Probability is Perfect, but We Can't Elicit it Perfectly

Anthony O'Hagan & Jeremy E. Oakley
Department of Probability and Statistics,
University of Sheffield, UK

October 10, 2002

1 Introduction

The challenge problems set out in Oberkampf, Helton, Joslyn, Wojtkiewicz and Ferson (this issue), hereafter referred to as OHJWF, raise a variety of interesting questions. The background to this workshop is the continuing growth in number, diversity and complexity of computer simulation codes being built to model a huge variety of important real-world systems. Users of such models are increasingly demanding to know about the uncertainties in the model outputs. Nobody seriously believes that the models are perfect. The model output will not predict exactly the real-world system behaviour, but how close can we expect it to be?

There are several possible contributory causes of the error in a model's predictions, and in order to understand and quantify the uncertainties in predictions it is important to recognise and analyse each of these components of uncertainty. One taxonomy of uncertainties in computer code outputs is given by Kennedy and O'Hagan (2001), and it is worth reviewing the main components briefly here.

- **Parameter uncertainty**. There is typically uncertainty about the values of some of the computer code inputs. We can think of those inputs as unknown parameters of the model. Generally, they specify features of a particular application context, but they may also be more global parameters, assumed to have a common value over a range of contexts or even in all contexts. They may even specify features of the model structure, and so serve to switch between different versions of the model.

- **Model inadequacy**. No model is perfect. Even if there is no parameter uncertainty, so that we know the true values of all the inputs required to make a particular prediction of the process being modelled,

the predicted value will not equal the true value of the process. The discrepancy is model inadequacy.

- **Residual variability**. The model is supposed to predict the value of some real process under conditions specified by the inputs. In practice, the process may not always take the same value if those conditions are repeated. We call this variation of the process even when the conditions are fully specified, residual variability.

- **Code uncertainty**. For some uses of computer codes it is not practical to run the code at all the relevant input configurations. An example is uncertainty analysis, where we wish to propagate uncertainty in inputs through the model, so are in principle interested in running it at all possible input configurations. One practical solution is to use a Monte Carlo approach (e.g. Latin hypercube sampling), and the sampling error in the Monte Carlo estimates is an instance of code uncertainty. There is an analogous uncertainty in estimates produced using the Bayesian method of Haylock and O'Hagan (1996).

OHJWF's challenge problems address only output uncertainty due to parameter uncertainty. This is a reasonable simplification for the purpose of focussing the debate in this workshop, but it is necessary to be aware of the other sources of uncertainty. In particular, any analysis that does not recognise model inadequacy is likely to grossly understate the probable magnitude of prediction errors. Furthermore, methods that may superficially appear practicable for quantifying parameter uncertainty may not look so good when we widen the discourse; how, for instance, would a proponent of fuzzy logic, Dempster-Shafer theory, etc., express the uncertainty due to model inadequacy?

Another restrictive simplification in the challenge problems is the assumption that all uncertain parameters are independent. The assumption will rarely hold in practice, and the uncertainty distribution of outputs can be very strongly influenced by correlation in model inputs. Effective elicitation and representation of dependence structures between parameters is an important but largely unexplored research area.

This paper is organised as follows. In Section 2, we consider aleatory and epistemic uncertainty, concluding that while the distinction between these is useful in practical terms, the boundaries are blurred. Indeed, a purely aleatory uncertainty is a rather rare beast, and may even be mythical. Section 3 considers the appropriate way to describe uncertainty. It is very firmly our opinion that the uniquely suitable construct is probability. However, we must recognise that in practice it is not feasible to quantify uncertainties precisely in probabilistic terms, and it is necessary to acknowledge imprecision in probability judgements. The real question, then, is how to measure

probabilities in practice with the maximum of precision and reliability. This is the elicitation problem.

Section 4 outlines some recent work on elicitation that explicitly allows for the imprecision in experts' judgements. The methodology is applied to one of the challenge problems in Section 5, although we also argue that the formulation of these problems is unrealistic and apparently based on a denial of the possibility of elicitation. The challenges issued by OHJWF address not only how to express a single expert's uncertainty in unknown model inputs, but also what to do when experts disagree, and how to derive and express the induced uncertainty in an output. Section 6 discusses the issue of reconciling information from multiple experts, while Section 7 considers how to propagate parameter uncertainty through the model. Some concluding remarks are given in Section 8.

2 Aleatory and epistemic uncertainties

OHJWF distinguish between aleatory and epistemic uncertainties, and this is useful for practical purposes. Aleatory uncertainties are described as arising from inherent variabilities or randomness in systems, whereas epistemic uncertainties are due to imperfect knowledge. The distinction is useful because epistemic uncertainties are in principle reducible by obtaining more or better information. Therefore if we know what portion of the uncertainty in the model outputs is due to epistemic sources of uncertainty, then we know that in principle this uncertainty is removable (or at least reducible), whereas that part due to aleatory uncertainties is irreducible. Sensitivity analysis (see Saltelli, Scott and Chan, 2001; Oakley and O'Hagan, 2002b) is a tool for exploring how different uncertainties influence the model output, and so can identify the scope for reducing uncertainty due to epistemic sources. It is also worth mentioning value of information analysis, where the consequences of epistemic uncertainties are costed through a utility function, allowing the value of reducing uncertainties to be compared with the cost of gaining that information. This has been used in economic models for the cost-effectiveness of medicines, in order to assess whether more experimental data should be acquired before reaching a decision on a new drug or treatment; see Claxton, Neumann, Araki and Weinstein (2001).

It is instructive to examine each of the four sources of uncertainty listed in Section 1, and to see which can be classified as aleatory or epistemic.

Parameter uncertainty is generally epistemic; we simply do not know what are the correct values for the input parameters. However, sometimes there is an aleatory component where we wish to use the model to predict the real-world process when some of the conditions specified in the inputs are uncontrolled and unspecified. For example, an atmospheric dispersion model predicts the dispersal of pollutants from a chemical accident, based

on parameters such as wind speed and direction. For a risk assessment of the chemical installation, we may wish to identify the risk of pollution reaching some sensitive location in the event of a future accident. We will not know the wind speed and direction, and the risk assessment requires these to be random.

Uncertainty about model inadequacy is unequivocally epistemic.

Now consider residual variability. This might appear to be inherently aleatory, but there are really two sources of uncertainty combined in one here. The process itself may be inherently unpredictable and stochastic, in which case we can regard the variability as aleatory. But it may also be that this variation would be eliminated (or at least reduced) if only we were able to recognise and to specify within the model some more conditions. In the atmospheric dispersion example, even if we know the wind speed and direction, the model will fail to predict exactly the quantity of pollutant deposited at a given location, simply because of the vagaries and unpredictability of the transport of the pollutants. However, we could conceivably improve the model by adding more detail, such as taking account of the detailed topography of the land and the type of vegetation cover at each point (which affects the deposition velocity). A more complex model could remove some of the apparently aleatory residual variability. The removed component was really epistemic, because it related to our lack of knowledge of topography, vegetation cover, and how these affect air flow and deposition.

This raises the question of whether there is any true randomness. This is the age-old argument of determinism; if I knew the exact position and velocity of every particle, could I not predict exactly the future? Heisenberg's uncertainty principle says I can't know both the position and velocity of every particle, but that is clearly an epistemic uncertainty! Is anything aleatory? If randomness is just a consequence of lack of knowledge (magnified by chaotic systems), then all uncertainty is epistemic. The true aleatory uncertainty becomes as much a mythical beast as Roger Cooke's hypothetical squizzel (Cooke, this issue)! In fact, quantum physics suggests that there is true randomness in some phenomena, but does this extend to the scales on which our models operate?

Finally, consider code uncertainty, for instance as it affects the results of an uncertainty analysis. Monte Carlo is obviously aleatory, isn't it? But the uncertainty is reducible by taking a larger Monte Carlo sample, so surely it is epistemic? The resolution of this paradox lies in another very old debate, that between the Bayesian and frequentist schools of statistical inference.

The uncertainty here concerns the objective of the uncertainty analysis, which is some feature of the probability distribution of the output, that is induced by input parameter uncertainty. For instance, it might be the variance of that distribution, a measure of overall output uncertainty. The output distribution itself might be aleatory (if the input parameter uncertainties are all aleatory) or epistemic, but the variance of the distribution

is just a number. That model output variance is not a random thing, and our uncertainty concerning it will definitely be epistemic. The Monte Carlo estimate appears to be driven by aleatory uncertainty because of the way the Monte Carlo method chooses random input configurations. It is a statistical method, and in particular it is a frequentist statistical method. All uncertainties in the frequentist approach have to be aleatory, and are derived from the randomness in the data, in this case the random choice of inputs. The Monte Carlo estimation variance *cannot* therefore describe uncertainty about the model output variance. It is the wrong kind of uncertainty measure. This is one of the many ways in which frequentist theory is unsatisfactory. Frequentist inferences are cleverly phrased so as to appear to make statements about the necessarily epistemic uncertainty that we have concerning parameters of statistical models, yet they are aleatory probability statements and really concern not the unknown parameters but the (known!) data. Confidence intervals are a confidence trick, and significance tests signify only how surprising the data might be.

Frequentist inference only acknowledges the frequency interpretation of probability, whereas Bayesian statistical methods are based on the personal (or subjective) interpretation of probability. The latter readily accommodates epistemic uncertainty. Bayesian prior and posterior distributions describe the epistemic uncertainties in the unknown parameters of the statistical model, and Bayesian inferences are clearly epistemic statements.

The randomness in the Monte Carlo method is as close to being truly aleatory as it is possible to get, yet the uncertainty in the outputs of the uncertainty analysis, such as the output variance, must be epistemic. The aleatory uncertainty in Monte Carlo is only a device. (And doesn't it seem odd to introduce extra uncertainty artificially in this way? Surely we wish to minimise uncertainty? The Bayesian approach of Haylock and O'Hagan, 1996, does not require randomly chosen inputs, and reports its results in honest epistemic form.)

3 Uncertainty, and how to measure it

OHJWF invite us to consider alternatives to probability as the way to describe and measure uncertainties. They assert that "representation, aggregation, and propagation of aleatory uncertainty is well established using traditional probability theory". They therefore invite us to consider how to do this for epistemic uncertainty and mixtures of epistemic and aleatory uncertainty, but fail to offer any reason to suppose that probability theory is not equally well established also for those uncertainties. Why should probability be adequate for one kind of uncertainty but not for another? I am not aware of any work on the foundations of probability that discriminates between the two.

On the contrary, the classic work by Savage (1956) laying the foundations for personal probability was very clearly intended to apply to all kinds of uncertainty. His approach was refined by De Groot (1970), who began by laying down a series of axioms about an individual's actions in the face of uncertainty. These axioms were intended to represent rules to which we would ideally wish to adhere, and behaviour in accordance with them is described as *coherent*. Violating one of these rules would be manifestly bad, either in the sense that it would be clearly illogical or else would lay us open to avoidable financial losses, or both. The logical consequence of these axioms is that the only way to avoid this kind of irrationality is to have a unique representation of uncertainty that follows the laws of probability. That is: only probability is a legitimate representation of uncertainty.

To Bayesian statisticians like De Groot, there is a sub-text that statistical inference must behave like Bayesian inference, in combining information from the data (possibly aleatory) with prior information (invariably epistemic) to produce posterior inferences (also epistemic). To describe epistemic uncertainties, or mixtures of the aleatory and epistemic, by anything other than probability is not an option if we want to be coherent in our actions.

Cooke (this issue) challenges the advocates of alternative representations of uncertainty to define them operationally. We challenge them also to say which of the De Groot or Savage axioms they wish to throw away. In what particular way do they wish to be illogical or to be someone else's money-pump?

Dennis Lindley has been challenging non-Bayesian statisticians in this way for almost half a century, and we make no apology for repeating the challenge here. However, we have always been persuaded more strongly by the argument that 'the proof of the pudding is in the eating'. That is, we are passionate advocates of Bayesian methods not primarily because of axiomatic arguments but because they obviously work better. They address and answer the real questions, and provide a logical over-arching framework for thinking that we find infinitely better than the web of ad-hockery that is frequentist inference. So we issue yet another challenge to the proponents of alternative uncertainty representations, that applying their methods should lead to manifestly sensible results, without the kind of counter-intuitive implications that Cooke cites for possibility theory.

An important question now requires an answer: if probability is so unquestionably the right representation of uncertainty, why are so many people looking for alternatives? After all, frequentist statisticians can at least be excused in the sense that they were not aware of Bayesian methods first and then perversely went in search of something less satisfactory!

We suspect that there are at least two reasons. One is proper intellectual curiosity — what if we change this or that; would we get something better? The pressures of modern academia, in particular, place a premium on at-

tracting the attention of one's peers, and one way to do this is by proposing something radically different. The second reason is that there are a number of ways in which people have expressed dissatisfaction with probability. We are back to the 'proof of the pudding' idea, here: no matter how persuasive the axioms might be, if their implications include some unacceptable consequences then there is something wrong.

However, we believe that all perceived deficiencies of probability are based on fundamental misunderstandings, and most of them concern the practical difficulties of using probabilities. Nobody said it would be easy to avoid being illogical, inconsistent or incoherent; none of us is perfect.

We know that the proper way to describe Your uncertainty about the number of birthday cards You will receive this year is through a probability distribution. That probability distribution should be unquestionably *Your* probability distribution for that uncertain quantity, representing Your beliefs about its likely value with prefect accuracy. Each individual probability, such as the probability that the number will exceed 20, should perfectly describe Your uncertainty about that event. The difficulty is how to measure those probabilities. If the number of cards is N, should Your probability that $N > 20$ be close to 0.1, or to 0.5? If it is close to 0.1, should it be represented as 0.1, 0.11, 0.101, 0.10023075, or what? The theory says that there is a precise value that uniquely represents Your knowledge, but how can You measure it?

Before proceeding further with these questions, let us emphasise that they apply to aleatory uncertainties just as much as to epistemic ones. O'Hagan (1988, Preface) says,

> "The practical question of assigning probability values in real problems is largely ignored in the traditional way of teaching probability. Either the problems are abstract, so that no [numerical] probabilities are required, or they deal with artificial contexts like dice and coin tossing in which probabilities are implied by symmetry, or else the student is told explicitly to assume certain values."

If a biased coin does not have probability 0.5 of falling 'Heads' up, how can we know what the probability actually is?

We do not have nice, accurate, mechanical devices for measuring probability, in contrast to measurement of physical things like length or weight. However, the fact that we cannot measure probabilities to arbitrary accuracy is not an argument for giving up and declaring the exercise impossible. And the fact that it is difficult to do *in practice* it is not in *any* sense an argument against probability being the unique way to represent uncertainty *in principle*.

The problem of measuring a person's probability distribution for one or more uncertain quantities is very familiar in Bayesian statistics as that of

specifying the prior distribution. The process is usually known as *elicitation*, and there is an appreciable body of literature about how to do it. So when it comes to specifying Your probability distribution for the number N of birthday cards You will receive, good elicitation will yield a probability distribution that represents Your beliefs and uncertainty as accurately as possible.

4 Elicitation

A useful introduction to the literature on elicitation is Kadane and Wolfson (1998). There are a variety of purposes for which elicitation may be required, and these lead to differences in methodology, but we will concentrate here on the situation where an analyst (the elicitor or the statistician) is facilitating an expert to formulate the expert's knowledge and beliefs about some unknown parameter (which may be a vector of parameters) in the form of a probability distribution. We will briefly consider other contexts in Section 6.

The fundamental technique for eliciting a probability distribution is first for the expert to specify a (usually small) number of specific features or summaries of the distribution, representing the principal aspects of their beliefs or knowledge of the parameter in question. Then a probability distribution is chosen to fit those elicited summaries. The final choice may be made on the grounds of availability (e.g. using such families of probability distributions as may be known to the analyst or provided in some relevant software) or convenience (e.g. having conjugacy or analytical tractability).

The justification for this apparently cavalier approach is that, once the expert has specified a number of key summaries, then all distributions which fit those summaries (and have the kind of smoothness that one would expect a person's probability distribution to have generally) would be very similar. For instance, once I have said that my beliefs about some integer-valued parameter A are unimodal with modal value 36 and such that I would be surprised to find it less than 20 or greater than 65 (meaning I am about 90% sure that A will lie in this range), then this is well fitted by a negative binomial distribution with parameters 10 and 0.2. Any other distribution satisfying even those few statements will either be very close to that negative binomial distribution or will look odd in some way.

Therefore the consequences of using the chosen distribution rather than any other distribution fitting the specification are likely to be negligible, for most purposes.

Whilst this argument is compelling in a pragmatic way, it leaves many open questions, such as the following.

- How can we be sure that the consequences are negligible, without explicitly evaluating them for all possible choices?

- How do we know which choices are plausible, in the sense that they do not 'look odd in some way'?
- The argument must depend on the expert having specified 'enough' information via the elicited summaries, so how do we know what is enough?
- Are some kinds of summaries better for this purpose than others?

These questions arise because there is uncertainty about what the expert's true probability distribution is and how far it might differ from the chosen distribution. Now in principle that uncertainty should also be represented by probabilities. This route can lead into philosophical wrangling or infinite regress, when we consider the nature of such probabilities and whether they can be elicited without introducing another level of uncertainty.

Such deeper questions can be set aside (although perhaps not really resolved) by adopting the following perspective. The expert's true probability distribution is unknown to the analyst. So we can represent the uncertainty about the expert's distribution as being the analyst's uncertainty. Oakley and O'Hagan (2002a) present a formal Bayesian analysis from this perspective. The analyst's prior beliefs about the expert's probability density function are represented by a prior distribution. These beliefs are then updated by Bayes' theorem, treating the expert's elicited summaries as data. Then the expert's probability density function can be estimated by the analyst's posterior mean, instead of making an arbitrary choice, and uncertainty in that choice can also be represented through the analyst's posterior distribution. More interpretation of this framework is given in Section 6.

Of course, in principle we might now go on to consider the uncertainty about how accurately the analyst can specify his or her beliefs about the expert's beliefs (and so on!), but this has no practical value. The objective of serious elicitation is to ascertain the expert's beliefs accurately, so that the (analyst's) uncertainty surrounding their specification is minimal. To then consider uncertainty about the analyst's specification would clearly be unproductive.

The analyst's prior distribution specifies that the expert's density function is expected to be quite smooth and similar to a normal distribution. However, the formulation is nonparametric, and so allows the expert's density function to take any form whatever, and the analyst's posterior distribution is thereby fully responsive to the expert's elicited summaries.

To illustrate the method, we consider a real elicitation reported in O'Hagan (1998). The expert in this example is actually a panel of experts (see Section 6 for more discussion of the case of multiple experts), and the unknown parameter in question was the shortest distance D (in miles) by road from Newcastle-upon-Tyne to Birmingham (two cities in England). The elicited

summaries are shown as the histogram in Figure 1; details of how they were elicited are given in O'Hagan (1998). For instance the first block in the histogram represents the expert's specification $P(D < 177.5) = 0.175$.

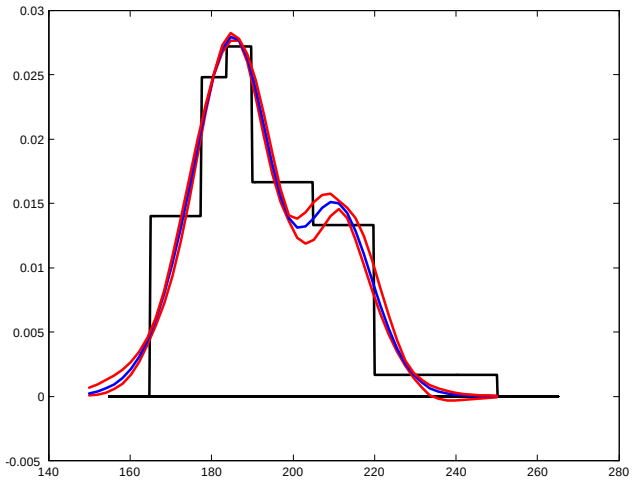

Figure 1. Driving distance example, exact elicitation

The blue curve is the analyst's posterior mean. Notice that, despite the prior belief that the expert's density function should approximate to a symmetric, unimodal normal density, and despite the histogram also being unimodal, this posterior mean is bimodal and asymmetric. The bimodality is a response to the sharp peak in the expert's histogram and the substantial probability that has been elicited for $D \in [205, 220]$ compared with the very much smaller probability for $D \in [220, 250]$.

The red curves are pointwise 95% intervals around the estimate. These show that there is rather little remaining uncertainty concerning the expert's density function, given the expert's elicited probabilities as shown in the histogram. The greatest area of residual uncertainty is over the second mode. The analysis suggests that for many purposes the elicited summaries are adequate to constrain the expert's density function well, but further specification would be desirable if the issue of the second mode is important.

However, all of this analysis is predicated upon the elicited summaries being perfectly accurate expressions of the expert's beliefs. In practice, of course this is an unrealistic assumption. The specified probability 0.175 that $D < 177.5$ is simply a judgement, and a judgement by someone unfamiliar with representing their knowledge in this way. If we allow that the elicited probabilities which were used to construct the histogram in Figure 1 might be inaccurate, with a standard deviation of 0.02 in each case, then we obtain an analysis shown in Figure 2.

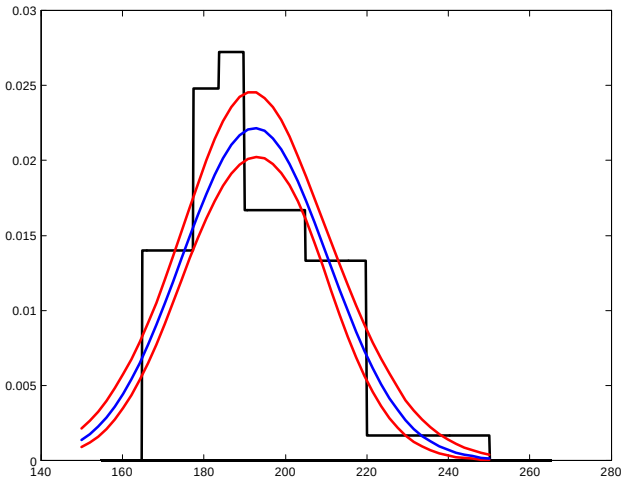

Figure 2. Driving distance example, elicitation subject to error

We now see that the evidence for bimodality has disappeared, and indeed it is plausible that the expert's distribution is very close to normality. The analysis of this example has been done by Delil Gomez Portugal Aguilar, and will be reported in a paper under preparation.

This is very much work in progress. We are working on formulations allowing a stronger belief on the part of the analyst in the expert's density being skew, and on quantifying the consequences of the uncertainty.

5 The 'challenge problems'

5.1 Parameters specified only to lie in an interval

Unfortunately, the challenge problems do not appear to represent a genuine attempt to elicit the beliefs of the experts in any form that could be useful in specifying a probability distribution. In many cases, we are told that the only information about a parameter is that it lies in some interval. How can this be the only information? We cannot imagine how any statistician, with any appreciation of the problem of eliciting probability distributions, could realistically make such an assertion.

For instance, we are apparently supposed to accept that these intervals are absolute, in the sense that the expert in question believes that there is no possibility at all that the parameter might lie outside. (Otherwise, why was the expert not asked to tell us what the chance was of the interval not containing the true parameter value?) We never knew any experts who were happy, on being challenged, to commit themselves to an interval with such an absolute assurance, unless it simply expressed the physical limits of the parameter - e.g. a length is bounded below by 0. Such an interval

conveys no *expert* information at all, and it is hardly sensible to consult an expert in order to elicit only the physical limits of the parameter.

Experts certainly have more information than this. No expert believes that the parameter is equally likely to be right up against its physical limits as it is to take some intermediate value, so why were they not asked to tell us where the parameter was more likely to lie? There are numerous examples of the successful elicitation of quite sophisticated information from experts. If OHJWF only expect to get an interval of this kind from an expert, then we think they are seriously under-using their access to experts!

We therefore believe that to claim that the only information available about a parameter is that it lies in some interval is to deny the possibility of eliciting expert information effectively. Nevertheless, we will accept the challenge and attempt to analyse such a case in the framework of Section 4.

We consider the simplest of the challenge problems, Problem 1 in the algebraic problem set. This problem is adequate to illustrate the general principles. There are two parameters, each asserted only as lying in a single interval. Thus, $a \in [0.1, 1.0]$, $b \in [0.0, 1.0]$. The only other information we have is that they are independent. Figure 3 shows what I have called the analyst's posterior distribution for the expert's density for each of these parameters.

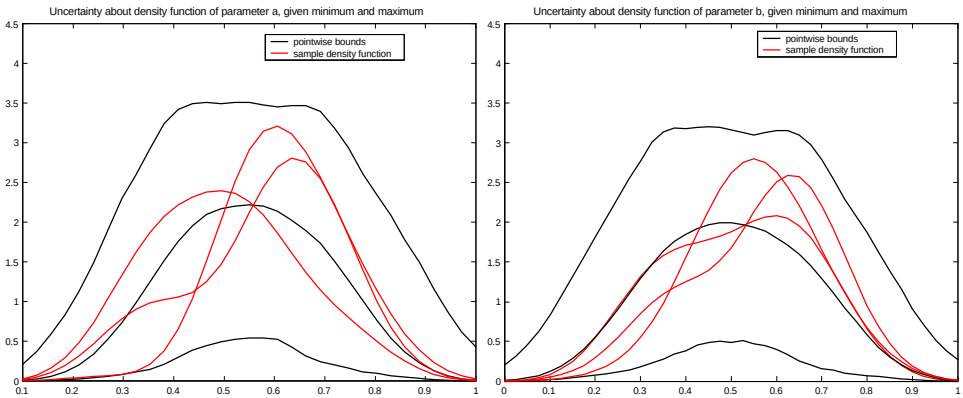

Figure 3. Algebraic challenge problem 1, parameters a (left) and b (right)

The central black curve shows the analyst's posterior mean, while the outer black curves give pointwise 95% bounds. Of course, the two pictures for parameters a and b are theoretically identical, differing only in the scale on the horizontal axis. The red curves show a few sample realisations from the analyst's posterior distribution. There is clearly a great deal of uncertainty remaining in this case, and the single elicited statement is nowhere near adequate to fix the expert's density function.

The context of the challenge problem provides a purpose for the elicitation, which is that the expert's distributions for a and b are to be propagated

through the simple algebraic model $y = (a + b)^a$ in order to specify the induced distribution for y. We do not know the expert's distribution, but we can make random draws from the analyst's distribution for each, as shown by the red curves in Figure 3, and for each pair of distributions we can evaluate the distribution of y. We can then determine whether the expert's knowledge has been elicited adequately for the purpose of quantifying any desired feature of the distribution of y.

Suppose, for instance, that we are interested in the expectation $\mu = E(y)$. Taking many realisations from our model, we obtain a large sample from the analyst's distribution for μ. We find that this has mean 1.05 and variance 0.004. That degree of uncertainty would in practice probably not be considered sufficiently small, and implies again that the elicitation has not been adequate to identify the expert's knowledge about a and b.

A more complete way to view that uncertainty is to examine the cumulative distribution function $F(y)$ of y. Figure 4 shows the analyst's expected value, with upper and lower 95% bounds, for $F(y)$ plotted against y.

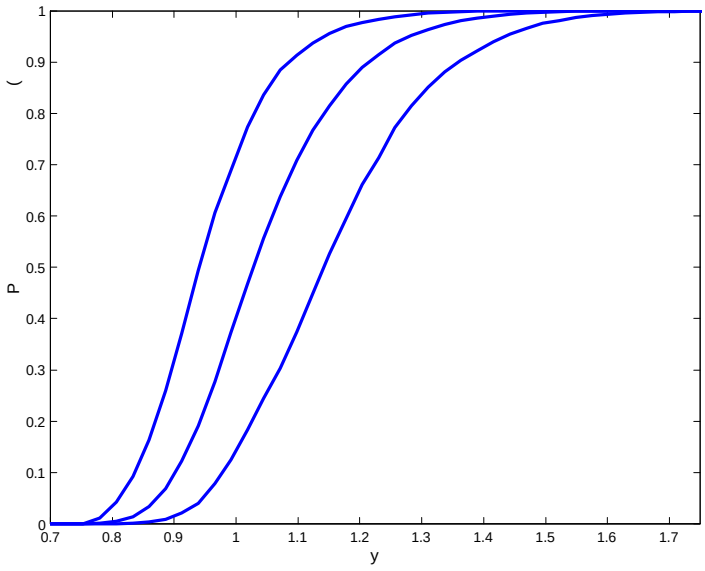

Figure 4. Algebraic challenge problem 1, output distribution uncertainty.

There is a substantial difference between our analysis of this example and most of the others presented at this meeting. This is particularly clear in Figure 4, where our analysis constrains the distribution of y much more than, for instance, the 'p-box' representation of Ferson and Hajagos (this issue). Implicitly, those other analyses take the information that, for instance, $a \in [0.1, 1]$ as allowing a to take absolutely any probability distribution over $[0.1, 1]$ without any notion that some of these distributions are more plausible representations of the expert's knowledge than others. Our analysis

differs precisely because we have added extra information to the problem, which enters through the analyst's beliefs about the expert's probability distribution for a.

- The analyst believes firmly that the expert's distribution admits a continuous density, so the expert's distribution for a is not concentrated on a single point, nor is it discrete and nor does it give non-zero probability to any single value in $[0.1, 1]$.
- The analyst thinks the expert's distribution is more likely to be broadly unimodal than strongly multimodal.
- The analyst thinks it highly unlikely that the expert's probability distribution is such that most of the probability is concentrated in a small subset of $[0.1, 1]$.

We believe that the extra information supplied by the analyst's prior distribution is entirely reasonable and defensible. The reason is simply that if any of the above beliefs were wrong then the expert would not have said that the only information they could give was that $a \in [0.1, 1]$. If any individual point could be singled out as having positive probability, the expert would have said so. If the expert thought that there was a region of relatively improbable values for a between two regions where it was much more likely to lie, he or she would have said so. If almost all the probability was concentrated in a small part of $[0.1, 1]$, the expert would have said so.

There is information in what the expert did not say as well as in what was said.

As a consequence, although Ferson and Hajagos's 'p-box' correctly identifies the absolute bounds on $F(y)$, our Figure 4 shows that it greatly overstates the uncertainty about the expert's information, since our 95% bounds fall well within the 'p-box'.

5.2 More complex cases

The OHJWF challenge problems also contain a number of specifications that we will call hierarchical. We are told that a parameter has a distribution of some form, characterised by some values that are often called hyperparameters. In the hierarchical examples, these hyperparameters are said to lie in intervals. Apart from the fact that the latter part of the specification is a non-problem for the same reasons as given above, we would like to know how such a specification could arise.

What does it mean for someone to say that they know that the uncertainty in this parameter should be described by a probability distribution of a certain form but that they do not know its hyperparameters?

This might be plausible if the probability distribution represents an aleatory uncertainty. It could conceivably be known from data on related parameters that such a distribution is appropriate. Then the hyperparameters might represent real physical quantities about which one might reasonably ask for an expert's knowledge. I do not know how to think about such a model if the uncertainty is not intrinsic. However, if such specifications are always aleatory, how can it be realistic to suggest triangular distributions? We know of no serious real-life context where a random quantity has such a distribution. The triangular distribution is only ever a simplified subjective assessment for an epistemic uncertainty. (It is better than a uniform distribution on an interval, but still suffers from having an unrealistically bounded domain.)

In practice, a triangular distribution results from the expert specifying a range and a mode, and then, following a failure to elicit anything more meaningful, someone assumes a triangular distribution to fit those values. In this spirit, we will apply the method of Section 4 to the specification of just the range and the mode.

Consider the parameter m in the ODE example. It is said to have a triangular distribution with minimum 10, maximum 12 and mode 11. We will use the information about range and mode but will not assume the triangular distribution. Figure 5 shows that the uncertainty now is very different from that shown in Figure 3 for the case of just interval information.

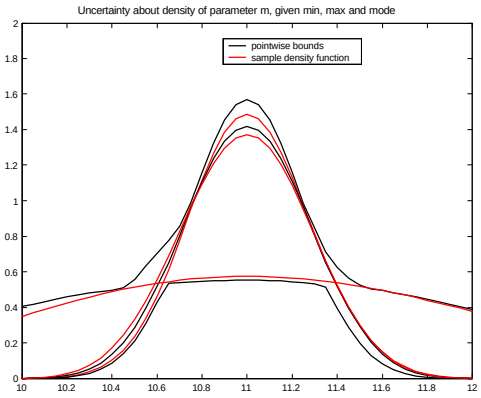

Figure 5. ODE challenge problem, parameter m

The mode at 11 has reduced the uncertainty, particularly around $m = 10.5$ and $m = 11.5$. The greatest element of uncertainty now is about spread, and realisations look like normal distributions just contained in the interval or else normal distributions truncated to fit the interval.

In conclusion, we find it very difficult to take the "challenge problems" seriously, because they are so unrealistic. Nevertheless, we have presented

some attempts to interpret some of the simpler challenge problems in the framework of Section 4.

6 Whose uncertainty?

A reasonable question to ask of the analyses that we have presented is, "Whose uncertainty are we interested in?" Is it the analyst's uncertainty? In practical elicitation there will often be a separation between the expert and the analyst, but it is not the analyst's beliefs that matter. The analyst is there to facilitate the elicitation of the expert's beliefs.

We have presented the method from the analyst's perspective only for ease of exposition. The intention throughout is to represent the uncertainty that remains concerning the expert's distribution for the parameter of interest, given the expert's elicited summaries of that distribution. It is part of the analyst's facilitating role to report back to the expert the implications of the expert's elicitation, and ideally to report (to the expert, or to the client who has commissioned the elicitation) at the end of the process the chosen distribution. We see the method as allowing the analyst to report back also the degree of uncertainty surrounding that distribution.

As explained in Section 5.1, the analyst introduces some extra information through the prior distribution. This was important in the analysis presented there of the case of an interval specification, and we argued that this was appropriate for such a problem. In more realistic elicitation, the process will include feeding back implications of the expert's elicited information, and in the early stages of the elicitation this may be coloured by the analyst's prior. The expert should then have an opportunity to refute any feature which has been introduced by the analyst. Furthermore, as the expert provides more information, so the influence of the prior distribution diminishes, so that we do not believe this feature of our approach has any practical drawbacks.

There are other, related questions about "whose uncertainty?"

It is common for expert information to be elicited to inform a decision-maker, who might be 'the client' in the preceding paragraph. In this case, the decision-maker is really interested in the expert's knowledge, and no matter what form that knowledge is expressed in it really functions as data for the decision-maker. It should be used to update the *decision-maker's* beliefs about the *parameter*. This is the framework of elicitation considered by French (1980) and Lindley (1982), but is in principle quite different from what we have described as the elicitation problem. Whatever summaries have been elicited as data from the expert are the decision-maker's raw data, and there is no point in then choosing a single distribution to fit those summaries, or in quantifying the uncertainty in that choice. If elicitation in the sense I have described here is needed for that context, it is elicitation

of the *decision-maker's* beliefs, both about the parameter and about the expert's expertise. Nevertheless, if the decision-maker has negligible prior information about the parameter and trusts the expert's expertise, then it is possible that the analyst's posterior distribution in our approach may indeed represent the decision-maker's posterior beliefs.

Finally, many of the challenge problems feature information from several experts, for instance in the form of several intervals (which may not be overlapping). It is common to consider the question of how to combine the various underlying distributions of the experts into a single distribution for the parameter. But whose distribution does that represent? The most likely scenario for using several experts is that in which their knowledge will inform a single decision-maker, and it is the decision-maker's beliefs that matter. In processing the information given by the various experts, the decision-maker will no doubt synthesise their implied distributions in some way. Our preference is to get the experts to exchange knowledge, with a view to deriving a consensus elicitation from them, rather than different, potentially conflicting elicitations from each expert.

Conflict is particularly important. If Expert A says that a parameter definitely lies in [0.1], while Expert B is certain that it lies in $[2, 3]$, then one (or both) of them is wrong. We should surely confront each with the views of the other to try to resolve the conflict. Or at least we need always to recognise the fallibility of expert knowledge.

7 Propagating parameter uncertainty

We have concentrated in this paper on the elicitation of an expert's probability distribution for an uncertain parameter. OHJWF also pose the challenge of how to propagate parameter uncertainties through the model to derive the uncertainty on the model output. This may be a serious conceptual challenge for some of the so-called representations of uncertainty that are advocated as alternatives to probability, but for probability the formal solution is absolutely natural and unquestioned. If we think of the model as a function f that turns an input vector $\mathbf{x}$ into an output $y = f(\mathbf{x})$ then it is a standard problem in probability theory: given the distribution of $\mathbf{x}$, what is the distribution of $f(\mathbf{x})$? For instance, if the model has a single input x, whose distribution is the standard normal distribution $N(0, 1)$, and if f is the square function, so that $y = x^2$, then it is a routine exercise in probability theory to show that y has a chi-square distribution with one degree of freedom.

There are no conceptual difficulties with propagation of probabilities, but there are practical complications. This mirrors our discussion in Section 3 — there is no doubt conceptually that probability is the unique correct way to represent uncertainty, but there are practical difficulties in doing so

accurately which we need address.

So it is with the propagation of probabilities through a model, when the inputs $\mathbf{x}$ are typically high-dimensional and when the model f is enormously more complicated than the simple square function! One good general purpose tool is the Monte Carlo method. We draw random input configurations $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ from the distribution of $\mathbf{x}$, run each configuration through the model to obtain $y_1 = f(\mathbf{x}_1)$, $y_2 = f(\mathbf{x}_2), \ldots, y_N = f(\mathbf{x}_N)$, then these are a sample from the distribution of y. If N is sufficiently large, typically several thousands, we will be able to estimate any desired feature of this distribution with negligible estimation error. (The estimation error is an instance of code uncertainty, as defined in Section 1.)

Monte Carlo is a perfectly adequate method, no matter how large the dimension of $\mathbf{x}$, and no matter how complex the function f (i.e. the model), provided that we can compute the sample y_1, $y_2, \ldots, y_N$ for a sufficiently large N in reasonable time with reasonable computing resources. We used Monte Carlo, for instance, in our propagation of uncertainty in Section 5.1.

Unfortunately, many computer codes are so complex that a single run may take minutes, hours or even days on very powerful computers. In this context, the thousands of runs needed for the Monte Carlo approach may become prohibitive. Even if the model runs in seconds, the computational demands of the Monte Carlo approach may make effective sensitivity analysis infeasible.

It is therefore worth mentioning the more powerful Bayesian approach developed in Haylock and O'Hagan (1996), Kennedy and O'Hagan (2001) and Oakley and O'Hagan (2002b). These methods make far more efficient use of each model run, and so are able to deliver all kinds of uncertainty analysis, sensitivity analysis, etc., from a single set of typically a few hundreds of model runs.

8 Conclusions

This workshop has been explicitly concerned with epistemic probability. The distinction between aleatory and epistemic uncertainty is useful in pragmatic terms, because it helps to identify which components of uncertainty are potentially reducible by gaining more information. It also helps to illuminate the difference between frequentist statistical theory, which applies only to aleatory uncertainty, and the Bayesian approach, where epistemic uncertainty is embraced and represented by personal probabilities. However, uncertainties which appear to be aleatory often have an epistemic component, and a purely aleatory uncertainty is a rather rare beast.

It is very firmly our opinion that the uniquely suitable representation of uncertainty, whether aleatory or epistemic, is probability. This is based both on axiomatic arguments and practical experience. However, we must

recognise that in practice it is not feasible to quantify uncertainties precisely in probabilistic terms, and it is necessary to acknowledge imprecision in probability judgements. It seems to us that some of the proponents of alternatives to probability frequently miss this distinction. To propose an alternative theory that does not follow axioms of sensible action in the face of uncertainty, simply because probabilities are difficult to measure in practice, is seriously misguided. We would rather have an approximate answer to the right problem than an exact answer to an entirely different one.

Instead of toying with alternative theories, we should be giving serious consideration to how to measure probabilities in practice with the maximum of precision and reliability. This is the elicitation problem.

We have outlined some recent work on elicitation that explicitly allows for the imprecision in experts' judgements. The methodology is applied to one of the challenge problems, although we also argue that the formulation of these problems is unrealistic and apparently based on a denial of the possibility of elicitation. This is work in progress, but seems to be a very promising approach with wide applicability.

Acknowledgements

The research reported here on elicitation was supported by the Engineering and Physical Sciences Research Council. We gratefully acknowledge the important contribution of Delil Gomez Portugal Aguilar in the preparation of this paper, as explicitly mentioned in the text. We also thank the organisers of the Workshop on Epistemic Probability for their support and the stimulus to write this paper.

References

Claxton, K., Neumann, P. J., Araki, S. and Weinstein, M. C. (2001). Bayesian value-of-information analysis. An application to a policy model of Alzheimer's disease. *Int J Technol Assess Health Care* **17**, 38–55.

Cooke, R. M. The anatomy of the squizzel: the role of operational definitions in representing uncertainty. *Reliability Engineering and System Safety* (this issue).

De Groot, M. H. (1970). *Optimal Statistical Decisions.* McGraw-Hill.

Ferson, S. and Hajagos, J. Don't open that envelope: solutions to the Sandia problems using probability boxes. *Reliability Engineering and System Safety* (this issue).

French, S. (1980). Updating of belief in the light of someone else's opinion. *J R Statist Soc* **A 143**, 43–48.

Haylock, R. G. and O'Hagan, A. (1996). On inference for outputs of computationally expensive algorithms with uncertainty on the inputs. In *Bayesian Statistics 5*, J. M. Bernardo *et al* (eds.). Oxford University Press, 629–637.

Kadane, J. B. and Wolfson, L. J. (1995). Experiences in elicitation. *The Statistician* **47**, 1–20.

Kennedy, M. C. and O'Hagan, A. (2001). Bayesian calibration of computer models (with discussion). *J R Statist Soc* **B 63**, 425–464.

Lindley, D. V. (1982). The improvement of probability judgements. *J R Statist Soc* **A 148**, 117–126.

Oakley, J. E. and O'Hagan, A. (2002a). Uncertainty in prior elicitations. *Research Report No. 521/02*, Department of Probability and Statistics, University of Sheffield.

Oakley, J. and O'Hagan, A. (2002b). Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Research Report No. 525/02*, Department of Probability and Statistics, University of Sheffield.

Oberkampf WL, Helton JC, Joslyn CA, Wojtkiewicz SF, Ferson S. Challenge problems: uncertainty in system response given uncertain parameters. *Reliability Engineering and System Safety* (this issue).

O'Hagan, A. (1988) *Probability: Methods and Measurement*. Chapman and Hall, London.

O'Hagan, A. (1998). Eliciting expert beliefs in substantial practical applications. *The Statistician* **47**, 21–35.

Saltelli, A., Chan, K. and Scott, M. (2001) (eds.). *Mathematical and Statistical Methods for Sensitivity Analysis*. Wiley.

Savage, L. J. (1956). *The Foundations of Statistics*. Wiley.