Evaluating Evidence when making Difficult Decisions

David A. Bell and David H. Glass
Faculty of Informatics,
University of Ulster,
Newtownabbey,
BT37 0QB, UK.

Abstract. *This paper presents the case for using the Dempster-Shafer theory of evidence and a derivative of it, to model the reasoning process in debates on difficult philosophical, theological and scientific questions. This gives a useful formal framework within which to enhance the debating process. A well-known theological debate and two scientific exemplars are given for illustration of the working and value of the proposed approach.*

Keywords: Dempster-Shafer, probability, relative evidential reasoning, uncertainty.

1. Introduction

When we are reasoning it is common to assume temporarily that we are absolutely certain of our facts and of the available methods, procedures and rules available for finding their consequences – also assumed to be certain.

For example, given the following assertions……

fact: the object is a swan
rule: swans are white

we could assume that they are absolutely dependable, and conclude that the object is white.

When we look carefully at this assumption we encounter some causes for challenging it. In fact, looked at with detachment, as far as is possible, it would seem to be exceedingly bold of us to make any totally confident assertions about states of affairs and to claim universally acceptable and applicable methods of establishing or recognising the truth of conclusions. There seem to be some ever-present deficiencies which have resisted all attempts at their removal. Our confidence in the conclusion above cannot be complete.

We ignore here the inherent limitation with our truth-finding mechanisms as pointed out by Godel in the 1930's. It is with a different kind of methodological and usage imperfection of our reasoning capabilities that we are concerned in this paper - many assertions cannot be known with certainty, and this manifests itself in a number of specific imperfections that everyone can recognise.

For example, *uncertainty* is ubiquitous. How sure is the fact-provider that the object is a swan? Does the fact-provider have good eyesight? Does he/she know much about birds? Does the rule cover all swans, including those from Australia? Judgement under uncertainty is demonstrably very fallible [1]. Moreover information and knowledge, using the terms in broad senses, are often *imprecise* or *vague*. When you use the word "white" what do you mean? How do we accommodate mud-covered swans or grey cygnets? Also we frequently have different sources or arguments, providing us with *inconsistent* inputs eg. a second fact from another source: "The object is a plastic bag floating on a lake" or "The object is a toy" – might become available from a better placed observer. Inconsistency is "inescapable for rational beings who base their choices on arguments".

Furthermore of course, any observer must admit to a large dollop of sheer *ignorance* when reasoning about world sub-systems of any complexity. Even the humblest decision making might involve objects never encountered before. There may be other facts, rules or possibilities we are not aware of. In addition, in much of our day-to-day thinking and decision making we have to acknowledge the *incompleteness* of our information and knowledge - that required data, or facts or rules are simply not available.

These deficiencies may not be very important for the example above, but what if the fact is “the disease is a cancer” and the stated rule is “ cancer is incurable”? In this case the deficiencies might be highly important – and desirable. Similarly if we are using facts, rules and arguments for some important philosophical discussion, as here, it is important to acknowledge the limitation to the representation of our thoughts, observations and other experiences and of what lies behind our sensations, our reflections, our use of powers of reasoning, our sources of counsel and perhaps our inspiration and our emotions.

There is a positive outcome of our acknowledgement of this persistent deficit. It can be used to verify and confirm a world-model which postulates inherent human limitations in perceptions and mental apparatus. We are forced to acknowledge that we are surrounded by enigmas and incompleteness and that our representations are somewhat under-developed.

At a more mundane level, we could easily imagine becoming overwhelmed by deficiency in practical decision and choice making. Yet somehow this has not happened. Despite uncertainty, aeroplanes regularly arrive at remote destinations with acceptable reliability, many ailments are diagnosed and treatments planned effectively, and even some weather and economic forecasts get it right more often than by chance. The question arises: is there any way we could formalise reasoning processes or otherwise make more visible for practical application how choices are arrived at? For our present purposes we would like the formalisation to help in debates in science, and incidentally in philosophy and religion.

We could hedge a little in our example above by changing the given fact to “this object is almost certainly a swan”, and the rule to “most swans are white”. But what do we mean by these added fuzzy-linguistic terms, and how do we allocate a fuzzy-linguistic [2] or other confidence level to our conclusion? If we use a numerical method for doing this, such as assigning a probability, where do we get the numbers, how do we combine them, and what do they really *mean*? And how do we take account of an uncertain second fact, and perhaps others?

We make some of these concepts a little clearer in the following sections and show how we can usefully include an explicit, elegant and mind-friendly representation of ignorance when making certain types of decision. Two possible advantages could be claimed for this. One, by confirmation of a limited world-model as above, would be helpful in keeping a realistic perspective which explicitly acknowledges limitations. The second would be to provide a modest practical addition to the tool-kit for addressing scientific and other questions which involve choosing between alternatives based on weighing (often limited) evidence for and against them.

The scientific literature includes discussions of many situations which include vast scenarios of time and space, for example, and which give rise to hard-to-answer questions. Examples are: establishing or discrediting a case for some theory on the origin of life on our planet (“what is the likelihood of information needed to generate new complex genetic structures becoming available by natural selection processes?”) and similarly for determining the credibility of some cosmological model (“what is the likelihood of the lumpiness in cosmic background radiation being due to dust cutting down the light that reaches earth?”). There is a degree of speculation here that is not present in frequently repeated lab experiments on a simple pendulum, for example.

We present an argumentation tool in section 2 and demonstrate it in sections 3 and 4 in the analysis of a well-known philosophical problem and two scientific examples, respectively, which have been greatly simplified for the purposes of this paper.

2. Evidential Reasoning

Handling the uncertainty that arises from the limitations in our factual knowledge as well as from the complexity of many decisions is a key feature of a human being’s reasoning ability. However it has serious limitations [1].

It is difficult enough to make judgements in many situations even if the object of our judgement is directly observable. It is an even more daunting task when uncertainty arises because we cannot clearly observe all we would like to.

The approach we outline is to resort to *arguments* to replace observations. However we then encounter the problem of resolving inconsistencies between conflicting statements. In this section we consider how arguments based on evidence items of various strengths can be combined/compared in order to help in decision making (ie choosing between alternative conclusions).

It is usually sensible to take a numerical approach where frequencies are available. The idea here is that if, for example, we know that 'nearly all' cases so far of some common phenomenon - say 95% - have some property, eg all swans are white, then we can expect that a new case will have the property. On a [0,1] interval we might say that our belief here is .95. However there are many fields of investigation for which frequencies are not available, and that is the motivation for the present study.

Our search has been for a simple, perspicacious method of weighing the evidence, based on theory or expert judgement or frequencies, for and against hypotheses has led us to consider the belief function formalism [3,4] of Dempster-Shafer (*DS*). This is a generalisation of the better-known formalism due to Bayes (*B*), so anything that can be done using the *B* formalisation could also be done using *DS*. Like *B, DS* is based on mathematical probability, but application to questions of interest is indirect. As a matter of fact, we have taken a step beyond *DS* in the Graded Relational Evidence (*GRE*) method in our search for a simple but effective argumentation method that takes full account of all available evidence. We do this by simply *comparing* the strengths of evidence statements rather than *measuring* them and allocating a number such as a probability. However we wish to focus attention on the *DS* numerical method at this point.

DS often allows us to conduct arguments that need fewer numerical inputs than *B*. But the main advantage of *DS* over *B* is that it does not require us to distribute our total probability over the elements of U, the Universe of Discourse or frame of discernment. We can withhold a portion of our belief, allocating it to *ignorance*. Furthermore, *B* uses *the insufficient reasoning principle* - in the absence of discriminating evidence, it encourages us to distribute our probability uniformly over the contenders. *DS* says ' NO! - only do this if there are definite grounds for doing so'.

In *DS* we make probability judgements on the basis of individual pieces of evidence, one by one. We may do the same with several items of evidence and combine the judgements. *B* requires assessment of degrees of belief on the basis of "all background knowledge" , whether it exists or not.

To introduce the *DS* approach, following [4], suppose we are trying to assess the implications of some evidence on the authenticity of a manuscript claimed to be a work authored by Newton. We have testimony from an expert that the text is authentic, and this induces us to attribute a probability of 80% to the statement "the text is authentic". We say then that we have 80% confidence in it, or that the expert's testimony supports it to a degree of 80%. Based on our confidence in the expert we translate his testimony into a degree of belief about authenticity of paper.

What about the other 20% of our belief? – using the *DS* approach we do not say , as Bayesians would, that "it is against the hypothesis about the manuscript - it is not in fact by Newton", but say that there is a 0 degree of belief to the alternative hypothesis for which we have no evidence. In fact in this situation the sensible thing to do is to withhold the leftover 20% of our belief - we allocate it to *ignorance*.

Similarly, suppose we get some new evidence – a particular observation combined with a piece of theory - which, we are 70% certain, could only be the result of the paper under consideration being authentic. How can we combine the 70% with the 80% earlier?

We could reason as follows. The 2 pieces of evidence are independent, so by high school mathematics, we can multiply probabilities to get four contributions to our decision.

(.8 x .7 = .56)	both items of evidence are reliable
(.8 x .3 = .24)	expert testimony is reliable, new evidence is not
(.2 x .7 = .14)	expert is not reliable, but new evidence is

(.2 x .3 = .06) neither item is reliable (this is the unassigned portion of belief after commitment to the other possibilities).

If at least 1 item is reliable, we have probability 0.56 + 0.24 + 0.14 or 0.94 that the document is authentic. This is the basis of the *orthogonal sum* operation.

We still have 0 belief that the text is not authentic – the 0.06 component of our belief is assigned to *ignorance*. When evidence items conflict the Dempster-Shafer formalism tells us to eliminate impossible outcomes and rescale the other 3 so that they become 1. The method is much more general than illustrated by this, but the details are beyond the scope of this paper (see [3]).

Incidentally much the same evaluative power can be gained without the allocation of numbers, by merely comparing strengths of arguments (asking questions such as: "is the combined weight of evidence items e_1 and e_2 taken together more than that of some item of evidence e_3 against our hypothesis?").

We have argued elsewhere [5] that such non-numeric methods are just as powerful as *DS*, eg, in many practical decision-making instances. We used numbers in our example above, but the relative evidence strengths were really what we were interested in as we came to our conclusion. The Graded Relative Evidence (*GRE*) uses a five-point scale of grades of evidence/argument strengths {*very weak*, *weak*, *average*, *strong*, *very strong*} rather than the numeric interval [0,1], used in *DS*. The advantage of this coarsening is that the reasoning is more accessible to non-experts. The method is theoretically grounded in the *DS* [6], and it provides a simple way of trading-off evidence – eg 2 *very weak* arguments balance 1 *weak* one, and a *very weak* argument together with a *strong* one balance a *very strong* one.

3. Application to a philosophical/theological debate

Using the *DS* theory of evidence we can combine 2 (or more) pieces of evidence that support a particular hypothesis as in the previous section. This gives us an extremely useful means of weighing evidence in philosophical/theological or scientific debates. Our suggestion is that it should be used as a formal framework to insist on discipline when considering the choices available in such debates.

In this section we illustrate our approach by a detailed example which has the feature of controversy that often characterises such discussions – but more importantly in it we must acknowledge much ignorance and often nebulous evidence. We look at it in more detail than the 2 examples in section 4 in order to demonstrate the full power and breadth of the approach.

This first example is worked through in some detail and it shows a number of different aspects of the value of (in particular) the *DS* evidence theory in this context. In the other two examples, in section 4, *GRE* is used to model the arguments in two scientific papers.

In the present debate we are concerned with an assessment of the evidential strengths of the arguments for the existence of God put forward by Swinburne [7] which could be summarised in the following words of Newman [8]:

" is not the being of a God reported to us by testimony, handed down by history, inferred by an induction process, brought home to us by a metaphysical necessity, urged on us by the suggestions of our conscience"

Swinburne's 'evidence' is accumulated to give the following more detailed pieces of evidence. They are all positive - he discards another two items, one negative and one positive, for stated reasons. We ignore this preprocessing (and any other pieces of evidence that others might suggest as important), as our aim is to demonstrate weighing of evidence, which has the incidental advantage of providing a succinct way to help in the assimilation of probabilistic arguments as a whole. The working assumption for the purposes of this illustration is that what Swinburne has assembled in the 'evidence base' is complete. So any conclusion that we come to will be along the lines – " if we accept Swinburne's evidence, here is a conclusion we can come to".

e_1 Cosmological - existence of a perceived universe

e_2 Teleological - conformity to order

e_3 Anthropological - existence of conscious beings

e_4 Human opportunities for co-operation in acquiring knowledge about the universe

e_5 The pattern of history

e_6 The existence of miracles

e_7 The occurrence of religious experience.

If we apply the orthogonal sum operation to the first two pieces of evidence, we reflect the reasoning of the example on the manuscript claimed to be by Newton above, and end up with four contributions to our decision. However, if instead of assigning definite numbers to the evidence strengths, we use e_1, and e_2 to denote the levels of support the items of evidence give to the hypothesis, we get a clear view of the structure of our arguments. The final one of the four contributions, corresponding to the 0.06 (that neither manuscript is reliable) is $(1-e_1)$ $(1-e_2)$. This is allocated to ignorance. If we get more pieces of evidence and end up with, say, 7 (see below) and follow the same reasoning procedure, then $(1-e_1)$ $(1-e_2)$ .. $(1-e_7)$ of our belief is "withheld" (ie allocated to ignorance).

Now, suppose that all pieces of evidence have a one-in-ten degree of support for our hypothesis – ie we can replace the variables e_i by numbers $e_1 = e_2 = \dots e_7 = 0.1$. Consider the effect of combining 2 pieces of evidence, then adding the third, and so on, stopping only when we get to a situation where < 50% of our belief is allocated to ignorance. Suppose we had just 2 items and $e_1 = e_2 = 0.1$. Then, by applying the *DS* rule as above, 0.81 is added to ignorance. For 3 items with $e_1 = e_2 = e_3 = 0.1$, 0.729 of our belief is allocated to ignorance. We get less than 50% ignorance when the 7th piece of evidence is added. Roughly speaking, this is the point at which the balance of evidence tips in favour of the hypothesis (as opposed to reserving judgement).

Now if $e_1 = e_2 \dots = 0.2$ we require only 4 items of evidence to get below 50% ignorance, but if $e_1 = e_2 \dots. = 0.01$, we would need 69 items before we could have < 50% ignorance. For what it's worth for a grasp of the scales involved in this, if the evidence strengths were 0.001 each we would need 693 items to cut ignorance down to this level.

We cannot allocate numbers like 0.1, 0.2, 0.01 to our teleological, cosmological, etc evidence above because we have little to compare them with. This is also the case in the sorts of scientific debates we are concerned with in this paper (examples are given in section 4.2 below). A frequentist approach is often not possible. (The evidence in our present example is by common consent even less tangible than in many of these scientific situations.)

But what we could do is make statements like:

"given the task of allocating support levels based on the seven given items of evidence e_i, we would only require 0.1 confidence in each to make the balance of evidence tip in favour of existence".

Well, is the level of 0.1 reasonable for each?
We leave this as an exercise for the reader!

Remember, of course, that we have ignored any additional evidence that might be put forward in support of the alternative hypothesis. Our purpose here is simply to illustrate the value of having a formal framework of *DS* theory to focus our thoughts in a disciplined manner. In the examples in section 4, contradictory evidence is accounted for.

While a result such as that found above can help us in our acceptance or otherwise of the hypothesis, it is not, in my opinion the best we can do using the *DS* theory of evidence here.

Indeed, suppose that we have precisely evidence items $e_1 \dots e_6$ as our evidential input, and also suppose that together, considering them collectively as a single piece of evidence, we have a very small confidence of, say, one in a thousand, 0.001, in the support they give for our hypothesis. An interesting little result can be obtained for the purposes of illustration if we are in a position to allocate 50% to the seventh piece of evidence - our own or others' experience - and combine the two items as follows:

6 items	e_1 ..	H	
→	e_6		….U
		0.001	0.999
experience			
|	$H^{0.5}$	0.0005	0.4995
|			
V	$U^{0.5}$	0.0005	0.4995

This indicates that since, under these conditions, less than 50% of our belief is allocated to ignorance, a rational choice would be to accept the hypothesis on the balance of the evidence.

We can put this into common parlance as follows:

“ Assuming all evidence has been submitted, if we have even a whisper of positive support from the first 6 pieces of evidence for our hypothesis, then merely sitting on the fence on our evaluation of the support offered by our own, and others’, religious experience would be sufficient to make the belief in existence of God rational”.

But, of course, we must emphasise again that we have not attempted to evaluate e_1 – e_6 properly in this paper, nor have we sought to enumerate all of the e_i s that could be available. Our objective is simply to demonstrate how the evidential reasoning approach can help us get some insights into the aggregate value of points made in debate by evaluating arguments, possibly put forward by others, in this case Swinburne.

Incidentally, if we use the calculus for evidence accumulation proposed for the *GRE* method [6], we would probably assess our evidence as follows:

e_1 Cosmological - existence of a perceived universe………*very weak*
e_2 Teleological - conformity to order ………*very weak*
e_3 Anthropological - existence of conscious beings ………*very weak*
e_4 Human opportunities for co-operation ………*very weak*
e_5 The pattern of history ……….*very weak*
e_6 The existence of miracle ………*very weak*
e_7 The occurrence of religious experience. ………*average strength*

In *GRE* the idea is to *trade-off* the pairs of arguments *pro* and *con* H, using the calculus. In the present example it is reasonable to retain our hypothesis as there are simply no items of contrary evidence in this (limited) study.

4. Two scientific examples

In this section we illustrate the broad use of *GRE* in scientific reasoning. The emphasis is now on comparing arguments rather than accounting for ignorance. The assessments of the evidence strengths are not meant to be authoritative!

Our first example is about “The Hominids of East Turkana” [9], discussed in an evidential reasoning context by Shafer [10]. It concerns evidence and arguments for choosing between hypotheses about the types of skull found near Lake Turkana in Kenya and the number (up to 3) of species they represent.

A theoretical argument (e_1) supported the conclusion B1 – “all one species”; absence of type I and type II among type III elsewhere (e_2) supported B2 {“I&II ;III” } (read as “ I and II are varieties of one species, III is a different species”); and “differences between types in the pair i,j” (e_{ij}) support various conclusions as summarised below.

e_1 -> B1	(1 *average* argument);
e_2 -> B2, B5	(2 *average* arguments);
e_{12} -> B3,B4,B5	(3 *weak* arguments);
e_{23} -> B2, B4, B5	(3 *average* arguments);
e_{13} -> B2, B3, B5	(3 *strong* arguments).

Where B3 is {“II&III;I”}; B4 is {“I&III;II”}; B5 is {“ 3 different species”}. These weights were deduced from the facts given in the original article.

B5 is the best supported conclusion with 4 arguments – 3 of which match those of the next best conclusion, B2. This roughly reflects the reasoning and conclusion of the palaeontologists.

The second scientific example is about the existence or otherwise of Einstein’s “Cosmological Constant” [11]. This concerns contenders for explaining the apparent changes in the expansion rate of the universe. We simplify the discussion somewhat by considering only two hypotheses or conclusions: (C) “ The universe is open, or it is made flat by some added form of energy not associated with ordinary matter”. The alternative hypothesis is that C is false.

A strong theoretical argument gives evidence (e_1) against C, but Krauss argues that due to a series of observations made recently, each weakly supporting C, the balance has swung in the direction of a “theoretically perplexing universe” – that is, to support C.

There are three types of observation: of the age of the universe (e_2), the density of matter (e_3) and the nature of cosmic structures (e_4). All contradict (very weakly) a flat universe, and an argument from Quantum Mechanics where virtual particles are unseen but have measurable effects adds further (weak) evidence (e_5).

e_1 -> *not* C	(*strong*);
e_2 -> C	(*very weak*);
e_3 -> C	(*very weak*);
e_4 -> C	(*very weak*);
e_5 -> C	(*weak*).

Using the *GRE* calculus, in which 4 *very weak* arguments would exactly balance a *strong* one, we get a conclusion similar to that of Krauss – that the support for C now just about outweighs that for the alternative.

5. Summary and Conclusion

We claim that our decision making procedures of the previous sections have both descriptive and practical utility in their application to discussing reality. They provide a discipline for constructing and evaluating arguments for and against things we may be inclined to believe in.

A secondary output from the discussion here is to enforce the conclusion that the confidence that we have in our conclusions in some areas such as the philosophical and cosmological examples in section 4 are of necessity limited. The ignorance we dealt with explicitly in section 3 is compounded by the fact that we may be omitting some pieces of evidence that are considered important in some quarters.

Obviously the conclusions obtained in these examples are of limited value and the weights could certainly be criticised constructively. There is much work that could be done to polish up the methods and extend their applicability.

It is a basic feeling shared by many serious thinkers that there is substantial value having a systematic and defensible method for evaluating whether some beliefs that are justified by evidence, and the contribution of this paper is in that direction. An observation after conducting the three exercises is that much of our evidence does not include direct and compelling sense data warrants, and that some of the resulting beliefs might even be more credible than ones based on data alone!

6. References

[1] A Tversky and D Kahnemann. Judgement under Uncertainty. *Science*, 185: 1124-1131, 1974.

[2] W Dubitsky, D Bell et al. How Old is 43 Years of Age?, in *Proc IJCAI Conf* 1997.

[3] J Guan and D Bell. *Evidence Theory and its Applications*, vol 1and 2. North-Holland, 1991, 1992.

[4] G Shafer and R Srivastava. The Bayesian and Belief-Function Formalisms. In *Uncertain Reasoning.* ed. by G Shafer and J Pearl. Morgan-Kaufmann, pp. 482-519, 1990.

[5] Z An, D Bell and J Hughes. A Non-computational Evidence Accumulation Structure, in *Proc 3rd Int Conf IPMU*, 1990.

[6] D Bell, Graded Relational Evidence, (to be published)

[7] R Swinburne. *The Existence of God.* Oxford, 1992.

[8] J H Newman. *The Idea of a University*. Blackwell,1855.

[9] A Walker and R Leakey. The Hominids of East Turkana. *Scientific American*, 239(2): 44-56, 1978.

[10] G Shafer. Languages and Designs for Probability Judgements. *Cognitive Science*, 9: 309-339, 1985.

[11] L M Krauss. Cosmological Antigravity. *Scientific American*, 280(1): 34-41, 1999.