]) + \phi([b|d]) + \beta. \tag{10}$$

This last form is central to any Cox-type approach to probability theory. Note that there was nothing particularly special about the original representation of degrees of plausibility by the real numbers $[a|d], [b|d], \ldots$ Their only purpose was to provide us with a ranking, an ordering of propositions according to how plausible they are. Since the function $\phi(x)$ is monotonic, the same ordering can be achieved using a new set of positive numbers,

$$\xi(a|d) \stackrel{\text{def}}{=} \phi([a|d]) + \beta, \quad \xi(b|d) \stackrel{\text{def}}{=} \phi([b|d]) + \beta, ... \tag{11}$$

instead of the old. The original and the regraduated scales are equivalent because by virtue of being invertible the function ϕ is monotonic and therefore preserves the ranking of propositions. However, the regraduated scale is much more convenient because, instead of the complicated rule (9), the OR operation is now represented by a much simpler rule, eq.(10),

$$\xi\,(a \vee b|d) = \xi\,(a|d) + \xi\,(b|d) \ , \tag{12}$$

which is just a sum. Thus, the new numbers are neither more nor less correct than the old, they are just considerably more convenient.

Perhaps one can make the logic of regraduation a little bit clearer by considering the somewhat analogous situation of introducing the quantity temperature as a measure of degree of "hotness". Clearly any acceptable measure of "hotness" must reflect its transitivity — if a is hotter than b and b is hotter than c then a is hotter than c — which explains why temperatures are represented by real numbers. But the temperature scales can be quite arbitrary. While many temperature scales may serve equally well the purpose of ordering systems according to their hotness, there is one choice — the absolute or Kelvin scale — that turns out to be considerably more convenient because it simplifies the mathematical formalism. Switching from an arbitrary temperature scale to the Kelvin scale is one instance of a convenient regraduation. ([8], p. 60)

In the old scale, before regraduation, we had set the range of degrees of belief from one extreme of total disbelief, $[\tilde{a}|a] = v_F$, to the other extreme of total certainty, $[a|a] = v_T$. At this point there is not much that we can say about the regraduated $\xi_T = \phi(v_T) + \beta$ but $\xi_F = \phi(v_F) + \beta$ is easy to evaluate. Setting $d = \tilde{a}\tilde{b}$ in eq.(12) gives

$$\xi(a \vee b|\tilde{a}\tilde{b}) = \xi(a|\tilde{a}\tilde{b}) + \xi(b|\tilde{a}\tilde{b}) \Rightarrow \xi_F = 2\xi_F \ , \tag{13}$$

and therefore

$$\xi_F = 0 \ . \tag{14}$$

4.3.3 The general sum rule

As mentioned earlier the restriction to mutually exclusive propositions in the sum rule eq.(12) can be easily lifted noting that for any two *arbitrary* propositions a and b we have

$$a \vee b = (ab) \vee (a\tilde{b}) \vee (\tilde{a}b) = a \vee (\tilde{a}b) \tag{15}$$

Since each of the two terms on the right are mutually exclusive the sum rule (12) applies,

$$\begin{aligned} \xi(a \vee b|d) &= \xi(a|d) + \xi(\tilde{a}b|d) + [\xi(ab|d) - \xi(ab|d)] \\ &= \xi(a|d) + \xi(ab \vee \tilde{a}b|d) - \xi(ab|d) \ , \end{aligned} \tag{16}$$

which leads to the general sum rule,

$$\xi(a \vee b|d) = \xi(a|d) + \xi(b|d) - \xi(ab|d) \ . \tag{17}$$

4.4 The product rule

Next we consider the function G in eq.(2) that represents AND. Once the original plausibilities are regraduated by ϕ according to eq.(11), the new function G for the plausibility of a conjunction reads

$$\xi(ab|c) = G[\xi(a|c), \xi(b|c), \xi(a|bc), \xi(b|ac)] \ . \tag{18}$$

The space of functions of four arguments is very large so we first narrow it down to just two. Then we require that the representation of AND be compatible with the representation of OR that we have just obtained. This amounts to imposing a consistency constraint that follows from the distributive properties of the Boolean AND and OR. A final trivial regraduation yields the product rule of probability theory.

The derivation proceeds in two steps. First, we separately consider special cases where the function G depends on only two arguments, then three, and finally all four arguments. Using commutivity, $ab = ba$, the number of possibilities can be reduced to seven:

$$\begin{aligned}
\xi(ab|c) &= G^{(1)}[\xi(a|c), \xi(b|c)] \\
\xi(ab|c) &= G^{(2)}[\xi(a|c), \xi(a|bc)] \\
\xi(ab|c) &= G^{(3)}[\xi(a|c), \xi(b|ac)] \\
\xi(ab|c) &= G^{(4)}[\xi(a|bc), \xi(b|ac)] \\
\xi(ab|c) &= G^{(5)}[\xi(a|c), \xi(b|c), \xi(a|bc)] \\
\xi(ab|c) &= G^{(6)}[\xi(a|c), \xi(a|bc), \xi(b|ac)] \\
\xi(ab|c) &= G^{(7)}[\xi(a|c), \xi(b|c), \xi(a|bc), \xi(b|ac)]
\end{aligned}$$

It is rather straightforward to show [28][8] that the only functions G that are viable candidates for a general theory of inductive inference are equivalent to type $G^{(3)}$,

$$\xi(ab|c) = G[\xi(a|c), \xi(b|ac)] \ . \tag{19}$$

The AND function G will be determined by requiring that it be compatible with the regraduated OR function F, which is just a sum. Consider three statements a, b, and c, where the last two are mutually exclusive, in the context of a fourth, d. Distributivity of AND over OR,

$$a\,(b \vee c) = ab \vee ac \ , \tag{20}$$

implies that $\xi\left(a\,(b \vee c)\,|d\right)$ can be computed in two ways,

$$\xi\left(a\,(b \vee c)\,|d\right) = \xi\left((ab|d) \vee (ac|d)\right) \ . \tag{21}$$

Using eq.(12) and (19) leads to

$$G[\xi\,(a|d)\,, \xi\,(b|ad) + \xi\,(c|ad)] = G[\xi\,(a|d)\,, \xi\,(b|ad)] + G[\xi\,(a|d)\,, \xi\,(c|ad)] \ ,$$

which we rewrite as

$$G\left(u, v+w\right) = G\left(u, v\right) + G\left(u, w\right) \ , \tag{22}$$

where $\xi\left(a|d\right) = u$, $\xi\left(b|ad\right) = v$, and $\xi\left(c|ad\right) = w$.

To solve the functional equation (22) we first transform it into a differential equation. Differentiate with respect to v and w,

$$\frac{\partial^2 G\left(u, v+w\right)}{\partial v \partial w} = 0 \ , \tag{23}$$

and let $v + w = z$, to get

$$\frac{\partial^2 G\left(u, z\right)}{\partial z^2} = 0 \ , \tag{24}$$

which shows that G is linear in its second argument,

$$G(u, v) = A(u)v + B(u) \ . \tag{25}$$

Substituting back into eq.(22) gives $B(u) = 0$. To determine the function $A(u)$ we note that the degree to which we believe in $ad|d$ is exactly the degree to which we believe in $a|d$ by itself. Therefore,

$$\xi(a|d) = \xi(ad|d) = G[\xi(a|d), \xi(d|ad)] = G[\xi(a|d), \xi_T] \ , \tag{26}$$

where ξ_T denotes complete certainty. Equivalently,

$$u = A(u)\xi_T \Rightarrow A(u) = \frac{u}{\xi_T} \ . \tag{27}$$

Therefore,

$$G\left(u, v\right) = \frac{uv}{\xi_T} \quad \text{or} \quad \frac{\xi\left(ab|d\right)}{\xi_T} = \frac{\xi\left(a|d\right)}{\xi_T}\frac{\xi\left(b|ad\right)}{\xi_T} \ . \tag{28}$$

The constant ξ_T is easily regraduated away: just normalize ξ to $p = \xi/\xi_T$. The corresponding regraduation of the sum rule, eq.(17) is equally trivial. The degrees of belief ξ range from total disbelief $\xi_F = 0$ to total certainty ξ_T. The corresponding regraduated values are $p_F = 0$ and $p_T = 1$.

4.5 Probabilities

In the regraduated scale the AND operation is represented by a simple product rule,

$$p\left(ab|d\right) = p\left(a|d\right) p\left(b|ad\right) \ , \tag{29}$$

and the OR operation is represented by the sum rule,

$$p\left(a \vee b|d\right) = p\left(a|d\right) + p\left(b|d\right) - p(ab|d) \ . \tag{30}$$

Degrees of belief p measured in this particularly convenient regraduated scale will be called "probabilities". The degrees of belief p range from total disbelief $p_F = 0$ to total certainty $p_T = 1$.
To summarize:

> *A state of partial knowledge — a web of interconnected rational beliefs — is mathematically represented by quantities that are to be manipulated according to the rules of probability theory. Degrees of rational belief are probabilities.*

Other equivalent representations are possible but less convenient; the choice is made on purely pragmatic grounds.

On meaning, ignorance and randomness

The product and sum rules can be used as the starting point for a theory of probability: Quite independently of what probabilities could possibly mean, we can develop a formalism of real numbers (measures) that are manipulated according to eqs.(29) and (30). This is the approach taken by Kolmogorov. The advantage is mathematical clarity and rigor. The disadvantage, of course, is that in actual applications the issue of meaning, of interpretation, turns out to be important because it affects how and why probabilities are used. It affects how one sets up the equations and it even affects our perception of what counts as a solution.

The advantage of the approach described above is that the issue of meaning is clarified from the start: the theory was designed to apply to degrees of belief. Consistency requires that these numbers be manipulated according to the rules of probability theory. This is all we need. There is no reference to measures of sets or large ensembles of trials or even to random variables. This is remarkable: it means that we can apply the powerful methods of probability theory to reasoning about problems where nothing random is going on, and to single events for which the notion of an ensemble is either absurd or at best highly contrived and artificial. Thus, probability theory is *the* method for consistent reasoning in situations where the information available might be insufficient to reach certainty: probability is *the* tool for coping with uncertainty and ignorance.

The degree-of-belief interpretation can also be applied to the probabilities of random variables. It may, of course, happen that there is an unknown influence that affects a system in unpredictable ways and that there is a good reason why this influence remains unknown, namely, that it is so complicated that the information necessary to characterize it cannot be supplied. Such an influence we can call 'random'. Thus, being random is just one among many possible reasons why a quantity might be uncertain or unknown.

5 What is information?

The term 'information' is used with a wide variety of different meanings (see *e.g.*, [1] [8] [10] [29] [30] [31]). There is the Shannon notion of information, which is meant to measure an amount of information and is quite divorced from semantics. There is also an algorithmic notion of information, which captures the notion of complexity and originates in the work of Solomonov, Kolmogorov and Chaitin. Here we develop an epistemic notion of information that is somewhat closer to the everyday colloquial use of the term — roughly, information is what we seek when we ask a question.

It is not unusual to hear that systems "carry" or "contain" information or that "information is physical". This mode of expression can perhaps be traced to the origins of information theory in Shannon's theory of communication. We say that we have received information when among the vast variety of messages that could conceivably have been generated by a distant source, we discover which particular message was actually sent. It is thus that the message "carries" information. The analogy with physics is immediate: the set of all possible states of a physical system can be likened to the set of all possible messages, and the actual state of the system corresponds to the message that was actually sent. Thus, the system "conveys" a message: the system "carries" information about its own state. Sometimes the message might be difficult to read, but it is there nonetheless.

This language — information is physical — useful as it has turned out to be, does not exhaust the meaning of the word 'information'. The goal of Shannon's information theory, or better, communication theory, is to characterize the sources of information, to measure the capacity of the communication channels, and to learn how to control the degrading effects of noise. It is somewhat ironic but nevertheless true that this "information" theory is unconcerned with the central Bayesian issue of how messages affect the beliefs of rational agents.

A fully Bayesian information theory demands an explicit account of the relation between information and the beliefs of ideally rational agents. The connection arises as follows. Implicit in the recognition that most of our beliefs are held on the basis of incomplete information is the idea that our beliefs would be "better" if only we had more information. Indeed, it is a presupposition of thought itself that some beliefs are better than others — otherwise why go through the trouble of thinking? Therefore a theory of probability demands a theory for updating probabilities.

The concern with 'good' and 'better' bears on the issue of whether probabilities are subjective, objective, or somewhere in between. We can argue that what makes one probability assignment better than another is that it better reflects something "objective" about the world. The adoption of better beliefs has real consequences: they provide a better guidance about how to cope with the world, and in this pragmatic sense, they provide a better guide to the "truth". Probabilities are useful to the extent that they incorporate some degree of objectivity.

On the other hand all probability assignments involve judgements and therefore some subjectivity is unavoidable. The long controversy over the objective or subjective nature of probabilities arises from accepting the sharp dichotomy that they are either one or the other with no room for intermediate positions. The Gordian knot is cut by simply declaring that *the dichotomy is false*. What we have is something like a spectrum. Some subjectivity is inevitable but objectivity is the desirable goal. The procedure for enhancing objectivity is through appropriate updating mechanisms that allow us to process information and incorporate its objective features into our beliefs. Barring some pathological cases Bayes' rule behaves precisely in this way. Indeed, as more and more data are taken into account the original (possibly subjective) prior becomes less and less relevant, and all rational agents become more and more convinced of the *same* truth. This is crucial: were it not this way Bayesian reasoning would not be deemed acceptable.

To set the stage for the discussion below assume that we have received a message — but the carrier of information could equally well have been input from our senses or data from an experiment. If the message agrees with our prior beliefs we can safely ignore it. The message is boring; it carries no news; literally, for us it carries no information. The interesting situation arises when the message surprises us; it is not what we expected. A message that disagrees with our prior beliefs presents us with a problem that demands a decision. If the source of the message is not deemed reliable then the contents of the message can be safely ignored — it carries no information; it is no different from noise. On the other hand, if the source of the message is deemed reliable then we have an opportunity to improve our beliefs — we ought to update our beliefs to agree with the message. Choosing between these two options requires a decision, a judgement. The message (or the sensation, or the data) becomes "information" precisely at that moment when as a result of our evaluation we feel that our beliefs require revision.

We are now ready to address the question: What, after all, is 'information'? The main observation is that the result of being confronted with new information is to restrict our options as to what we are honestly and rationally allowed to believe. This, I propose, is the defining characteristic of information.

> *Information, in its most general form, is whatever affects and therefore constrains rational beliefs.*

Since our objective is to update from a prior distribution to a posterior when *new* information becomes available we can state that

> *New information is what forces a change of rational beliefs.*

New information is a set of constraints on the family of acceptable posterior distributions. Our definition captures an idea of information that is directly related to changing our (rational) minds: information is the driving force behind the process of learning. Incidentally, note that although we did not find it necessary to talk about amounts of information, whether measured in units of bits or otherwise, our notion of information will allow precise quantitative calculations. Indeed, constraints on the acceptable posteriors are precisely the kind of information the method of maximum entropy (to be developed below) is designed to handle.

An important aspect of this epistemic notion of information is that the identification of what qualifies as information — as opposed to mere noise — already involves a judgement, an evaluation; it is a matter of facts as much as a matter of values. Furthermore, once a certain proposition has been identified as information, the revision of beliefs acquires a moral component; it is no longer optional: it becomes a moral imperative.

The act of updating is a type of dynamics — the study of change. In Newtonian dynamics the state of motion of a system is described in terms of its momentum — the "quantity" of motion — while the change from one state to another is explained in terms of an applied force. Similarly, a state of belief is described in terms of probabilities — a "quantity" of belief — and the change from one state to another is due to information. Just as a force or an impulse is that which induces a change from one state of motion to another, so *information is that which induces a change from one state of belief to another*. Updating is a form of dynamics — and vice versa: in [7] quantum dynamics is derived as an updating of probabilities subject to the appropriate information/constraints.

What about prejudices and superstitions? What about divine revelations? Do they constitute information? Perhaps they lie outside our restriction to the beliefs of *ideally rational agents*, but to the extent that their effects are indistinguishable from those of other sorts of information, namely, they affect beliefs, they should qualify as information too. False information is information too. Assessing whether the sources of such information are reliable or not can be a difficult problem. In fact, even ideally rational agents can be affected by false information because the evaluation that assures them that the data was competently collected or that the message originated from a reliable source involves an act of judgement that is not completely infallible. Strictly, all judgements that constitute the necessary first step of one inference process, are themselves the end result of a previous inference process that is not immune from uncertainty.

What about limitations in our computational power? Such practical limitations are unavoidable and they do influence our inferences. Should they be considered information? No. Limited computational resources may affect the numerical approximation to the value of, say, an integral, but they do not affect the actual value of the integral. Similarly, limited computational resources may affect the approximate imperfect reasoning of real agents and real computers but they do not affect the reasoning of those ideal rational agents that are the subject of our present concerns.

6 The design of entropic inference

Once we have decided, as a result of the confrontation of new information with old beliefs, that our beliefs require revision the problem becomes one of deciding how precisely this ought to be done. First we identify some general features of the kind of belief revision that one might consider desirable, of the kind of belief revision that one might count as rational. Then we design a method, a systematic procedure, that implements those features. To the extent that the method performs as desired we can claim success. The point is not that success derives from our method having achieved some intimate connection to the inner wheels of reality; success just means that the method seems to be working. Whatever criteria of rationality we choose, they are meant to be only provisional — they are not immune from further change and improvement.

Typically the new information will not affect our beliefs in just one proposition — in which case the updating would be trivial. Tensions immediately arise because the beliefs in various propositions are not independent; they are interconnected by demands of consistency — the sum and product rules we derived earlier. Therefore the new information also affects our beliefs in all those "neighboring" propositions that are directly linked to it, and these in turn affect their neighbors, and so on. The effect can potentially spread over the whole network of beliefs; it is the whole web of beliefs that must be revised.

The one obvious requirement is that the updated beliefs ought to agree with the newly acquired information. Unfortunately, this requirement, while necessary, is not sufficiently restrictive: we can update in many ways that preserve both internal consistency and consistency with the new information. Additional criteria are needed. What rules is it rational to choose?

6.1 General criteria

The rules are motivated by the same pragmatic design criteria that motivate the design of probability theory itself — universality, consistency, and practical utility. But this is admittedly too vague; we must be more specific about the precise way in which they are implemented.

6.1.1 Universality

The goal is to design a method for induction, for reasoning when not much is known. In order for the method to perform its function we must impose that it be of *universal* applicability. Consider the alternative: We could design methods that are problem-specific, and employ different induction methods for different problems. Such a framework, unfortunately, would fail us precisely when we need it most, namely, in those situations where the information available is so incomplete that we do not know which method to employ.

We can argue this point somewhat differently: It is quite conceivable that different situations could require different problem-specific induction methods. What we want to design here is a general-purpose method that captures what all those problem-specific methods have in common.

6.1.2 Parsimony

To specify the updating we adopt a very conservative criterion that recognizes the value of information: what has been laboriously learned in the past is valuable and should not be disregarded unless rendered obsolete by new information. The only aspects of one's beliefs that should be updated are those for which new evidence has been supplied. Thus we adopt a

Principle of Minimal Updating: *Beliefs should be updated only to the extent required by the new information.*

The special case of updating in the absence of new information deserves special attention. It states that when there is no new information an ideally rational agent should not change its mind.[8] In fact, it is difficult to imagine any notion of rationality that would allow the possibility of changing one's mind for no apparent reason. This is important and it is worthwhile to consider it from a different angle. Degrees of belief, probabilities, are said to be subjective: two different individuals might not share the same beliefs and could conceivably assign probabilities differently. But subjectivity does not mean arbitrariness. It is not a blank check allowing the rational agent to change its mind for no good reason.

Minimal updating offers yet another pragmatic advantage. As we shall see below, rather than identifying what features of a distribution are singled out for updating and then specifying the detailed nature of the update, we will adopt design criteria that stipulate what is not to be updated. The practical advantage of this approach is that it enhances objectivity — there are many ways to change something but only one way to keep it the same.

The analogy with mechanics can be pursued further: if updating is a form of dynamics, then minimal updating is a form of inertia. Rationality and objectivity demand a considerable amount of inertia.

6.1.3 Independence

The next general requirement turns out to be crucially important: without it the very possibility of scientific theories would not be possible. The point is that in every scientific model, whatever the topic, if it is to be useful at all, we must assume that all relevant variables have been taken into account and that whatever was left out — the rest of the universe — does not matter. To put it another way: in order to do science we must be able to understand parts of the universe without having to understand the universe as a whole. Granted, it is not necessary that the understanding be complete and exact; it must just be adequate for our purposes.

The assumption, then, is that it is possible to focus our attention on a suitably chosen system of interest and neglect the rest of the universe because they

[8] We refer to ideally rational agents who have fully processed all information acquired in the past. Humans do not normally behave this way; they often change their minds by processes that are not fully conscious.

are "sufficiently independent". Thus, in any form of science the notion of statistical independence must play a central and privileged role. This idea — that some things can be neglected, that not everything matters — is implemented by imposing a criterion that tells us how to handle independent systems. The requirement is quite natural: *Whenever two systems are a priori believed to be independent and we receive information about one it should not matter if the other is included in the analysis or not.* This amounts to requiring that independence be preserved unless information about correlations is explicitly introduced.[9]

Again we emphasize: none of these criteria are imposed by Nature. They are desirable for pragmatic reasons; they are imposed by design.

6.2 Entropy as a tool for updating probabilities

Consider a variable x the value of which is uncertain; x can be discrete or continuous, in one or in several dimensions. It could, for example, represent the possible microstates of a physical system, a point in phase space, or an appropriate set of quantum numbers. The uncertainty about x is described by a probability distribution $q(x)$. Our goal is to update from the prior distribution $q(x)$ to a posterior distribution $P(x)$ when new information becomes available. The information is in the form of a set of constraints that defines a family $\{p(x)\}$ of acceptable distributions and the question is: which distribution $P \in \{p\}$ should we select?

Our goal is to design a method that allows a systematic search for the preferred posterior distribution. The central idea, first proposed in [16],[10] is disarmingly simple: to select the posterior first rank all candidate distributions in increasing *order of preference* and then pick the distribution that ranks the highest. Irrespective of what it is that makes one distribution preferable over another (we will get to that soon enough) it is clear that any ranking according to preference must be transitive: if distribution p_1 is preferred over distribution p_2, and p_2 is preferred over p_3, then p_1 is preferred over p_3. Such transitive rankings are implemented by assigning to each $p(x)$ a real number $S[p]$ in such a way that if p_1 is preferred over p_2, then $S[p_1] > S[p_2]$. The functional $S[p]$ will be called the entropy of p. The selected distribution (one or possibly many, for there may be several equally preferred distributions) is that which maximizes the entropy functional.

The importance of this particular approach to updating distributions cannot be overestimated: it implies that the updating method will take the form of a variational principle — the method of Maximum Entropy (ME) — involving a certain functional — the entropy — that maps distributions to real numbers and

[9] The independence requirement is rather subtle and one must be careful about its precise implementation. The robustness of the design is shown by exhibiting an alternative version that takes the form of a consistency constraint: *Whenever systems are known to be independent it should not matter whether the analysis treats them jointly or separately.* [8][18]

[10] [16] deals with the more general problem of ranking positive additive distributions which also include, *e.g.*, intensity distributions.

that is designed to be maximized. *These features are not imposed by Nature; they are all imposed by design.* They are dictated by the function that the ME method is supposed to perform. (Thus, it makes no sense to seek a generalization in which entropy is a complex number or a vector; such a generalized entropy would just not perform the desired function.)

Next we specify the ranking scheme, that is, we choose a specific functional form for the entropy $S[p]$. Note that *the purpose of the method is to update from priors to posteriors* so the ranking scheme must depend on the particular prior q and therefore the entropy S must be a functional of both p and q. The entropy $S[p, q]$ describes a ranking of the distributions p *relative* to the given prior q. $S[p, q]$ is the entropy of p *relative* to q, and accordingly $S[p, q]$ is commonly called a *relative entropy*. This is appropriate and sometimes we will follow this practice. However, since all entropies are relative, even when relative to a uniform distribution, the qualifier 'relative' is redundant and can be dropped.

The functional $S[p, q]$ is designed by a process of elimination — a process of *eliminative induction*. First we state the desired design criteria; this is the crucial step that defines what makes one distribution preferable over another. Then we analyze how each criterion constrains the form of the entropy. As we shall see the design criteria adopted below are sufficiently constraining that there is a single entropy functional $S[p, q]$ that survives the process of elimination.

This approach has a number of virtues. First, to the extent that the design criteria are universally desirable, then the single surviving entropy functional will be of universal applicability too. Second, the reason why alternative entropy candidates are eliminated is quite explicit — at least one of the design criteria is violated. Thus, *the justification behind the single surviving entropy is not that it leads to demonstrably correct inferences, but rather, that all other candidate entropies demonstrably fail to perform as desired.*

6.3 Specific design criteria

Three criteria and their consequences for the functional form of the entropy are given below. Proofs are given in [8].

6.3.1 Locality

DC1 *Local information has local effects.*

Suppose the information to be processed does *not* refer to a particular subdomain $\mathcal{D}$ of the space $\mathcal{X}$ of xs. In the absence of any new information about $\mathcal{D}$ the PMU demands we do not change our minds about probabilities that are conditional on $\mathcal{D}$. Thus, we design the inference method so that $q(x|\mathcal{D})$, the prior probability of x conditional on $x \in \mathcal{D}$, is not updated. The selected conditional posterior is[11]

$$P(x|\mathcal{D}) = q(x|\mathcal{D}) \ . \tag{31}$$

We emphasize: the point is not that we make the unwarranted assumption that keeping $q(x|\mathcal{D})$ unchanged is guaranteed to lead to correct inferences. It need not; induction is risky. The point is, rather, that in the absence of any evidence to the contrary there is no reason to change our minds and the prior information takes precedence.

The consequence of DC1 is that non-overlapping domains of x contribute additively to the entropy,

$$S[p,q] = \int dx \, F\left(p(x), q(x), x\right) \ , \tag{32}$$

where F is some unknown function — not a functional, just a regular function of three arguments.

Comment:

If the variable x is continuous the criterion DC1 requires that information that refers to points infinitely close but just outside the domain $\mathcal{D}$ will have no influence on probabilities conditional on $\mathcal{D}$. This may seem surprising as it may lead to updated probability distributions that are discontinuous. Is this a problem? No.

In certain situations (*e.g.*, physics) we might have explicit reasons to believe that conditions of continuity or differentiability should be imposed and this information might be given to us in a variety of ways. The crucial point, however — and this is a point that we keep and will keep reiterating — is that unless such information is in fact explicitly given we should not assume it. If the new information leads to discontinuities, so be it.

Comment: Bayes' rule

The locality criterion DC1 includes Bayesian conditionalization as a special case. Indeed, if the information is given through the constraint $p(\mathcal{D}) = 1$ — or more precisely $p(\tilde{\mathcal{D}}) = 0$ where $\tilde{\mathcal{D}}$ is the complement of $\mathcal{D}$ so that the information does not directly refer to $\mathcal{D}$ — then $P(x|\mathcal{D}) = q(x|\mathcal{D})$, which is known as Bayesian conditionalization. More explicitly, if θ is the variable to be inferred on the basis of information about a likelihood function $q(x|\theta)$ and observed data x', then the update from the prior q to the posterior P,

$$q(x,\theta) = q(x)q(\theta|x) \to P(x,\theta) = P(x)P(\theta|x) \tag{33}$$

consists of updating $q(x) \to P(x) = \delta(x - x')$ to agree with the new data and invoking the PMU so that $P(\theta|x') = q(\theta|x')$ remains unchanged. Therefore,

$$P(x,\theta) = \delta(x - x')q(\theta|x) \ . \tag{34}$$

Marginalizing over x gives

$$P(\theta) = q(\theta|x') = q(\theta)\frac{q(x'|\theta)}{q(x')} \ , \tag{35}$$

[11] We denote priors by q, candidate posteriors by lower case p, and the selected posterior by upper case P.

which is Bayes' rule. Thus, *entropic inference is designed to include Bayesian inference as a special case.* Note however that imposing locality is not identical to imposing Bayesian conditionalization — locality is more general because it is not restricted to absolute certainties such as $p(\mathcal{D}) = 1$.

6.3.2 Coordinate invariance

DC2 *The system of coordinates carries no information.*

The points $x \in \mathcal{X}$ can be labeled using any of a variety of coordinate systems. In certain situations we might have explicit reasons to believe that a particular choice of coordinates should be preferred over others and this information might have been given to us in a variety of ways, but unless it was in fact given we should not assume it: the ranking of probability distributions should not depend on the coordinates used.

The consequence of DC2 is that $S[p,q]$ can be written in terms of coordinate invariants such as $dx\, m(x)$ and $p(x)/m(x)$, and $q(x)/m(x)$:

$$S[p,q] = \int dx\, m(x)\Phi\left(\frac{p(x)}{m(x)}, \frac{q(x)}{m(x)}\right) \ . \tag{36}$$

Thus the single unknown function F which had three arguments has been replaced by two unknown functions: Φ which has two arguments, and the density $m(x)$.

To grasp the meaning of DC2 it may be useful to recall some facts about coordinate transformations. Consider a change from old coordinates x to new coordinates x' such that $x = \Gamma(x')$. The new volume element dx' includes the corresponding Jacobian,

$$dx = \gamma(x')dx' \quad \text{where} \quad \gamma(x') = \left|\frac{\partial x}{\partial x'}\right|. \tag{37}$$

Let $m(x)$ be any density; the transformed density $m'(x')$ is such that $m(x)dx = m'(x')dx'$. This is true, in particular, for probability densities such as $p(x)$ and $q(x)$, therefore

$$m(x) = \frac{m'(x')}{\gamma(x')} \ , \quad p(x) = \frac{p'(x')}{\gamma(x')} \quad \text{and} \quad q(x) = \frac{q'(x')}{\gamma(x')} \, . \tag{38}$$

The coordinate transformation gives

$$\begin{aligned} S[p,q] &= \int dx\, F\left(p(x), q(x), x\right) \\ &= \int \gamma(x')dx'\, F\left(\frac{p'(x')}{\gamma(x')}, \frac{q'(x')}{\gamma(x')}, \Gamma(x')\right), \end{aligned} \tag{39}$$

which is a mere change of variables. The identity above is valid always, for all Γ and for all F; it imposes absolutely no constraints on $S[p,q]$. The real constraint arises from realizing that we could have *started* in the x' coordinate frame, in which case we would have ranked the distributions using the entropy

$$S[p',q'] = \int dx' \, F\left(p'(x'), q'(x'), x'\right) \, , \tag{40}$$

but this should have no effect on our conclusions. This is the nontrivial content of DC2. It is not that we can change variables, we can always do that; but rather that the two rankings, the one according to $S[p,q]$ and the other according to $S[p',q']$ must coincide. This requirement is satisfied if, for example, $S[p,q]$ and $S[p',q']$ turn out to be numerically equal, but this is not necessary.

6.3.3 Locality (again)

Next we determine the density $m(x)$ by invoking the locality criterion DC1 once again. A situation in which no new information is available is dealt by allowing the domain $\mathcal{D}$ to cover the whole space of xs, $\mathcal{D} = \mathcal{X}$ and DC1 requires that in the absence of any new information the prior conditional probabilities should not be updated, $P(x|\mathcal{X}) = q(x|\mathcal{X})$ or $P(x) = q(x)$. Thus, when there are no constraints the selected posterior distribution should coincide with the prior distribution, which is expressed as

DC1′ *When there is no new information there is no reason to change one's mind and one shouldn't.*

The consequence of DC1′ (a second use of locality) is that the arbitrariness in the density $m(x)$ is removed: up to normalization $m(x)$ must be the prior distribution $q(x)$, and therefore at this point we have succeeded in restricting the entropy to functionals of the form

$$S[p,q] = \int dx \, q(x) \Phi\left(\frac{p(x)}{q(x)}\right) \, . \tag{41}$$

6.3.4 Independence

DC3 *When two systems are a priori believed to be independent and we receive independent information about each then it should not matter whether one is included in the analysis of the other or not.*

Consider a composite system, $x = (x_1, x_2) \in \mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2$. Assume that all prior evidence led us to believe the systems were independent. This belief is expressed through the prior distribution: if the individual priors are $q_1(x_1)$ and $q_2(x_2)$, then the prior for the whole system is $q_1(x_1)q_2(x_2)$. Further suppose that new information is acquired such that $q_1(x_1)$ would by itself be updated to $P_1(x_1)$ and that $q_2(x_2)$ would be itself be updated to $P_2(x_2)$. DC3 requires that $S[p,q]$ be such that the joint prior $q_1(x_1)q_2(x_2)$ updates to the product $P_1(x_1)P_2(x_2)$ so that inferences about one system do not affect inferences about the other.

The consequence of DC3 is that the remaining unknown function Φ is determined to be $\Phi(z) = -z \log z$. Thus, probability distributions $p(x)$ should be ranked relative to the prior $q(x)$ according to their relative entropy,

$$S[p, q] = -\int dx\, p(x) \log \frac{p(x)}{q(x)}. \tag{42}$$

Comment:

We emphasize that the point is not that when we have no evidence for correlations we draw the firm conclusion that the systems must necessarily be independent. They could indeed have turned out to be correlated and then our inferences would be wrong. Induction involves risk. The point is rather that if the joint prior reflected independence and the new evidence is silent on the matter of correlations, then the prior takes precedence and there is no reason to change our minds. This is parsimony in action: a feature of the probability distribution — in this case, independence — will not be updated unless the evidence requires it.

6.4 The ME method

At this point our conclusions are summarized as follows:

The ME method: *We want to update from a prior distribution q to a posterior distribution when there is new information in the form of constraints $\mathcal{C}$ that specify a family $\{p\}$ of allowed posteriors. The posterior is selected through a ranking scheme that recognizes the value of prior information, the irrelevance of choice of coordinates, and the privileged role of independence. Within the family $\{p\}$ the preferred posterior P is that which maximizes the relative entropy $S[p, q]$ subject to the available constraints. No interpretation for $S[p, q]$ is given and none is needed.*

We emphasize that the logic behind the updating procedure does not rely on any particular meaning assigned to the entropy, either in terms of information, or heat, or disorder. Entropy is merely a tool for inductive inference; we do not need to know what entropy means; we only need to know how to use it.

The derivation above has singled out *a unique $S[p, q]$ to be used in inductive inference*. Other "entropies" such as those associated with the names of Renyi or Tsallis might turn out to be useful for other purposes — perhaps as measures of some kinds of information, or measures of discrimination or distinguishability among distributions, or of ecological diversity, or for some altogether different function — but they are unsatisfactory for the purpose of updating in the sense that they do not perform according to the design criteria DC1-3.

6.5 Deviations from maximum entropy

There is one last issue that must be addressed before one can claim that the design of the method of entropic inference is more or less complete. Higher entropy represents higher preference but there is nothing in the previous arguments to tell us by how much. Suppose the maximum of the entropy function is not particularly sharp, are we really confident that distributions that are ranked close to the maximum are totally ruled out? We want a quantitative measure of the extent to which distributions with lower entropy are ruled out. The discussion below follows [32].

The problem is to update from a prior $q(x)$ given information specified by certain constraints $\mathcal{C}$. The constraints $\mathcal{C}$ specify a family of candidate distributions $p(x) = p(x|\theta)$ which can be conveniently labelled with some finite number of parameters θ^i, $i = 1 \dots n$. Thus, the parameters θ are coordinates on a statistical manifold Θ_n specified by $\mathcal{C}$. The distributions in this manifold are ranked according to their entropy $S[p, q] = S(\theta)$ and the preferred posterior is the distribution $p(x|\theta_0)$ that maximizes the entropy $S(\theta)$.

The question we now address concerns the extent to which $p(x|\theta_0)$ should be preferred over other distributions with lower entropy or, to put it differently: To what extent is it rational to believe that the selected value ought to be the entropy maximum θ_0 rather than any other value θ? This is a question about the probability $p(\theta)$ of various values of θ.

The original problem which led us to design the ME method was to assign a probability to x; we now see that the full problem is to assign probabilities to both x and θ. We are concerned not just with $p(x)$ but rather with the joint distribution $p_J(x, \theta)$; the universe of discourse has been expanded from $\mathcal{X}$ (the space of xs) to the product space $\mathcal{X} \times \Theta_n$ (the space of xs and θs).

To determine the joint distribution $p_J(x, \theta)$ we make use of essentially the only method at our disposal — the ME method itself — but this requires that we address the two standard preliminary questions: first, what is the prior distribution, what do we know about x and θ before we receive the information in the constraints $\mathcal{C}$? And second, how do handle this new information $\mathcal{C}$ that constrains the allowed $p_J(x, \theta)$?

This first question is the subtler one: when we know absolutely nothing about the θs we do not know how they are related to the xs, and we know neither the constraints $\mathcal{C}$ nor the space Θ_n they determine. At best we just know that the θs are points in some unspecified space Θ_N of sufficiently large dimension N. A joint prior that reflects this state of ignorance is a product, $q_J(x, \theta) = q(x)\mu(\theta)$. We will assume that the prior over x is known — it is the same prior we had used when we updated from $q(x)$ to $p(x|\theta_0)$. We will also assume that $\mu(\theta)$ represents a uniform distribution, that is, it assigns equal probabilities to equal volumes in Θ_N.

Next we tackle the second question: what are the constraints on the allowed joint distributions $p_J(x, \theta) = p(\theta)p(x|\theta)$? The new information is that the θs are not any arbitrary points in some unspecified large space Θ_N but are instead constrained to lie on the smaller subspace Θ_n that represents those distributions $p(x|\theta)$ satisfying the constraints $\mathcal{C}$. This space Θ_n is a statistical manifold and there exists a natural measure of distance given by the information metric g_{ij}. The corresponding volume elements are given by $g_n^{1/2}(\theta)d^n\theta$, where $g_n(\theta)$ is the determinant of the metric [8] [33]. Therefore, on the constraint manifold Θ_n the uniform prior for θ is proportional to $g_n^{1/2}(\theta)$ and the corresponding joint prior is $q_J(x,\theta) = q(x)g_n^{1/2}(\theta)$.

To select the preferred joint distribution $P(x,\theta)$ we maximize the joint entropy $\mathcal{S}[p_J, q_J]$ over all distributions of the form $p_J(x,\theta) = p(\theta)p(x|\theta)$ with $\theta \in \Theta_n$. It is convenient to write the joint entropy as

$$\begin{aligned} \mathcal{S}[p_J, q_J] &= -\int_{\mathcal{X}\times\Theta_n} dx\, d^n\theta\, p(\theta)p(x|\theta)\, \log \frac{p(\theta)p(x|\theta)}{g_n^{1/2}(\theta)q(x)} \\ &= -\int_{\Theta_n} d^n\theta\, p(\theta) \log \frac{p(\theta)}{g_n^{1/2}(\theta)} + \int_{\Theta_n} d^n\theta\, p(\theta) S(\theta), \end{aligned} \tag{43}$$

where

$$S[p,q] = S(\theta) = -\int_{\mathcal{X}} dx\, p(x|\theta) \log \frac{p(x|\theta)}{q(x)}. \tag{44}$$

Then, maximizing (43) with respect to variations $\delta p(\theta)$ such that $\int d^n\theta\, p(\theta) = 1$, yields

$$0 = \int_{\Theta_n} d^n\theta \left(-\log \frac{p(\theta)}{g^{1/2}(\theta)} + S(\theta) + \log\zeta \right) \delta p(\theta)\,. \tag{45}$$

(The required Lagrange multiplier has been written as $1 - \log\zeta$.) Therefore the probability that the value of θ should lie within the small volume $dV_n = g_n^{1/2}(\theta)d^n\theta$ is

$$P(\theta)d^n\theta = \frac{1}{\zeta}\, e^{S(\theta)} dV_n \quad \text{with} \quad \zeta = \int_{\Theta_n} dV_n\, e^{S(\theta)}. \tag{46}$$

Equation (46) is the result we seek. It tells us that, as expected, the preferred value of θ is the value θ_0 that maximizes the entropy $S(\theta)$, eq.(44), because this maximizes the *scalar* probability density $\exp S(\theta)$.[12] It also tells us the degree to which values of θ away from the maximum θ_0 are ruled out.

The previous discussion allows us to refine our understanding of the ME method. ME is not an all-or-nothing recommendation to pick that single distribution that maximizes entropy and rejects all others. The ME method is more nuanced: in principle all distributions within the constraint manifold ought to be included in the analysis; they contribute in proportion to the exponential of their entropy and this turns out to be significant in situations where the entropy maximum is not particularly sharp.

Going back to the original problem of updating from the prior $q(x)$ given information that specifies the manifold $\{p(x|\theta)\}$, the preferred update within the family $\{p(x|\theta)\}$ is $p(x|\theta_0)$, but to the extent that other values of θ are not totally ruled out, a better update is obtained marginalizing the joint posterior

[12] The density $\exp S(\theta)$ is a scalar function; it is the probability per unit invariant volume $dV = g_n^{1/2}(\theta)d^n\theta$.

$P_J(x,\theta) = P(\theta)p(x|\theta)$ over θ,

$$P_{\rm ME}(x) = \int_{\Theta_n} d^n\theta\, P(\theta)p(x|\theta) = \int_{\Theta_n} dV_n \frac{e^{S(\theta)}}{\zeta} p(x|\theta) \ . \tag{47}$$

In situations where the entropy maximum at θ_0 is very sharp we recover the old result,

$$P_{\rm ME}(x) \approx p(x|\theta_0) \ . \tag{48}$$

When the entropy maximum is not very sharp eq.(47) is the more honest update.

The summary description of the ME method in the previous subsection can now be refined by adding the following line:

> *The ME posterior $P_{\rm ME}$ is a weighted average of all distributions in the family $\{p\}$ specified by the constraints $\mathcal{C}$. Each p is weighted by the exponential of its entropy $S[p,q]$.*

Physical applications of the extended ME method are ubiquitous. For macroscopic systems the preference for the distribution that maximizes $S[p,q]$ can be overwhelming but for small systems such fluctuations about the maximum are common. Thus, violations of the second law of thermodynamics can be seen everywhere — provided we know where to look. For example, eq.(46) agrees with Einstein's theory of thermodynamic fluctuations and extends it beyond the regime of small fluctuations. Another important application, developed in [7], is quantum mechanics — the ultimate theory of small systems.

7 Summary

Science requires a framework for inference on the basis of incomplete information. We showed how to design tools to represent a state of partial knowledge as a web of interconnected beliefs with no internal inconsistencies; the resulting scheme is probability theory. Then we argued that in a properly Bayesian framework the concept of information must be defined in terms of its effects on the beliefs of rational agents. The definition we have proposed – that information is a constraint on rational beliefs – is convenient for two reasons. First, the information/belief relation is explicit, and second, such information is ideally suited for quantitative manipulation using entropic methods. Finally, we designed a method for updating probabilities. The design criteria are strictly pragmatic; the method aims to be of universal applicability, it recognizes the value of information both old and new, and it recognizes the special status that must be accorded to considerations of independence in order to build models that are actually usable. The result — the maximum entropy or ME method — is a framework in which entropy is the tool for updating probabilities. The ME method unifies both MaxEnt and Bayes' rule into a single framework of inductive inference and allows new applications. Indeed, much as the old MaxEnt method provided the foundation for statistical mechanics, recent work has shown that the extended ME method provides an entropic foundation for quantum mechanics.

Within an informational approach it is not possible to sharply separate the subject matter or contents of science from the inductive methods of science; science includes both. This point of view has two important consequences. The first is that just as we accept that the contents of science will evolve over time, to the extent that contents and methods are not separable, we must also accept that the inference methods are provisional too — the best we currently have — and are therefore susceptible to future change and improvement.

The second consequence stems from the observation that experiments do not vindicate individual propositions within a theory; they vindicate the theory as a whole. Therefore, when a theory turns out to be pragmatically successful we can say that it is not just the contents of the theory that have been corroborated but also its methods of inference. Thus, the ultimate justification of the entropic methods of inference resides in their pragmatic success when confronted with experiment.

8 Discussion

The accidents of history have caused the term 'pragmatism' to acquire connotations that are not fully satisfactory. But we do not have to literally adopt Peirce's version of pragmatism or James' or Putnam's. It is not necessary that we agree with everything or even most of what these authors have said; not only did they not agree with each other, but their views evolved, and their later views disagreed with those held in their own youth. Instead, the proper pragmatic attitude is to pick and choose useful bits and pieces and try to stitch them into a coherent framework. To the extent that this framework turns out to be useful we have succeeded and that is all we need.

The vague notion of truth that I have favored — truth as a compliment — is not Peirce's. It might be closer to James' or Putnam's and even Einstein's — but it need not be identical to their notions either: the real purpose is not a theory of truth itself but rather those applications that may be tackled through a pragmatically designed framework for inference. In other words, the theory of truth is not the real goal; it is only an intermediate obstacle, perhaps even a distraction, on the way to the real problems: Can we do quantum mechanics? Can we do economics? Or, borrowing Floridi's words, can we "model the world in such a way to make sense of it and withstand its impact"?

I have argued in favor of an informational pragmatic realism but I have not attempted its systematic development. Some of its features can be summarized as follows: There is a world out there which we must navigate. Physical models are inference schemes; they are instruments to help us succeed. Within such models we find elements that purport to represent entities such as particles or fields. We also find other elements that are tools for manipulating information — probabilities, entropies, wave functions. All models inevitably involve concepts and categories of our own construction chosen for our own pragmatic reasons.

To the extent that a model is reliably successful we say that its entities are real. Beyond that it is meaningless to assert that these entities enjoy any special relation to the world that might be described as "referring" or "corresponding" to something independent of us.

Within the informational approach to science the notion of a structural realism in which the world consists of relations and structures only — without any entities or objects — makes no sense. The information or lack thereof is information about something — both the entities and the informational tools must appear in the models. Having said that, the central point of the informational approach to physics [5] [7] [8] is precisely that the formal rules for manipulating information — Bayesian and entropic methods — place such strong constraints on the formal structure of theories that a label of *informational structural pragmatic realism* may, in the end, be quite appropriate.

Therefore I welcome Floridi's reconciliation of the epistemic and the (non-eliminativist) ontic versions of structural realism. Indeed, the recognition that a model describes a structure at a given level of abstraction brings Floridi close to Putnam's internal realism. 'Internal' because the entities and structures in our world are defined *within* a given level of abstraction or conceptual framework which is chosen by us for our purposes. And 'realism' because the conceptual framework is not arbitrary; it is not independent of the world; the chosen entities and structures have to be useful and succeed in the real world. An important difference, however, is that Putnam's internal realism makes no mention of information. This is the gap that can hopefully be closed by the informational pragmatic realism advocated here.

Another aspect of Floridi's structural realism that I find appealing is the idea that the distinction between what constitutes intrinsic nature and what constitutes structure is not easily drawn — that relata are not logically prior to relations, that they come together, all or nothing, in a single package. Such an idea is already deeply ingrained in physics. For example, a quantity such as electric charge refers on one hand to an intrinsic property of a particle — an electron without its charge would just not be an electron — and on the other hand electric charge describes interactions, the relations of the particle to other charged particles. It is totally inconceivable to claim that a particle could possible have an electric charge and yet not interact with other particles according to the laws of electromagnetism. Therefore, it is perfectly legitimate to assert that an electron is, to use Floridi's term, a structural object — its intrinsic properties are defined by the structure of its relations to other particles. Furthermore, as shown in [7] the electric interactions can indeed be approached from a purely informational perspective. They are described through information — that is, constraints — on the allowed motions that the particle can undertake.

The definition of information that I have proposed — information as a constraint on rational beliefs — differs from Floridi's definition as well-formed, meaningful, and truthful data. But there is considerable overlap. Indeed, within the entropic/Bayesian framework a mere set of numbers — or data — does not by itself constitute information. It is necessary that the data be embedded within a model. This is what endows the data with significance, with meaning.

The model provides a connection between the data and the other quantities one wishes to infer — a relation that is usually established through what in statistics is called a likelihood function. Moreover, as discussed in section 5, the data becomes "information" precisely at that moment when we feel justified in allowing its effects to propagate throughout the web of beliefs. Such data deserve the compliment 'true'. And thus data is information provided it is well-formed and meaningful by virtue of being embedded in a model, and is truthful by virtue of an appropriate judgement.

Pragmatism is a form of empiricism too, at least in the sense that,

> "...it is contented to regard its most assured conclusions concerning matters of fact as hypotheses liable to modification in the course of future experience..." ([34], p.vii)

van Fraasen's empiricism, either in its earlier form of *constructive empiricism* or in the later form of *empiricist structuralism*, appears to adopt a correspondence model of truth and this places him squarely in the anti-realist camp. Nevertheless, if we ignore the labels and just look at what he actually wrote we find significant points of contact with both structural and pragmatic realism. van Fraasen has argued that in a semantic approach to theories as mathematical models if one mathematical structure can represent the phenomena then any other isomorphic structure can also do it:

> "...models represent nature only up to isomorphism – they only represent structure." [35]

This has been called the underdetermination *problem* but we will, more optimistically, regard it as an *opportunity*. It offers the possibility of imposing additional pragmatic virtues such as explanatory power, simplicity, etc., that go beyond mere empirical adequacy, as criteria for the acceptance of theories. This opportunity has been advantageously pursued by Ellis [22], and Floridi ([1], p.358-360). van Fraasen too recognizes that the acceptance of a theory rests on pragmatic considerations that go beyond mere belief in empirical adequacy. ([13], p. 12-13).

From our pragmatic perspective two of van Fraasen's ideas appear particularly appealing. One has to do with clarifying the meaning of empirical adequacy [35] [36]. The problem is that under a correspondence model of truth one might naively attempt to achieve an agreement between the model and the phenomena as they are in themselves — which leads to the same old problem of reference. van Fraasen skillfully evades this problem by asserting that empirical adequacy is not adequacy to the phenomena-in-themselves but rather to the *phenomena as described by us.* This is as pragmatic as it gets! Empirical adequacy involves the comparison of two descriptions; one is supplied by the theory and the other is our description — a "data model" — of certain selectively chosen aspects that are relevant to our interests, a description in terms of concepts invented by us because they are adequate to our purposes. van Fraasen sums up as follows:

> "... in a context in which a given model is my representation of a phenomenon, there is no difference between the question of whether a theory fits that representation and the question of whether it fits the phenomenon." [36]

It is difficult to imagine that either James or Putnam would have objected.

The second appealing idea concerns the transition from an older theory to a newer theory that is presumably better — van Fraasen calls it the issue of royal succession in science [12]. This topic was briefly mentioned in section 1 with reference to the pessimistic meta-induction and the no-miracles arguments.

Any systematic procedure for theory revision would naturally attempt to preserve those features of the old theory that made it work. Which leads to the question "what makes a good theory good?" Whatever it is — it could perhaps be that it captures the correct structure — we can call it "true" or "real" but this is just a name, a compliment that merely indicates success. Explaining the predictive success of science by "realism" or by "truth" is somewhat analogous to explaining that opium will put you to sleep because it has "dormitive" powers.

Instead I find van Fraasen's pragmatic argument much more persuasive. The success of science is not a miracle — it is a matter of how scientific theories are designed: beyond retaining the empirical successes of the old theory we want more and the new theory must provide us with new empirical successes. This is a necessary requirement for the new theory to be accepted. Thus today's science is bound to be more successful than yesterday's — otherwise we would not have made the switch.

Within the pragmatic framework advocated here royal succession is explained by adopting a relaxed form of van Fraasen's criteria. The acceptance criteria are extended beyond mere empirical success to include other pragmatic virtues: the new theory must retain the pragmatic successes of the old and add a few of its own. The point is that it is not strictly necessary that the new theory must lead to new empirical successes; it might, for example, just lead to better explanations, or be more computationally convenient.

I conclude that an *informational pragmatic realism* has the potential of incorporating many valuable pragmatic insights derived from internal realism, informational structural realism and empiricist structuralism into a single coherent doctrine and thereby close the gap between them.

Acknowledgements: I am grateful to C. Cafaro, N. Caticha, A. Giffin, A. Golan, P. Goyal, K. Knuth, C. Rodríguez, M. Reginatto, and J. Skilling for many useful discussions on entropic inference; and also to A. Beavers and L. Floridi for the invitation to participate in this symposium.

References

[1] L. Floridi, *The Philosophy of Information* (Oxford U. Press, Oxford 2011).

[2] L. Floridi, "A defence of informational structural realism", Synthese **161**, 219 (2008).

[3] L. Floridi, "Scepticism and the Search for Knowledge: a Peirceish Answer to a Kantian Doubt", Trans. C.S. Peirce Soc. **30**, 543 (1994).

[4] L. Floridi, "Information, Possible Worlds, and the Cooptation of Scepticism", Synthese **175.1**, 63 (2010).

[5] E. T. Jaynes: Phys. Rev. **106**, 620 and **108**, 171 (1957); *E. T. Jaynes: Papers on Probability, Statistics and Statistical Physics*, ed. by R. D. Rosenkrantz (Reidel, Dordrecht, 1983).

[6] A. Caticha: Phys. Lett. **A244**, 13 (1998); Phys. Rev. **A57**, 1572 (1998); Found. Phys. **30**, 227 (2000).

[7] A. Caticha, "Entropic Dynamics, Time, and Quantum Theory", J. Phys. **A 44**, 225303 (2011).

[8] A. Caticha, *Entropic Inference and the Foundations of Physics* (monograph commissioned by the 11th Brazilian Meeting on Bayesian Statistics – EBEB-2012 (USP Press, São Paulo, Brazil 2012); online at http://www.albany.edu/physics/ACaticha-EIFP-book.pdf.

[9] R.T. Cox, "Probability, Frequency and Reasonable Expectation", Am. J. Phys. **14**, 1 (1946).

[10] E. T. Jaynes, *Probability Theory: The Logic of Science*, ed. by L. Bretthorst (Cambridge U.P., 2003).

[11] H. Putnam, *Reason, Truth and History* (Cambridge U. Press, Cambridge 1981).

[12] B. C. van Fraasen, "Structure: Its Shadow and Substance", Brit. J. Phil. Sci. **57**, 275 (2006).

[13] B. C. van Fraasen, *The Scientific Image* (Clarendon, Oxford 1980).

[14] J. E. Shore and R. W. Johnson, IEEE Trans. Inf. Theory **IT-26**, 26 (1980); IEEE Trans. Inf. Theory **IT-27**, 26 (1981).

[15] P. M. Williams, "Bayesian Conditionalization and the Principle of Minimum Relative Information", Brit. J. Phil. Sci. **31**, 131 (1980).

[16] J. Skilling, "The Axioms of Maximum Entropy" in *Maximum-Entropy and Bayesian Methods in Science and Engineering*, G. J. Erickson and C. R. Smith (eds.) (Kluwer, Dordrecht, 1988).

[17] C. C. Rodríguez, "Entropic priors" in *Maximum Entropy and Bayesian Methods*, edited by W. T. Grandy Jr. and L. H. Schick (Kluwer, Dordrecht, 1991).

[18] A. Caticha and A. Giffin, "Updating Probabilities", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, Ali Mohammad-Djafari (ed.), AIP Conf. Proc. **872**, 31 (2006) (arxiv.org/abs/physics/0608185).

[19] H. Putnam, *Mathematics, Matter, and Method*, Vol. 1 (Cambridge U. Press, Cambridge 1975).

[20] D. Papineau (ed.), *The Philosophy of Science* (Oxford U. Press, Oxford 1996).

[21] P. Godfrey-Smith, *Theory and Reality* (U. of Chicago Press, Chicago 2003).

[22] B. Ellis, "What science aims to do" in *Images of Science* ed. by P. Churchland and C. Hooker (U. of Chicago Press, Chicago 1985); reprinted in [20].

[23] H. Putnam, "How to be an internal realist and a transcendental idealist (at the same time)" in *Language Logic and Philosophy*, Proc. of the 4th International Wittgenstein Symposium (Kirchberg/Wechsel, Austria 1979).

[24] A. Fine, *The Shaky Game – Einstein Realism and the Quantum Theory* (U. of Chicago Press, Chicago 1996).

[25] H. P. Stapp, "The Copenhagen Interpretation", Am. J. Phys. **40**, 1098 (1972).

[26] N. Bohr, *Atomic Theory and the Description of Nature* (Cambridge U.P., 1934; reprinted by Ox Bow Press, 1987).

[27] N. Bohr, *Essays 1932-1957 on Atomic Physics and Human Knowledge* (Wiley, 1958; reprinted by Ox Bow Press, 1987).

[28] A. Caticha, "Quantifying Rational Belief", *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, ed. by P. Goggans *et al.*, AIP Conf. Proc. **1193**, 60 (2009) (arXiv.org/abs/0908.3212).

[29] T. Cover and J. Thomas, *Elements of Information Theory* (Wiley, New York 1991).

[30] A. Golan, "Information and Entropy in Econometrics – A Review and Synthesis", Foundations and Trends in Econometrics **2**, 1–145 (2008).

[31] P. W. Adriaans and J.F.A.K. van Benthem (eds.), *Handbook of Philosophy of Information* (Elsevier, 2008).

[32] A. Caticha, "Maximum entropy, fluctuations and priors", *Bayesian Methods and Maximum Entropy in Science and Engineering*, ed. by A. Mohammad-Djafari, AIP Conf. Proc. **568**, 94 (2001) (arXiv.org/abs/math-ph/0008017).

[33] S. Amari, *Differential-Geometrical Methods in Statistics* (Springer-Verlag, 1985).

[34] W. James, *The Will to Believe* (Dover, New York 1956).

[35] B. C. van Fraasen, “Structure and Perspective: Philosophical Perplexity and Paradox”, in *Logic and Scientific Methods*, p. 511, ed. by M. L. Dalla Chiara *et al.* (Kluwer, Netherlands 1997).

[36] B. C. van Fraasen, “Representation: The Problem for Structuralism”, Philosophy of Science **73**, 536 (2006).