JEFF PARIS

COMMON SENSE AND MAXIMUM ENTROPY

ABSTRACT. This paper concerns the question of how to draw inferences *common sensically* from uncertain knowledge. Since the early work of Shore and Johnson (1980), Paris and Vencovská (1990), and Csiszár (1989), it has been known that the *Maximum Entropy Inference Process* is the only inference process which obeys certain common sense principles of uncertain reasoning. In this paper we consider the present status of this result and argue that within the rather narrow context in which we work this complete and consistent mode of uncertain reasoning is actually characterised by the observance of just a *single* common sense principle (or *slogan*).

1. INTRODUCTION

The practical need to understand uncertain reasoning, that is, how to represent and draw conclusions from uncertain knowledge or beliefs, arose about 15–20 years ago with attempts to build what were called at that time *Expert Systems*. To take the particular case of, say, a general medical expert system, the idea was to provide a computer with medical knowledge base so that in order to diagnose a patient's illness it would be enough to simply type into the computer the patient's symptoms, signs, gender, age etc. (in other words the information apparently available to the doctor at a consultation) and have the computer (now referred to as an Expert System) use its 'knowledge' to give a, possibly qualified, diagnosis, much in the way a doctor does. The two key problems then are acquiring the 'knowledge' with which to fill the computer and then programming the computer to correctly apply this knowledge to the problem in hand.

As a first step towards achieving this we could argue that on the face of it the doctor him, or her, self acts just like such an expert system, so it should be enough to extract 'the general knowledge the doctor uses in diagnosing' and to discover what the trick is of combining this with the patient's signs, symptoms etc. to come up with a diagnosis.

Now if we were to ask the doctor what knowledge he/she was using during any particular consultation the initial answer would probably be some set of rules or causal links, such as,

Memory loss in the elderly is not uncommon

Synthese **117:** 75–93, 1999.

Constipation is sometimes the result of memory loss

and, possibly, simple statements about frequencies of various conditions, for example,

Most of my patients are hypochondriacs.

Pressing the doctor to be a little more precise here, s/he may be prevailed upon to re-express such assertions in terms of the probabilities s/he would be willing to assign to various events. For example s/he might rephrase this last statement to:

I'd give probability 4/5 to the next patient through my door being a hypochondriac.

Considerations such as these have led to the simple picture, or model, of an intelligent agent's knowledge, here the agent is the doctor, as consisting of a (finite) set, K say, of assertions of the form

Probability that X *holds given that* Y *holds is* c or $\mathrm{Prob}(X|Y) = c$

or, even, simply,

Probability that X *holds is* d or $\mathrm{Prob}(X) = d$

where c, d etc. lie in the interval [0, 1], and, in the case of the above-mentioned doctor's knowledge, X, Y, etc., stand for various combinations of signs, symptoms, conditions etc. That is, the agent's knowledge is pictured as simply a set K of probabilities, or more precisely as constraints on a probability function and its corresponding conditional probability function,

$$\mathrm{Prob}(\ ) : S\mathcal{L} \to [0, 1]$$

$$\mathrm{Prob}(|) : S\mathcal{L} \times S\mathcal{L} \to [0, 1],$$

where $S\mathcal{L}$ are the set of *sentences*, or *propositions*, built up from some finite set $\mathcal{L}$ of propositional variables using the connectives *and* ($\wedge$), *or* ($\vee$), and *not* ($\neg$).[1]

Two points should be emphasised here. Firstly in this model we are assuming that K *is* the the sum total of the agent's knowledge, it is not a summary of the agent's knowledge, it is *all* there is. This assumption (referred to as the Watt's Assumption in (Paris 1994)) will be crucially important in what follows.

Secondly, we are thinking of these probabilities as *personal* to the agent, as reflected, say, in the agent's willingness to bet on the outcomes. Of course in the case we have been considering of a doctor's knowledge about the patients attending his/her clinic, there would appear to be a notion of a *true*, or *correct*, probability, for example in the case of hypochondriacs the *proportion* of patients attending the clinic who suffer from hypochondria. However we shall make no assumption here that the agent's (in this case the doctor's) probabilities agree with any 'correct' ones (if indeed such even exist).

In practice, of course, such figures as actually supplied by a doctor may fall short of the defining requirement for being probabilities (on $S\mathcal{L}$), that is that the functions Prob( ), Prob(|) take values in the interval [0, 1], Prob( ) satisfies that for $\theta, \varphi \in S\mathcal{L}$,

(P1) if $\models \theta$ then $\text{Prob}(\theta) = 1$,

(P2) if $\models \neg(\theta \wedge \varphi)$ then $\text{Prob}(\theta \vee \varphi) = \text{Prob}(\theta) + \text{Prob}(\varphi)$,

and that Prob(|) is related to Prob( ) via the identity

$$\text{Prob}(\theta|\varphi) \cdot \text{Prob}(\varphi) = \text{Prob}(\theta \wedge \varphi),$$

(which forces $\text{Prob}(-|\varphi)$ to also satisfy the defining conditions (P1)–(P2) for a probability function provided $\text{Prob}(\varphi) \neq 0$).

Nevertheless, we shall idealise here and make the assumption that our intelligent agent's knowledge base has this form. Having done this we might now ask, how much practical use is K to the agent? For example, in the case we are considering of a medical expert system, if we elicited from the doctor a knowledge base K as above, (which is supposedly a copy of the doctor's knowledge), would this be sufficient to determine all further probabilities (say of various conditions given various combinations of signs symptoms etc.), and hence enable the computer to simulate the doctor as far as diagnosis is concerned?

Unfortunately, it would be very unlikely in practice that K alone would suffice to determine all other probabilities. In other words there would usually be a range of different probabilities which could consistently be assigned on the basis of this knowledge base, equivalently, there could be many different probability functions Prob on $S\mathcal{L}$ satisfying the constraints K.

The reason for this is that the function Prob is determined by the vector

$$\text{Prob} \leftrightarrow \langle \text{Prob}(\alpha_1), \ldots, \text{Prob}(\alpha_{2^n}) \rangle$$

of its values on the *atoms* α_i of the language, where, for $\mathcal{L} = \{p_1, \ldots, p_n\}$, the atoms are all the sentences of the form

$$\pm p_1 \wedge \pm p_2 \wedge \ldots \wedge \pm p_n.$$

Now in even a very modest expert system n will be in the tens so that 2^n will already be astronomic and it would (in general) be very much the exception if K, which would rarely contain more than $O(n^k)$ constraints for some small k, would determine Prob uniquely.

Assuming then, as we do, that K sums up *all the agent's knowledge*, the agent must go beyond simply the constraints, as such, in K if s/he is to determine a total, personal, probability function on SL. One way the agent might achieve this is by appealling to some *higher principles, or rules, of uncertain reasoning* to generate, or infer, new knowledge from the knowledge base K s/he does possess. And where to find such principles? Well, we would claim, that some such 'principles' already have a name. They are called *common sense*, and it is in this direction that we now turn.

2. THE SYMMETRY PRINCIPLE

In the mid 1980's Alena Vencovská and myself, began making a collection of so called *common sense principles* of uncertain reasoning (see Paris and Vencovská 1989, 1990, 1996a, 1997; Paris 1994). Some of these were, apparently, new, some already existed, under different names, in the literature, in particular in the work of Shore and Johnson from the early 1980's. Over the intervening years further investigation into the relation between these apparently diverse principles have shown that they can actually all be viewed as special examples of one grand principle, which seems, self evidently, *common sensical*. This unifying principle was clearly stated (but afforded only the status of a slogan!) by Bas van Fraassen (1989) in the late 1980's as follows:

THE SYMMETRY PRINCIPLE. Essentially similar problems should have essentially similar solutions.

The principle applies to our situation in that our 'problem' (more precisely the agent's problem) is one of assigning probabilities (subject to the constraints put on these by the knowledge). The principle says that whatever way probabilities are assigned essentially similar probabilities should be assigned in essentially similar situations. So the principle says nothing, directly, about *what* probabilities should be assigned, only about

the *process* of assigning probabilities. Nevertheless, as we shall see, it does. indirectly, have a profound influence on what probability values can be assigned.

By taking various interpretations of what we mean by 'essentially similar' this one principle yields a number of other, common sense, principles. The plan for the remainder of this section is to state some such principles *informally* and give everyday examples which are intended (*only*) to illustrate the underlying idea. In the section following the key consequences of abiding by these principles, and a 'practical' example will be given, again within an informal setting. In the final section a precise mathematical formulation of these principles and consequences will be presented.

We start our list of principles with the:

RENAMING PRINCIPLE. Changing the names we call things should not change the probabilities we assign to them [– because this does not *essentially* change the problem].

Example. What probability should we assign to a standard six sided die landing six?

Since the problem is essentially the same if we interchange six with any other number on the die the probability we assign to a six coming up should (according to the Symmetry Principle) be the same as the probability we assign to any particular one of the other five numbers. Since exactly one of them must come up each of these (and in particular a six) should get probability 1/6.

Notice that this argument is based entirely on the symmetries in the situation, it has nothing to do with the the frequency of sixes in a series of throws of the die (which might somehow be considered to reflect the *correct* probability).

IRRELEVANT INFORMATION PRINCIPLE. Knowledge entirely irrelevant to the problem in hand can he ignored [– because it is *inessential*].

Example. Suppose I asked you what probability you would assign to Manchester United repeating their phenomenal successes of recent years by winning the league again this season, and to help you in your deliberations I supplied you with the 1955 tide tables for Lowestoft and the surrounding coastal resorts.

Doubtless you would, quite reasonably, feel that these 1955 tide tables for Lowestoft and the surrounding coastal resorts were irrelevant, or *inessential*, information which could safely be disregarded in corning to your conclusions. And doubtless, if you were a keen supporter of some

other Premiership team you would also tell me what to do with my 1955 tide tables for Lowestoft and the surrounding coast resorts!

OBSTINACY PRINCIPLE. Receiving evidence supporting what one already thinks should not cause one to alter one's views [– because that evidence is effectively redundant, so inessential].

Example. Suppose I am absolutely sure that my window cleaner did not actually clean my windows this morning. My neighbour, who spent the day lolling in a deck chair on the back lawn, confirms that he certainly did not clean my back windows. In that case I should continue to believe that he did not clean my front windows either.

In terms of the Symmetry Principle my neighbour's evidence does not *essentially* change the 'problem' I have, (of what probability I should assign to the window cleaner having not cleaned my front windows), since I had come to the conclusion, in any case, that he had not cleaned my back windows. In other words, since my neighbour was not telling me anything I did not already believe, her information, in that sense, gave me nothing new, and so should certainly not cause me to alter my beliefs.

RELATIVISATION PRINCIPLE. Very roughly, the probabilities one would give if some event occurred should only depend on the knowledge one would have if that event occurred.

From the point of view of the Symmetry Principle the knowledge I have about the world if the event *does not* occur is irrelevant, or *inessential*, when it comes to the problem of deciding what probabilities to assign when the event *does* occur. [This principle assumes that the knowledge comes conveniently divided into these forms, together with the probability of the event itself.]

Example. Suppose I stagger out of a bar at closing time and for the life of me I cannot remember what day it is, any day of the week is as likely as any other as far as I am concerned. Should I wait at the bus stop, or should I attempt to walk home? The trouble is, if it is Sunday I would be wasting my time waiting because the last bus on Sunday is 1.00 pm in the afternoon. On the other hand there are late buses on all the other nights and if I knew it was not Sunday I would feel confident enough about getting a bus that I would wait, rather than run the risk of having the pavement rear up at me on the long walk home. Just then I see a sign on the bus stop to say that the 1.00 pm bus on Sundays will be subject to serious delays for the next two weeks. Appreciating the full relevance of this expression I might now

be encouraged to wait, no matter, but in any case it should not cause me to alter my view that *if it is not Sunday then it is worth waiting*.

EQUIVALENCE PRINCIPLE. Two knowledge bases which express the same knowledge (in the sense of the constraints they impose) should engender the same probability assignments [– because they are essentially similar].

Example. Suppose my room mate tells me that today he plans to toss a (fair) coin. If it comes down heads he says he will then toss a second coin and fatalistically accept that he should continue with his resolution to quit smoking only if this second coin also lands heads. In this case the probability I should assign to him breaking his resolution today should be the same as if he had told me that he planned to toss both the coins and would continue his resolution if they land HH, whilst breaking it if they land HT, since as far as I am concerned my knowledge in the two cases is essentially the same.

CONTINUITY PRINCIPLE. Microscopic changes in the knowledge base should not cause macroscopic changes in the probabilities assigned [– because the two problems are essentially similar and so should have essentially similar solutions].

Example. I have been playing poker all night with a most amazingly lucky opponent. Towards daybreak however, I start to wonder if he is not maybe making his own luck, in short, cheating. At that point he deals me three queens, and I again anticipate that, as has almost invariably happened before this session, I will be led on by the strength of my hand to raise the stakes higher and higher, only for him to come out on top with even better cards.

But for once I win the hand. Should this allay my fears that he has been cheating? Well, very slightly perhaps, but really only as a drop in the ocean given the magnitude of his good fortune overall.

In terms of the Symmetry Principle the problem of assigning a probability to my opponent being a cheat after this hand is, given the long night behind us, very similar to the same problem before that hand, and should therefore have a very similar solution. In other words, this comparatively microscope new piece of evidence should not have a macroscopic effect on the probabilities assigned.

WEAK INDEPENDENCE PRINCIPLE. [A precise formulation of this principle will be given in the Section 4. For the present we shall content ourselves with an informal example illustrating the underlying idea.]

Example. I read the paper and learn that according to the statistics my chance of winning a fortune on the lottery in my lifetime is 200,000 to 1 whereas my chance of being murdered is 200 to 1. At this point I form an opinion as to the chances of my winning the lottery and subsequently being murdered. Reading on I learn that my chances of being killed as the result of a nuclear accident are currently estimated to be only 50,000 to 1. Comforting as this may be, this new revelation should not alter the opinion I had previously formed, since, as far as I am aware, it is independent of my winning the lottery.

In terms of the Symmetry Principle this additional information gives nothing new relative to the problem of assigning a probability to winning the lottery and being murdered, and so should have no effect. Notice that here the additional evidence, in short that my probability of being killed as the result of a nuclear accident is 50,000 to 1, is not 'irrelevant' in the sense of the irrelevant Information Principle, since clearly I cannot both be killed as the result of a nuclear accident and be murdered. [This distinction will be clearer later when we give formal versions of these principles.]

3. THE MAIN THEOREM

In the previous section we presented some seven 'common sense' principles of uncertain reasoning as special cases of the Symmetry Principle. Whilst these were presented informally they can, as we shall see in the next section, be given a precise mathematical formulation.

Clearly observance of each of these principles restricts the probabilities that can be assigned. So much so, that, remarkably, within the formal mathematical framework we are able to prove (Paris and Vencovská 1990, 1997):

THEOREM 1. Observance of the seven principles of Renaming, Irrelevant Information, Relativisation, Obstinacy, Equivalence, Continuity, and Weak Independence, all of which may be viewed as special cases of the Symmetry Principle, allows no freedom of choice in the probability values which can be assigned (on the basis of such knowledge bases K, as above).

In other words, observance of these principles, completely determines all other probabilities.

In fact, it turns out that these probability forced to assign on the basis of such a knowledge base K and adherence to the seven principles are precisely those given by the *Maximum Entropy solution* to K. To explain this, recall that the probability function Prob is determine by the vector in $\mathbb{R}^{2^n}$,

$$\langle \mathrm{Prob}(\alpha_1), \ldots, \mathrm{Prob}(\alpha_{2^n}) \rangle.$$

The *Maximum Entropy solution* to the knowledge (i.e. constraints) K is that solution Prob to K for which the entropy,

$$-\sum_{i=1}^{2^n} \mathrm{Prob}(\alpha_i) \cdot \log(\mathrm{Prob}(\alpha_i)),$$

is maximal.

This notion is, of course, already well known in thermodynamics and information theory. The fact that this old friend appears also here as a distant consequence of the Symmetry Principle is both pleasing and intriguing.

In the next section a precise mathematical formulation of the principles and the above theorem will be given. Firstly, however, we will present, informally, a simple, and, admittedly, highly contrived, example of this theorem and the way the principles can be made to work.

For this example let us suppose that you are locked away in some dark dungeon with two fellow prisoners, A and B say. All you have been told about the situation is that

(1) At least one of you or A will be spared.
(2) The Governer plans to roll a fair die. If it comes up 6 both you and B will be executed. Otherwise at least one of you will be spared.

So, the question is, what probability, should you give to your being spared? To answer this using just these 'common sense' principles, suppose first that you only knew (1). Then there are exactly three outcomes here:

O_1: You and A both spared,

O_2: You spared, A executed,

O_3: You executed, A spared.

because the knowledge (1) tells you that 'You and A both executed' won't happen. Rephrased in this way the knowledge amounts to just

Exactly one of O_1, O_2, O_3 will happen.

Now there is complete symmetry here between O_1, O_2, O_3. Thus, on the basis of this knowledge, the problem of giving a probability to O_1 is essentially the same as the problem of giving a probability to O_2 (or O_3). Hence, by Renaming, since one of the three of them must occur we conclude each has should have probability 1/3. Thus, invoking Equivalence, the probability that you should assign to your being executed (on the basis of (1)) i.e., that O_3 will occur, should be 1/3, and the probability that you should assign to being spared, i.e. the probability that one of O_1, O_2 will occur, should be 1/3 + 1/3 = 2/3.

The principles then forces us to come to the conclusion that, purely on the basis of knowledge (1), your probability that you will be spared should be 2/3, equivalently, your probability that you will be executed should be 1/3.

But what about the other piece of knowledge, (2), that we have so far ignored, that the Governor will roll a die and both you and *B* will be executed just if it lands 6?

Well, for a moment, consider instead the knowledge base:

(1) At least one of you or *A* will be spared.
(3) The probability that you will be executed is 1/3.

Since, as we have already seen, (3) was exactly what should have concluded on the basis of (1) alone, (3) is actually an *inessential* addition to this knowledge base, so that, by Obstinacy,

> Your conclusions on the basis of (1) alone should agree with those formed on the basis of (1) + (3).

Now (1)+(3) can be re-expressed as

(1) At least one of you or *A* will be spared.
(3) The probability that either you will be executed and *B* will be executed, or you will be executed and *B* will be spared is 1/3.

Clearly this actually says nothing about *B*, indeed there is again complete symmetry here between *B* being executed and being spared; if we were to transpose '*B* executed' everywhere with '*B* spared' we would get back exactly the same knowledge base. Thus, as an application of Renaming (and Equivalence), both '*you executed and B executed*' and '*you executed and B spared*' should get the same probability on the basis of (1) + (3). Furthermore, since these two probabilities should together sum to 1/3 , we see that they should each get value 1/6. But, as we have already seen, (3) is an inessential addition to (1), so by Obstinacy you should give '*you executed and B executed*' probability 1/6 on the basis of (1) alone.

But now, since the Govenor's die is fair, (2) amounts to saying that the probability that both you and B will be executed is 1/6, which is exactly what we have just reasoned that it should be on the basis of (1) alone. Hence adding (2) to (1) only adds inessential information, inessential because you (should) believe it in any case. So, as an example again of the Obstinacy Principle, the probability you give to 'you being spared' on the basis of the knowledge base (1) + (2) should be the same as the probability you give it on the basis of (1) alone – which we have already reasoned should be 2/3, and we have our answer.

Of course this was a dreadfully simple (and contrived) example. However, the fact is, as the theorem tells us, that for *all* such knowledge bases (and we will make this precise in the next section), appealling to our seven common sense principles will always lead one to a unique probability assignment. Indeed, the proofs of the main theorem in (Paris and Vencovská 1990) and theorem 6 of (Paris and Vencovská 1997) (see also (Paris and Vencovská 1996b)) show that we have even more than that: For all practical purposes (more precisely, provided our knowledge could be expressed using explicitly only rational numbers) we can actually deduce these unique values using step by step arguments such as in the above example.

Furthermore, although we initially introduced the whole discussion from the point of view of someone attempting to build a medical expert system, it is clear from the example that this result applies in general to any intelligent agent who's knowledge is of this form. In particular, if we assume that *our* knowledge is of this form (as it was in the prisoner example), then it applies to us.

One comforting consequence of this then, is that if we all have the same, or similar knowledge, about some issue and we all act commonsensically, then we should all come to similar conclusions. Turning this on its head then, if someone seriously disagrees with you, *you* can explain this either by *their* lack of background knowledge, or because *they* are not using common sense!

In view of this theorem, it would now seem that our problems are now solved: If our expert system, or intelligent agent, has a knowledge base K of this form then the application of common sense to K uniquely determines all other probabilities (and so enables such an expert system to give consistent, indeed, *ideally sensible*, qualified diagnoses for any combination of signs, symptoms, etc.).

So, is that it then, nothing more to say, all sewn up? Unfortunately not.

The flaw is that whilst the theorem tells us that there are unique, ideal, probabilities we should give if we are to obey 'common sense' (and have our knowledge in this form), the theorem give *us* no obvious *practical* way

of computing these probabilities, *in general*. In short the general problem of assigning *any* probabilities consistent with K (not necessarily these maximum entropy values even) is computationally intractable (assuming $P \neq NP$), even if we only want approximate values, and then only 'most of the time' (assuming $RP \neq NP$), see (Maung and Paris 1990; Paris 1994).

So here we find ourselves in a quandary. We know what we should do, indeed to do anything else requires us to act in contravention of common sens, yet, at the same time, we can't do it except in rather special, simple, circumstances. Circumstances which, in the real, highly complicated world we live in, we would almost never actually encounter.

And, if acting 'common-sensically' is identified with intelligence, *as we maintain it should be when the knowledge is of this form*, then this amounts to saying that in such situations we cannot hope to act intelligently, except possibly in rare flashes of inspiration!

4. MATHEMATICAL FORMULATION

The results of the earlier sections will now be reformulated, precisely, within a formal mathematical framework, following the presentation given in (Paris and Vencovská 1997).

Throughout let L stand for the countably infinite set of propositional variables $p_1, p_2, \ldots, p_n, \ldots, n \in \mathbb{N}$, let SL denote the sentences of L built up using the connectives $\wedge$, $\vee$, $\neg$ and let $SL(p_{i_1}, \ldots, p_{i_n})$ denote the set of sentences of the finite sublanguage of L with (distinct) propositional variables $p_{i_1}, \ldots, p_{i_n}$. We will use θ, φ, ψ etc. to denote sentences. Let $B(p_{i_1}, \ldots, p_{i_n})$ stand for the Lindenbaum Algebra on $SL(p_{i_1}, \ldots, p_{i_n})$, so the elements of this algebra are the equivalence classes

$$\bar{\theta} = \{\varphi \in SL(p_{i_1}, \ldots, p_{i_n}) | \varphi \equiv \theta\}$$

where $\theta \in SL(p_{i_1}, \ldots, p_{i_n})$ and $\equiv$ stands, as usual, for logical equivalence and the operations of complement ($\neg$), meet ($\wedge$), and join ($\vee$) are defined by

$$\neg(\bar{\theta}) = \overline{(\neg\theta)},\ \bar{\theta} \wedge \bar{\varphi} = \overline{(\theta \wedge \varphi)},\ \bar{\theta} \vee \bar{\varphi} = \overline{(\theta \vee \varphi)}.$$

As usual, g is an *isomorphism* between $B(p_{i_1}, \ldots, p_{i_n})$ and $B(p_{j_1}, \ldots, p_{j_m})$, written

$$g : B(p_{i_1}, \ldots, p_{i_n}) \cong B(p_{j_1}, \ldots, p_{j_m}),$$

if g is a bijection from $\{\bar{\theta}|\theta \in SL(p_{i_1}, \ldots, p_{i_n})\}$ to $\{\bar{\theta}|\theta \in SL(p_{j_1}, \ldots, p_{j_m})\}$ (so, necessarily, $n = m$), such that for all $\theta, \varphi \in SL(p_{i_1}, \ldots, p_{i_n})$,

$$g(\overline{\neg\theta}) = \neg g(\bar{\theta}),\ g(\overline{\theta \wedge \varphi}) = g(\bar{\theta}) \wedge g(\bar{\varphi}),$$

$$g(\overline{\theta \vee \varphi}) = g(\bar{\theta}) \vee g(\bar{\varphi}).$$

An *automorphism* of $B(p_{i_1}, \ldots, p_{i_n})$ is an isomorphism of $B(p_{i_1}, \ldots, p_{i_n})$ with itself. The (distinct) *atoms* of the Boolean algebra $B(p_{i_1}, \ldots, p_{i_n})$ are the equivalence classes of the 2^n atoms of $SL(p_{i_1}, \ldots, p_{i_n})$ (recall these are the sentences of $L(p_{i_1}, \ldots, p_{i_n})$ of the form $\pm p_{i_1} \wedge \pm p_{i_2} \wedge \cdots \wedge \pm p_{i_n}$). Clearly, from the Disjunctive Normal Form Theorem every element of $B(p_{i_1}, \ldots, p_{i_n})$ is a unique join of these atoms. Hence, since an automorphism of $B(p_{i_1}, \ldots, p_{i_n})$ must map atoms to atoms, it is clear that every automorphism determines, and conversely is determined by, a unique permutation of the atoms of $B(p_{i_1}, \ldots, p_{i_n})$ equivalently of $L(p_{i_1}, \ldots, p_{i_n})$. In what follows therefore we shall specify automorphisms simply by their actions on the atoms.

Recall that from the earlier sections w is a *probability function* on SL (and similarly for $SL(p_{i_1}, \ldots, p_{i_n})$) if $w : SL \to [0, 1]$ and for all $\theta, \varphi \in SL(SL(p_{i_1}, \ldots, p_{i_n}))$,

(P1) if $\models \theta$ then $w(\theta) = 1$

(P2) if $\models \neg(\theta \wedge \varphi)$ then $w(\theta \vee \varphi) = w(\theta) + w(\varphi)$

As shown in (Paris 1994) page 10 (for example), simple consequences of (P1)–(P2) are that for $\theta, \varphi \in SL(SL(p_{i_1}, \ldots, p_{i_n}))$,

(i) If $\models \theta$ then $w(\theta) = 1$ and $w(\neg\theta) = 0$.
(ii) If $\theta \models \varphi$ then $w(\theta) \leq w(\varphi)$, and if $\theta \equiv \varphi$ then $w(\theta) = w(\varphi)$.
(iii) $w(\theta \vee \varphi) = w(\theta) + w(\varphi) - w(\theta \wedge \varphi)$.

In what follows w, w_0 etc. will be used for probability functions.

Notice that if $\psi \in SL(p_{i_1}, \ldots, p_{i_n})$ then, by the Disjunctive Normal Form Theorem,

$$\psi \equiv \bigvee_{j=1}^{r} \beta_{i_j},$$

for some distinct atoms β_{i_j} of $SL(p_{i_1}, \ldots, p_{i_n})$, so by (P1)–(P2) and (ii), for w a probability function on $SL(p_{i_1}, \ldots, p_{i_n})$,

$$w(\psi) = \sum_{j=1}^{r} w(\beta_{i_j}).$$

By (i) then,

$$\sum_{j=1}^{2^n} w(\beta_j) = 1,$$

and we see that w is determined by

$$\langle w(\beta_1), \ldots, w(\beta_{2^n})\rangle \in \left\{ \overrightarrow{x} \,|x_1, \ldots, x_{2^n} \geq 0,\ \sum_{j=1}^{2^n} x_j = 1 \right\},$$

and conversely every $\overrightarrow{x}$ in this set determines a unique probability function w, a fact we mentioned earlier.

For w a probability function on $SL(p_{i_1}, \ldots, p_{i_n})$ with $\varphi \in SL(p_{i_1}, \ldots, p_{i_n})$, $w(\varphi) \neq 0$, the conditional probability function $w(-|\varphi) : SL(p_{i_1}, \ldots, p_{i_n}) \to [0, 1]$ is, as usual, defined by

$$w(\theta \mid \varphi) = \frac{w(\theta \wedge \varphi)}{w(\varphi)}.$$

To avoid any problems in conditionals with possibly zero denominators we shall adopt the convention in what follows that expression of the form

$$w(\theta|\varphi) = \gamma,$$

stand for

$$w(\theta \wedge \varphi) = \gamma\, w(\varphi).$$

Formally, a *Knowledge Base* on $SL(p_{i_1}, \ldots, p_{i_n})$, in the sense of this paper, is a set finite set of constraints,

$$\{w(\theta_i) = \gamma_i \mid i = 1, \ldots, m\} + \{w(\varphi_i|\psi_i) = \delta_i \mid i = 1, \ldots, r\},$$

where the $\theta_i, \varphi_i, \psi_i \in SL(p_{i_1}, \ldots, p_{i_n})$ and the γ_i, δ_i are real, which is consistent, i.e., satisfied by some probability function w_0 on $SL(p_{i_1}, \ldots, p_{i_n})$ (or, equivalently, any larger language). If $r = 0$ here, i.e., there are no conditional constraints, we call such knowledge bases 'elementary'. We use $CL(p_{i_1}, \ldots, p_{i_n})$ ($CEL(p_{i_1}, \ldots, p_{i_n})$) etc. to denote the set of such (elementary) knowledge bases. Notice that if $\{p_{i_1}, \ldots, p_{i_n}\} \subseteq \{p_{j_1}, \ldots, p_{j_s}\}$ then $CL(p_{i_1}, \ldots, p_{i_n}) \subseteq CL(p_{j_1}, \ldots, p_{j_s})$.

We now introduce the notion of an *inference process*, that is a function, N, such that for any finite nonempty subset $\{p_{i_1}, \ldots, p_{i_n}\}$ of L and

$K \in CL(p_{i_1}, \ldots, p_{i_n})$, $N(\{p_{i_1}, \ldots, p_{i_n}\}, K)$ is a probability function on $SL(p_{i_1}, \ldots, p_{i_n})$ which satisfies K.

The idea behind the definition of an inference process is that we are identifying a rational agent's knowledge in a particular area with a set of constraints $K \in CL(p_{i_1}, \ldots, p_{i_n})$ and supposing that, when required to assign (subjective) probabilities on the basis of K, the agent does so in such a way that they are mutually consistent with K and with each other. Since these values together determine a probability function on $SL(p_{i_1}, \ldots, p_{i_n})$ the agent's assignment procedure is effectively *picking* a probability function to satisfy K. In other words we can, for our purposes, identify the agent with an inference process, N.

The value of doing this is that 'common sense' requirements on the agent's assignment process can be captured in terms of *principles* which the function N should ideally satisfy. (It is important to appreciate in these principles that the constraint set K is supposed to sum up *all* the (assigning agent's) knowledge (in the particular area).) We refer the reader to (Paris and Vencovská 1990) and (Paris 1994) for a fuller discussion and justification of these principles.

For the purpose of this paper we are interested in the *Maximum Entropy Inference Process*, ME, which is defined as follows: Given a language with propositional variables $\{p_{i_1}, \ldots, p_{i_n}\}$, $K \in CL(p_{i_1}, \ldots, p_{i_n})$, let $\beta_1, \ldots, \beta_{2^n}$ run through the atoms of $SL(p_{i_1}, \ldots, p_{i_n})$. Then $ME\ (\{p_{i_1}, \ldots, p_{i_n}\}, K)$ is that probability function w on $SL(p_{i_1}, \ldots, p_{i_n})$ satisfying K for which the entropy,

$$-\sum_{i=1}^{2^n} w(\beta_i) \log w(\beta_i),$$

is maximal.

We are now ready to state formal versions of our 'common sense' principles of uncertain reasoning given, informally, earlier. These have their origins in (Paris and Vencovská 1990) although the versions given here incorporate a number of simplifications proved in (Paris 1994; Paris and Vencovská 1996a, 1997). We shall state these principles for N defined on CL. The analoguous principles for CEL, which we shall refer to in the main theorem, are obtained simply by replacing CL throughout by CEL (and dropping any conditional constraints).

IRRELEVANT INFORMATION PRINCIPLE. Let $K_1 \in CL(p_{i_1}, \ldots, p_{i_n})$, $K_2 \in CL(p_{j_1}, \ldots, p_{j_m})$ with $\{i_1, \ldots, i_n\} \cap \{j_1, \ldots, j_m\} = \emptyset$. Then for $\theta \in SL(p_{i_1}, \ldots, p_{i_n})$,

$$N(\{p_{i_1}, \ldots, p_{i_n}\},\ K_1)(\theta) =$$

$$= N(\{p_{i_1}, \ldots, p_{i_n}, p_{j_1}, \ldots, p_{j_m}\},\ K_1 + K_2)(\theta).$$

The principle of Irrelevant Information as presented here provides us with a very valuable simplification. Namely, by taking K_2 to be empty we see that for N satisfying this principle $N(\{p_{i_1}, \ldots, p_{i_n}\}, K_1)(\theta)$ does not depend on the particular overlying language $\{p_{i_1}, \ldots, p_{i_n}\}$ chosen (a property known as *Language Invariance* in earlier papers (Paris and Vencovská 1989, 1990; 1997; Paris 1994)). Since we are interested in inference processes satisfying this principle, we shall henceforth therefore omit explicit mention of the argument $\{p_{i_1}, \ldots, p_{i_n}\}$ of N whenever this does not cause confusion.

EQUIVALENCE PRINCIPLE. If $K_1, K_2 \in CL(p_{i_1}, \ldots, p_{i_n})$ are equivalent in the sense that a probability function w satisfies K_1 just if it satisfies K_2, then $N(K_1) = N(K_2)$.

RENAMING PRINCIPLE. Let $g : B(p_{i_1}, \ldots, p_{i_n}) \cong B(p_{j_1}, \ldots, p_{j_n})$ and suppose that $K_1 \in CL(p_{i_1}, \ldots, p_{i_n})$, $K_2 \in CL(p_{j_1}, \ldots, p_{j_n})$ are such that

$$K_1 = \{w(\theta_i) = b_i | i = 1, \ldots, m\} + \{w(\theta'_j | \theta''_j) = c_j | j = 1, \ldots, r\}$$

$$K_2 = \{w(\varphi_i) = b_i | i = 1, \ldots, m\} + \{w(\varphi'_j | \varphi''_j) = c_j | j = 1, \ldots, r\}$$

where $g(\overline{\theta_i}) = \overline{\varphi_i}$ for $i = 1, \ldots, m$, and $g(\overline{\theta'_j}) = \overline{\varphi'_j}$, $g(\overline{\theta''_j}) = \overline{\varphi''_j}$ for $j = 1, \ldots, r$. Then $N(K_1)(\theta) = N(K_2)(\varphi)$ whenever $g(\bar{\theta}) = \bar{\varphi}$.

RELATIVISATION PRINCIPLE. Suppose that $K_1, K_2 \in CL(p_{i_1}, \ldots, p_{i_n})$ are respectively the sets of constraints

$$\{w(\theta_j \wedge \varphi) = b_i, w(\varphi) = c,\ w(\psi_t \wedge \neg\varphi) = d_t | i = 1, \ldots, m,$$

$$t = 1, \ldots, r\},$$

$$\{w(\theta_j \wedge \varphi) = b_i,\ w(\varphi) = c | i = 1, \ldots, m\}.$$

Then for $\theta \in SL(p_{i_1}, \ldots, p_{i_n})$, $N(K_1)(\theta \wedge \varphi) = N(K_2)(\theta \wedge \varphi)$.

OBSTINACY PRINCIPLE. If $K_1, K_2 \in CL(p_{i_1}, \ldots, p_{i_n})$ and $N(K_1)$ satisfies K_2, then $N(K_1) = N(K_1 + K_2)$.

WEAK INDEPENDENCE PRINCIPLE. If K_1, K_2 are, respectively, the sets of constraints

$$\{w(p_1) = a,\ w(p_2) = b\},$$

$$\{w(p_1) = a,\ w(p_2) = b,\ w(p_3) = c,\ w(p_1 \wedge p_3) = 0\},$$

then $N(K_1)(p_1 \wedge p_2) = N(K_2)(p_1 \wedge p_2)$.

CONTINUITY PRINCIPLE. For $K \in CL(p_{i_1}, \ldots, p_{i_n})$,

$$K = \{w(\theta_i) = b_i | i = 1, \ldots, m\} + \{w(\varphi_j|\psi_j) = c_j | j = 1, \ldots, r\}$$

and $\chi \in SL(p_{i_1}, \ldots, p_{i_n})$, $N(K)(\chi)$ is a continuous function of the b_i, $i = 1, \ldots, m$.

(This version of Continuity is somewhat weaker than that appearing in (Paris and Vencovská 1989, 1990, 1997; Paris 1994).)

All these principles hold for ME. Indeed, by combining, and sharpening slightly, the results in (Paris and Vencovská 1990, 1996b, 1997) we have the following result characterising ME (for three alternate characterisations with a similar 'axiomatic' flavour see the work of Shore and Johnson (1980); Csiszár (1989); and Kern-Isberner (forthcoming).

THEOREM. Let N be an inference process which satisfies the principles of Irrelevant Information, Equivalence, Renaming, Relativisation, Obstinacy, Weak Independence and Continuity on $CL(CEL)$. Then N agrees with ME on $CL(CEL)$.

A natural question to ask at this point is whether we might give be able to give a *mathematical formulation* of the Symmetry Principle and prove our principles from that, rather than relying on a host of different interpretations of what is meant by 'similar' in order to derive them. Unfortunately, we have been unable to find such a master principle. One, ostensibly reasonable, candidate is the following:

MARK I SYMMETRY PRINCIPLE. Let B be a subalgebra of the Lindenbaum algebra $\overline{SL(p_{i_1}, \ldots, p_{i_n})}$ and let $K_1 \subseteq K_2 \in CL(P_{i_1}, \ldots, p_{i_n})$ be such that the set of constraints

$$\{w(\theta) = a_\theta \mid \bar{\theta} \in B,\ a_\theta = N(K_1)(\theta)\}$$

is consistent with K_2. Then, for $\bar{\theta} \in B$,

$$N(K_1)(\theta) = N(K_2)(\theta).$$

[Notice this set of constraints is effectively finite.]

Here then the intention is that the 'additional information in K_2 beyond K_1' would not require the probabilities the agent had assigned (using N) to sentences θ, with $\bar{\theta} \in B$, on the basis of K_1 alone, to alter, and so, as far as such θ were concerned, K_2 really added nothing 'essentially new' and so gave no reason for assigning a different value.

Unfortunately this 'symmetry principle' cannot hold for any inference process because it implies the inconsistent *Atomicity Principle* as stated in (Paris 1994). This is, perhaps, rather surprising (disquieting?) given the apparent similarity of the above argument justifying this principle and the one used earlier to justify the Obstinacy Principle on the basis of the Symmetry Principle.

5. CONCLUSION

In this paper we have, by combining a number of earlier results and ideas, demonstrated that in the case where an agent's knowledge K consists simply of a set of values of the agent's subjective probabilities and conditional probabilities there are a handful of common sense principles of uncertain reasoning which, if obeyed, uniquely determine the probabilities the agent may assign on the basis of K. Furthermore, these unique values are precisely those given by the solution to K which maximises entropy.

All these principles may be viewed as special cases of the Symmetry Principle, that *essentially similar problems should have essentially similar solutions*, by considering different interpretations of 'essentially similar'. As such they may be argued to accord with common sense, not only in their own right, but also as children of the emminently common sensical Symmetry Principle.

This, of course, raises the question of how and why we judge something to be 'common sense', a question we have largely glossed over, taking it to be self evident that our principles have this property. Certainly it would be comforting to have some precise, mathematical perhaps, and widely accepted, formulation of what constitutes common sense.

NOTES

[1] Notice that we do not allow assertions of statistical independence in our knowledge bases.

REFERENCES

Csiszár, I.: 1989, 'Why Least Squares and Maximum Entropy? An Axiomatic Approach to Inverse Problems', *Mathematics Institute of the Hungarian Academy of Sciences*, Preprint No. 19/1989.

Kern-Isberner, G.: forthcoming, 'Characterising the Principle of Minimum Cross-Entropy within a Conditional-Logical Framework', to appear in *Artificial Intelligence*.

Maung, I. and J. B. Paris: 1990, 'A Note on the Infeasibility of Some Inference Processes', *International Journal of Intelligent Systems* **5**, 595–604.

Paris, J. B.: 1994, *The Uncertain Reasoner's Companion – A Mathematical Perspective*, Cambridge University Press, Cambridge, UK.

Paris, J. B. and A. Vencovská: 1989, 'Maximum Entropy and Inductive Inference', in J. Skilling (ed.), *Maximum Entropy and Bayesian Methods*, Kluwer Academic Publishers, pp. 397–403.

Paris, J. B. and A. Vencovská: 1990, 'A Note on the Inevitability of Maximum Entropy', *International Journal of Approximate Reasoning* **4**, 183–224.

Paris, J. B. and A. Vencovská: 1996a, 'Principles of Uncertain Reasoning', in A. Clark et al. (eds.), *Philosophy and Cognitive Science*, Kluwer Academic Press, pp. 221–59.

Paris, J. B. and A. Vencovská: 1996b, 'Some Observations on the Maximum Entropy Inference Process', *Technical Report L1-96*, Department of Mathematics, Manchester University, UK.

Paris, J. B. and A. Vencovská: 1997, 'In Defence of the Maximum Entropy Inference Process', *International Journal of Approximate Reasoning* **17**, 77–103.

Shore, J. E. and R. W. Johnson: 1980, 'Axiomatic Derivation of the Principle of Maximum Entropy and the Principle of Minimum Cross-Entropy', *IEEE Transactions on Information Theory* **IT-26**(1), 26–37.

Van Fraasssen, Bas.: 1989, *Laws and Symmetry*, Clarendon Press, Oxford, UK.

Department of Mathematics
Manchester University
Manchester, M13 9PL
UK