Introducing structured expert judgement

Defining Structured Expert Judgement

An *expert* is generally considered to be someone who has built up knowledge of a particular domain, giving them insight that is not shared by the general population. They may be recognized for their expertise through academic qualifications, through years of practical experience, or by peer recognition. Expertise in a particular domain does not necessarily imply that the expert is also good at making assessments of uncertainty in that area, that he/she is expert in related areas, or that he/she agrees with others who are also recognized as experts.

The word *judgement* is used to indicate that an expression of opinion is being given by the expert relating to statements about the real world. Such an expression of opinion does not need to be the “gut instinct” of the expert, but would ideally be a well-considered view that takes account of the extent of scientific or other knowledge about previous experience, and may rest on the use of quantitative modelling tools. The form of judgement will depend on the questions that the expert is to respond to, but may range from judgements about which modelling approaches are most appropriate, to quantitative assessments of uncertainty ranges.

A *structured* approach to expert judgement is one that seeks to minimize any biases and sources of ambiguity in the process of collecting expert data, and which ensures that the process is as transparent as possible. In particular the role of expert is normally separated from the roles of decision maker (by which we mean the person or group who will use the expert data to take actions) and from the role of analyst/facilitator (by which we mean the person or group responsible for ensuring that the process is well run.

Desiderata for a structured process include:

* Identification: A well-defined procedure for identifying potential experts and ensuring that diverse expert views are represented within the group.
* Role separation: Experts, decision-makers and analyst/facilitator roles are separated.
* Neutrality: Treating all experts equally, a-priori.
* Eliminating ambiguity: Ensuring that divergences of expert views are not caused by ambiguity in the way questions are posed.
* Maintaining plausible diversity: Ensuring that divergences in expert views are not eliminated if these rest on different plausible scientific perspectives, including a recognition of incomplete scientific understanding.
* Training: Providing appropriate training so that experts understand what they are being asked to assess – in particular if they are being asked to assess probabilities or uncertainty ranges, that they understand what these are.
* Neutrality: Encouraging experts to state their true opinions, for example by ensuring that individual expert behaviours do not affect the views of others, or by ensuring there are no hidden incentives/costs to experts to state something other than their true opinion.
* Empirical control: Where possible providing empirical control of expert assessments
* Transparency: Ensuring that rationales are given for expert assessments, that any calculations made are reproducible, and that any assumptions made are recorded.
* Accountability: The sources of expert opinion should be identified to the stakeholders.

Problem types for which SEJ may be appropriate

Expert views are used at a number of different stages of understanding and assessing problems. This can include

* Structuring problems: Assessing what factors are relevant to a particular problem and – where necessary – helping to provide a better definition of the problem.
* Identifying strategies or options: Helping to define options that the decision maker could choose from, that might be expected to have an impact on the problem.
* Screening: Selecting from a larger group of problems those ones which are likely to be the most significant, according to some scale which cannot be measured precisely enough except at high cost in time, money or other constrained resources.
* Making assessments of uncertainty: Quantifying or categorising the degree of uncertainty associated to a particular outcome, prior to that outcome being known to the decision makers.

The work of COST Action IS1304 focusses mainly on the final problem, though not exclusively. Expert Judgement is often used when there is insufficient directly relevant data available for the problem at hand, but there may be partially relevant data for similar problems which is available. The expertise required is then the ability to extrapolate to that new situation and assessing a range of credible outcomes that should (with high probability) capture what will actually be experienced in the new situation.

Uncertainty and probability

The word “uncertain” is used in everyday language in a number of different ways. I could be uncertain about which choice to make when confronted with a difficult decision problem, or during a conversation I might be uncertain about what information the other person is trying to convey. The first problem is about weighing up between choices, and the second one is about dealing with ambiguity. Neither of these is what is meant here when we talk about uncertainty: in the first case the uncertainty is resolved by making a decision, while in the second case the uncertainty is resolved by discussion around the meaning of words to reduce the ambiguity.

In general, we are interested in uncertainty about some aspect of the state of the world, either something that is to happen in the future or something that is now or in the past but has not been observed. A statement making an assertion about the unknown state of the world needs to be sufficiently well detailed that is either going to be true or false. For example, “The average temperature in London in 2025 is above 10°C.” is fairly clear, but could be further clarified by saying that air temperature will be measured every day at the terminal rooftop measurement station of London City Airport at every whole hour on every day, and the arithmetic mean of those measurements is the “average temperature in London”. Probably for most purposes this would be enough to resolve to a reasonable level any residual ambiguity in the meaning of the statement. It is now a statement that will either be true or false, and we only have to wait until the end of 2025 to find out. In the meantime, we are uncertain about whether this will happen or not and this uncertainty can be expressed using probability.

The theoretical foundations of probability are often expressed in purely mathematical terms through the axioms of Kolmogorov, but there is a rich tradition (with names like Ramsey, Von Neumann and Morgenstern, de Finetti, Savage..) which provides both a theoretical framework consistent with Kolmogorov and an interpretation of probability in terms of “rational preferences” (these are preferences expressed by an individual – for example - when asked to compare possible outcomes for the unknown state of the world to the outcomes of a “reference lottery”, like coin toss outcomes, which has known properties). This is really important as it enables us to assess probability for one-off events. For example, for geometric reasons we accept that a coin which is physically symmetric should have equal probabilities of heads and tails when it is tossed. By asking myself whether I think it is more or less likely that the coin toss gives an outcome of *H*  or that a particular event happens (eg the average midday temperature in London in 2025 being larger than 10°C) I can assess coherently whether or not the probability of that event is greater or less than 0.5 (in my opinion).

In practice, most elicitation protocols do not use the idea of direct comparison with reference lotteries in order to actually elicit probabilities, but reference lotteries are good examples to illustrate clearly that it makes sense to talk about the probability of one-off events.

The upshot of this is that when we are interested in the uncertainty of a well-defined event then this can expressed in terms of probability. However, there is no reason why an expert should not change their mind about the probability if they get more information about the event in question, and there is no reason why two different experts should make the same assessment about the probability of a well-defined event. Since many of our decision problems are ones in which the decision maker is not an expert, we need to find good processes by which decision makers can put their trust in assessments made by experts. This trust issue is central to many of the challenges of structured expert judgement.

Probability and information

The famous probabilist Bruno de Finetti said that “Probability does not exist”. There is no physical measurement that could be made to demonstrate that the probability of a one-off event is (say) 0.5. Probability can change depending on information available: a very crude example is when a coin toss is made in front of a group of people and the result revealed to half of them. For that subgroup the probability of heads has gone from 0.5 to either 0 or 1 depending on the outcome, but for the other subgroup the probability of heads is still 0.5.

This points to a potentially important reason for differences in expert assessments: some may have access to more information than others and therefore an expert elicitation process which allows the sharing of information prior to assessments could reduce potential impact of this issue as a source of reasons for differences between experts. Note however that experts may differ in the degree to which they think that a particular piece of information is relevant to the situation they are considering, and when they take it into account there could be more divergence between their views rather than less.

The key point – one that scientists and policy makers often struggle with – is that probability is not something “out there” in the real world that can be measured without using judgement and/or interpretation. (Even with the “gold standard” of randomised control trials in medicine, judgements are made about how to group different parts of the population under study, which are to some extent arbitrary and which would change “the probability” associated to a specific individual).

When using experts we are not then seeking to identify “the objectively correct probability” – this cannot exist - but are seeking a probability which can plausibly be adopted by decision makers as appropriate to the context in hand. The context can be very important – private sector decision makers accountable to a commercial board may operate in a very different context to public sector decision makers subject to stronger public scrutiny.

Group assessments of probability

Even when experts have a common understanding of the event whose probability is to be assessed, and even when they have exchanged insights about information (potentially) relevant to the event, they may still have a different view about the probability.

For a decision maker, different assessments of probability are generally seen to be problematic and unhelpful. First is it is not clear whether there are good or bad reasons for the lack of consensus: is there a genuine lack of scientific understanding, do the experts really have good knowledge of this area, are they good at expressing their scientific understanding into probability assessments? Second, a range of uncertainty assessments may not help the decision maker with the specific decision to be made when this a one-off (eg do we evacuate?)

Because of this, most approaches to expert judgement assessment include a process or method to bring together different expert assessments into a single “group” probability assessment.

There are two broad categories of methods for aggregation of expert probability assessments: mathematical pooling methods and consensus-generating methods. There is a strong methodological difference between these approaches and academics often hold strong views about one or the other approach. The consensus-generating approach relies on the idea that with enough discussion and a shared understanding, a group of experts will eventually come to agreement about what the uncertainty in a particular outcome is. This view which is criticized by those who point out that a shared understanding does not mean a shared view about outcomes and who argue that behaviours within the group could lead to dominant or articulate individuals influencing group assessments, or even that the need to find consensus dominates over the need to address the actual uncertainty: it is little use that the experts agree if they all turn out to be wrong. The mathematical pooling methods would normally ensure that there is common understanding of the questions and usually uses weights to combine the individual expert assessments of uncertainty. Equal weights are often used, as are “performance based weights” that reflect the (relative) ability of experts to assess uncertainty. The pooling methods are criticized on the grounds that the meaning of the pooled assessments is not clear, that there may be residual differences of understanding between experts, and that it can be difficult to find appropriate calibration questions on which the experts’ ability is assessed.

There are strong analogies between “model averaging” as performed within the AI community and the mathematical pooling approaches. Consensus within the AI community is that the performance of pooled models improves that of individual models, and this is also to be seen in most examples of pooled experts – with performance based weighting generally providing a superior improvement beyond that shown by equal weights. However, the analogy is by no means perfect as the types of problems assessed by AI models are quite different and for these there is a larger evidence base (“training set”) used to calibrate models.

Expert Judgement Processes

Despite debate of the relative merits of approaches to combine expressed expert views, there is broad agreement that the actual elicitation should be preceded by a number of stages in order to ensure that the data collected from the experts is as “clean” as it possibly can be. These stages include

* Work with problem owners and modellers to ensure that the topics for expert assessment cover the requirements of the study. This may involve carrying out a scoping exercise to decide what will be included and what not.
* Develop questions and review to ensure that all quantities, conditions and assumptions are unambiguous. Test that the questions are not subject to framing biases by restating in different ways if appropriate.
* Carry out a dry run process in which a small number of experts can test and give feedback on the process and the questions. In particular it is important to check the questions for potential ambiguities
* Revise the process and the questions as necessary.
* Train the experts to ensure that they have a good understanding of probability assessment, and check for ambiguity problems.

Preference judgements

The discussion so far has been about uncertainties and probability quantification. Decision problems normally also involve making preference judgements about outcomes. These can include both judgements about preference trade-offs between potentially competing outcomes (for example, cost and performance) and about attitude to risk. These kinds of judgements are generally considered to be the domain of the decision maker rather than of any set of experts.

However, there are cases when expert judgement techniques could have application to preferences. Some examples are:

* Societal decision making is often carried out within a broad cost-benefit framework, but the societal context often requires the inclusion of “intangibles” or other aspects which are not traded and do not have a “market price”. In this context there may be a role for expert assessment in judging how or what the valuations should be.
* Decision makers may, for a variety of reasons, want to know what other stakeholders preferences are (even if they recognize that they cannot keep all the stakeholders happy). Hence there is a potential role for expert judgement in assessing stakeholder preferences.
* In an adversarial setting, the decision makers need to understand what the preferences of the adversaries could be. Since the trade-offs and objectives of the adversaries are usually unknown, there is a role for expert assessment of their preferences.

Legitimacy and expert performance

A key challenge for expert judgement methods is the justification of the outputs to users and stakeholders. Why should they adopt the outputs as legitimate representations of uncertainty? The answer is complex. If users and stakeholders agree a-priori that the experts are indeed experts, and are happy that the process is effective in managing potential biases etc, then they could be expected to adopt the outcomes of the process. Acceptance of the expertise and the process a-priori leads to acceptance of the outcomes a-posteriori.

However, this is not always the case, and in areas of policy controversy there can be scepticism about either expertise or independence of the experts. In that kind of situation measures of expert performance can be used to provide an independent demonstration of the way the expert has performed. The main method that seeks to measure expert performance is the Classical model of Cooke (though others exist). One could argue that if stakeholders and/or users are not willing to fully accept expertise a-priori, but are only willing to accept it contingent on demonstrated performance, then the Classical method provides a route to justify legitimacy of the overall expert outputs even in this case.

The key societal challenge of using Structured Expert Judgement is a satisfactory response to the claim that

Key literature

There are several surveys of expert judgement methods in the literature. This introduction was therefore designed to survey the wider issues of Structured Expert Judgement in a way that complements the more detailed academic literature surveys. The idea that a process should be created, in which expert opinions can be systematically, usefully, and accurately gathered, became best known after second world war with the development of the Delphi method. However, since then it has been widely used

A wide survey of work in SEJ up to 1991 is given by Cooke (1991), with further methods described by Bedford and Cooke (2001). The Classical Model of Cooke described there and implemented in EXCALBUR software, is one of the main methods used for SEJ, and a data set of quantitative expert judgements from more than 70 applications is publicly available. O’Hagan, Buck et al (2006) provided a different perspective, and have gone on to develop the SHELF method which uses a behavioural aggregation rather than the expert testing approach used in the Classical model. A discussion of some of the pro’s and con’s is given by Gosling (2014), though he does not address the legitimacy questions raised above. A much wider review of methods for combining different expert opinion is given in Daneshkah et al (2011). A broad overview of processes and elicitation methods is given by Hora (2007). Specific guidance procedures for Cooke’s Classical Model are given by Cooke and Goossens (1999). Specific guidance for the SHELF approach can be found at [www.tonyohagan.co.uk/shelf/](http://www.tonyohagan.co.uk/shelf/).

References

Bedford T. and Cooke R.M., *Probabilistic Risk Analysis: Foundations and Methods*, Cambridge University Press, 2001.

Cooke R.M., *Experts in Uncertainty*, Oxford University Press, 2001

Cooke R.M., and Goossens, L.J.H. Procedures Guide for Structured Expert Judgement, EUR18820 , European Commission, 1999. Available at ftp://ftp.cordis.europa.eu/pub/fp5-euratom/docs/eur18820\_en.pdf

Daneshkhah, A., Revie, M., Bedford, T., Walls, L. and Quigley, J. 2011. “Eliciting Subjective Probability Distributions from Groups.” *Wiley Encyclopedia of Operations Research and Management Science.*

Gosling J.P., Methods for eliciting expert opinion to inform health technology assessment, Vignette on SEJ methods for MRC, UK, 2014., accessed at [www.mrc.ac.uk/documents/pdf/methods-for-eliciting-expert-opinion-gosling-2014/](http://www.mrc.ac.uk/documents/pdf/methods-for-eliciting-expert-opinion-gosling-2014/)

Hora, S.C., Eliciting probabilities from experts, in *Advances in Decision Analysis: From Foundations to Applications Edwards, W., Miles, R.F., Jr., von Winterfeldt, D. (Eds),* Cambridge University Press, 2007.

O'Hagan A., Buck C. E., et al *Uncertain Judgements: Eliciting Experts' Probabilities*, Wiley, 2006.