Engineering judgement in reliability and safety and its limits: what can we learn from research in psychology

Lorenzo Strigini
Centre for Software Reliability, City University
Northampton Square, London EC1V OHB, UK
Tel +44 171 477 8245 Fax +44 171 477 8585
E-mail: strigini@csr.city.ac.uk

Abstract

Engineering judgement has an important role in safety or reliability assessment. This paper focuses on the use of engineering judgement for integrating diverse evidence into an assessment of the safety or reliability of a product. In many cases of stringent safety requirements, this form of engineering (or "expert") judgement, i.e., "informal inference from complex evidence", is the crucial resource for the decision maker, for lack of more solid, objective evidence. This dependence on judgement is especially evident in the assessment of the unreliability due to possible design faults in complex products, and computer software in particular. Although engineering judgement plays an essential role in the assessment, there are good reasons to doubt the ability of experts in some of the judgement tasks in which they are usually employed. Experimental research both about the way humans think and integrate evidence, and about the performance of experts in tasks similar to engineering judgement, support the idea that the ability of experts may be overrated. This paper summarises some literature about common fallacies and ways to guard against them, and argues for a more disciplined use of expert judgement.

Keywords: expert judgement, judgement under uncertainty, cognitive bias, probabilistic reasoning, heuristics, dependability assessment, safety case, design faults.

1. Introduction

The problem of ensuring very high levels of safety or reliability in all kinds of engineered products is drawing increasing attention, driven by an increased awareness of safety issues, increasingly stringent requirements, and technological advances (in particular, but not only, the growing role of computer software in all kinds of products and engineering processes) which require an evaluator to estimate the probability of design-caused failures. In all safety evaluation tasks, a role is recognised for "engineering judgement", or "expert judgement". In cases like the evaluation of safety-critical software, where the required levels of reliability may be so high that they cannot be practically demonstrated by applying standard reliability evaluation methods, engineering judgement tends to be treated as either the ultimate or the sole basis for evaluation [Littlewood and Strigini 1993; Strigini 1994]. (A similar *de facto* attitude is common in tasks like the choice of software engineering methods for the industry in general [Fenton, 1994]. Differing positions about the role of expert judgement exist in the risk assessment community. A taxonomy of the problems and decisions necessary in using expert judgement is found in [Chhibber et al. 1992].

This survey considers one of the questions raised by the widespread reliance on engineering judgement. When the role of an "expert engineer" or "expert" (as I will call any expert of an engineering field relevant to the product being assessed - e.g., aircraft structures, control software or chemical reactors) is *to integrate disparate evidence* from the project at hand with information from the expert's own experience, how good are the

experts and how can they best be used? More narrowly yet, if we exclude errors caused by fatigue, dishonesty, and other well recognised causes, which causes remain that are inherent in the way the human mind works, and how can an assessor, trying to use the expert engineer's judgement in decision-making, guard against them?

This is in no way a report of original research work in psychology. Rather, it is a summary of a few compilations (at the level of advanced textbooks) of current knowledge, as read by a researcher in computer dependability, complemented with discussion and references about its significance for dependability assessment. Its aim is to point out, with reference to the issue of dependability judgements in general, but in particular for software and other products subject to design faults: i) the relevance of some problems which are known and already receive some limited consideration in some other specific fields (like nuclear risk assessment and management); and ii) possible defences against these problems.

The expert engineer's task in these cases is often one of complex statistical inference performed in an intuitive or semi-conscious way. Examples of statements by people called upon to evaluate safety-critical software are "Our confidence in the software stems from the observed excellence of its development process" or "All the observed imperfections in the software process could individually be seen as acceptable exceptions, but together they build the impression that one cannot trust software built this way". The assessor or the expert engineer (often the same person) is confronted with a wealth of evidence about the details of the design, the design methods used, the quality assurance organisation, and the results of testing, none of which by itself comes even close to proving the desired conclusion, e.g., that a system has a certain minuscule probability of failure per hour. The net of cause-and-effect chains, deductions and inferences which binds the evidence with the conclusion to be reached is presumably rational, but possibly so complex as to defy analytical description. However, it is widely felt that people are good at deriving decisions from such complex mazes of evidence and reasoning. Experts in a discipline, in particular, are skilled in integrating the evidence, through a partially unconscious algorithm learnt from experience, to obtain appropriate solutions. Or are they?

There are at least two rational methods for answering this question. One is to measure the past performance of the experts, and draw inferences about how trustworthy a new prediction is. This approach has a number of practical problems when dealing with rare events and small numbers of predictions. There is also some widely quoted evidence of dramatic failures by multiple experts in individual controlled experiments.

The second way of reasoning about the experts' abilities is to try and understand *how* they obtain their predictions, and whether their method is reliable, or fit for its purpose. Researchers both in the field of "expert systems" and in branches of psychology have found that people tend to be unaware of the algorithms they use, even in tasks where they are remarkably successful. So, researchers need to conjecture these mechanisms and algorithms by observing behaviour and building models of it. They have developed models of the mechanisms by which the human mind processes information for the various tasks to which it is applied. This research uses controlled experiments and all the normal scientific precautions and is, in this sense, trustworthy. Much of this research points to people not being "correct" processors of statistical evidence whenever they do not consciously apply the rules of probability calculus (also called the "normative" procedure). This is not a conclusive research result: both new experiments and re-interpretations of previous ones have been shown to counter some of the more pessimistic conclusions, and the opinions of scholars about the general proficiency of humans in intuitive statistical processing vary between optimism and pessimism. Another consideration is, of course, that it is difficult to accept theories derived from small-sample experiments as appropriate for humans at large, and for specific experts in particular. On the other hand, the theories supported by these experiments may be the best now available about the functioning of experts' minds as well. At this stage of scientific knowledge, it seems quite safe to conclude that *a priori* trust in the human ability to perform tasks of intuitive statistical inference is unjustified: decisions should not be based on the

judgement produced by a human in such a task unless proper scrutiny of the specifics (of the expert and of the problem) supports trust in that individual judgement. In some cases where engineering judgement has a crucial role, strong enough evidence for trust seems normally to be missing.

In the following sections, I will first examine the concept of expertise, and in which sense we rely on it when relying on "engineering judgement"; then, in Section 3, I will list some of the known biases and fallacies in human information processing and the ways they may affect the task of engineering judgement for dependability assessments. In Section 4, I will consider the modelling and assessment of expert judgement; in Section 5, safeguards and precautions will be discussed.

2. Requirements on the expert

2.1. Roles of the expert engineer

The roles of an expert engineer in assessment (as opposed to roles in design) may be various. The expert may be called upon to state, e.g., one of the following opinions (or its contrary):

- that the design methods used are appropriate for the task, based on the current state of the art;
- that no unacceptable failure mechanism (design fault) is left in the design;
- that a safety analysis does not contain dangerous commission errors (e.g., a gate of the wrong type in a fault tree);
- that a safety analysis does not contain dangerous omissions (e.g., an input to an OR gate in a fault tree);
- that a certain value (or range of values) for an input parameter of a safety analysis, e.g., the probability of a certain human error, is realistic;
- that, based on the disparate evidence available about a system, a certain global judgement is appropriate, e.g., that a system is acceptable for use; or that the probability of failures of a certain type is less than an established threshold (this is the case of most interest for this paper).

Interestingly, all these cases contain an element of probabilistic reasoning. The first three tasks can in part be discharged by applying quasi-algorithmic tasks of proof, list-checking and such, but of course a probabilistic element is present in answering questions like "How confident am I that I did not omit any element in the checklist?". This probabilistic element is more explicit in the last three items in the list. When pruning a fault tree to discard events that would make it unmanageably complex, one has to choose events whose total probability is low enough not to cause excessive error: experts may do this by extrapolating from their own (more or less direct) knowledge of past events in more or less similar circumstances. This operation is a sub-case of the next one, estimating the value of a probabilistic parameter. Last, the combination of disparate evidence is the archetypal judgement operation. There is a set of inputs that relate to the issue at hand (e.g., whether the operation of a plant should be authorised). These input data are evidence for or against the issuance of an authorisation. The combination of evidence is typically a combination of causal (or deterministic) and probabilistic reasoning, where the latter seems to pose the more serious problems, while both contribute to the complexity of the task. In the rest of this paper, I will concentrate on the problems of intuitive statistical inference: determining how much the knowledge of the past successes of a design organisation should contribute to the confidence in its last design, which branches in a fault tree are truly negligible in terms of contribution to a failure probability, etc. Diverse tasks are involved: determining a correlation between two factors, estimating the probability of an event, etc. The existing literature does cover such diverse tasks. In

Section 3, I will cursorily summarise interesting results, without going into much detail regarding each task.

2.2. The nature of expertise

The advantages of asking the opinions of experts rather than of "lay" people seem to be twofold: i) experts know more facts about the problem of interest; and ii) experts are accustomed to reasoning about these problems, and hence they will use the facts better. Regarding this second presumed advantage, however, the current understanding of mental processes seems to require some caution [Reason 1990].

Compared to non-experts, experts seem to exhibit improved performance in skill-based and rule-based tasks (the "lower" levels of activity, where less conscious intellectual activity is required; these are also the levels where errors are less frequent and more easily detected by the author of the error) [Reason 1990]. In knowledge-based tasks (i.e., those requiring "higher-level" intellectual activity of consciously analysing the problem) experts seem to make the same kinds of mistakes as non-experts. Experts enjoy two main advantages, in comparison with non-experts: i) they possess large collections of problem-solving rules, stored in their minds, which are appropriate for the class of problems where they are expert; and ii) they are able to see an individual problem in more abstract terms, obtaining a mental model more appropriate than that of a lay person for finding keys to problem-solving rules. It can therefore be expected that an expert will apply knowledge-based activity less often than a lay person, and the expert's errors, if any, will consist in applying a well-learned rule to the wrong situation. The question arises here of which tasks, among those delegated to "engineering judgement", are such that experts may be expected to be highly reliable in performing them.

As we have seen, experts may be used to apply well-defined (to a large extent procedural) skills to data (e.g., predict the behaviour of an electrical circuit from a circuit diagram, or build a probabilistic model of a certain kind of accident), or to apply generic "judgement" to insufficient data. In the former case, we worry mostly about possible mistakes in applying a correct procedure: we should be concerned with an organisation of the job that does not require excessive mental work before an intermediate result is recorded, abundant chances to double-check the procedure used, a physical presentation of the data which is not stress- or error-inducing, etc.

For the task of finding errors in a design or analysis we can probably expect experts to be fairly reliable: they probably proceed (more or less consciously) by looking for cues pointing to the existence of a flaw, and humans are good at using cues. Of course, flaws that the expert has not encountered before may go unnoticed. This problem may be overcome by adding to the error-seeking a more systematic analysis of the design or argument: the simple systematic application of a known process (like mathematical deduction) may suffice. However, this is not usually sufficient for checking a complex design or argument: complexity itself gets in the way, and, furthermore, errors of omission may make the procedure ineffective by undermining the very axioms of the reasoning process. Errors of omission are difficult to find (a fault-tree example is in [Slovic et al. 1982]), and yet many arguments must depend on exhaustive enumeration, e.g., of failure cases in a fault tree, or simply of the set of checks to be run on a design. An expert's judgement that the fault tree is complete probably depends on the expert's previous experience about the appearance of complete (or incomplete) fault trees for similar problems, the typical omissions and the lines of thought which led to finding them: unless the expert notices "patterns of omission" that he/she has learned to recognise, some kind of statistical inference is again required.

For most of the tasks just discussed, there is thus a basis for trusting expert engineers to be effective at them, thanks to peculiarly human information-processing skills, although not completely reliable in terms of doing a complete job. Tasks like conjecturing the probabilities of rare events, or drawing inferences about and from the correlations among factors in our experience, are definitely not in this category. Many tasks in which experts

excel seem to rely on a powerful pattern-matching mechanism which can pick up the right cues in the presentation of the problem, to trigger appropriate rules of reasoning or behaviour. To trust a human expert in a task of intuitive statistical inference, we should instead postulate an ability to automatically count and classify events (when an expert is asked to estimate probabilities or correlations of events that he/she has observed in the past), and to apply (or emulate) probabilistic decision algorithms (when the expert is asked to combine evidence into a synthetic assessment).

As we shall see, there is very good evidence that our mental mechanisms for dealing with probabilities (or the intuitive concepts that we consciously represent as probabilities) are liable to serious errors, even in comparatively frequent situations: the mechanisms themselves do not follow the formal calculus of probability; they are apparently a reasonable approximation of it as far as producing everyday decisions which are either correct or not catastrophically wrong, but not necessarily for producing the kind of judgement that we are considering here. Natural selection, one may consider, must have been most effective in eliminating strongly disadvantageous behaviour, but not at fine-tuning the capability for optimal decisions (which can give only marginal advantage in a world where most decisions have to be taken on the basis of very uncertain data anyway).

2.3. Expertise in engineering judgement

Expertise seems to consist in having developed the ability for intuitive solutions of problems, i.e., an acquaintance with the patterns of evidence in a class of problem. In an expert, observing a new problem in the class prompts a reliable process of pattern recognition leading to the recall of the right solution or solution rule. The question arises of how one can become an expert in intuitive engineering judgement[1] about dependability issues. There seem to be some necessary conditions for this kind of expertise to be obtained:

- a previous exposure to a sufficiently large sample of problems from the class of interest: the number of problems observed must be large enough both to activate the human mechanisms for "learning from experience", and for providing these mechanisms with a presumably fair sample of the real population, and this latter condition seems to be the more difficult to meet of the two;
- an exposure to their correct solutions (or at least an indication of the errors made by the apprentice expert). There are different ways for apprentices to learn whether their solutions are correct. The intuition of how an engineering design should be structured to produce the desired result is checked by analysing or testing the completed design, a rather reliable method. An intuition of which design is *best* for its purpose is more difficult to check, as in many cases it would require an ability to generate and check [a classification of] the relevant population of alternate designs. A politician or a chess player may learn both through the positive and negative reinforcement of victories and defeats (which may, however, be deceptive), and through an analytical comparison between the intended and the obtained results in light of the existing external influences. In the case of dependability predictions, the feedback must come from the observed outcome exhibiting statistical properties matching the prediction; but fully informative feedback may be difficult to obtain, e.g., in the case of assessing the probability of a single, non-repeatable event;
- the availability of information about these problems that is sufficiently relevant to a solution. If the evidence available is always very weak, it seems that a good

[1] As distinct from experts in statistical inference. These, of course, are a resource for drawing inference in those cases where the evidence and the inference process are made explicit. The expert engineers themselves could become experts in statistical inference, but this is not a common situation.

apprentice expert will only learn to correctly derive weak predictions, which will not be of much use.

So, to trust an instance of engineering judgement of the form (familiar in the debate about safety-critical software) "this product is likely to have a probability of failing of less than 10^{-9} per hour" *on the basis of the utterer's expertise* would seem to require that the latter has observed a number of similar products which convincingly exhibited that probability of failure. So, we will not find expertise a very solid basis for trusting such a statement. What is worse, even trusting less extreme opinions may be difficult, if these are the results of intuitive combination of evidence. Imagine an expert predicting an MTTF of one year; he/she may have observed many systems with an MTTF of one year, and is assimilating the product under examination to one of those, instinctively using cues in the product's structure and environment: how do we know that these cues are appropriate (that the expert is using a good sample for his/her prediction)? However, the one-year MTTF can be demonstrated via rigorous procedures, which make the intuitive judgement process less necessary. It is when we ask for statements which are difficult to support by empirical evidence and rigorous reasoning that we give intuitive engineering judgement a crucial role. When dealing with judgements of very high dependability, it is more reasonable to trust, on the basis of expertise, someone who claims that a system will *not* be as reliable as required, as there is a better chance that the person has indeed observed project failures of this kind. Yet another problem is whether the expert has formed his/her pattern-recognition habits on the basis of valid evidence. It may well be that the evidence practically available has very little predictive value. There are at least two tests for recognising this danger: i) has the expert usually been right, where "usually" must be interpreted in terms of statistical significance (and difficult questions still remain, like "Is an expert on safety of electro-mechanical equipment still trustworthy when judging software-based equipment?"), and ii) even if that is not the case, is he/she able to explain his/her use of the evidence, so that the correctness of the inference processes may be checked independently?

3. Weaknesses in human judgement

I now list some of the mechanisms that seem to determine erroneous judgement. This chapter is mostly based on [Kahnemann et al. 1982]. The reader should be aware that my selection of sources thus favours the "pessimistic" view of human abilities. In the presence of very scattered evidence which does not cover the sets of tasks and experts that one has to deal with, one should, in my opinion, consider these experimental results as useful in two ways (similar arguments are in [Ayton 1993]). First, they are counterexamples refuting the conjecture that human intuitive inference can generally be trusted: hence, to decide how much trust to give it in an individual case, one must consider the details of that specific case. The second use of these results is as pointers to observed fallacies and their possible causes and remedies: even if the pessimistic results of a specific experiment did not apply to human performance in general (because they are the result of unrepresentative peculiarities of that experimental set-up), the decision maker should be aware of the risk of stating an individual task for the expert in a way that reproduces those peculiarities and is likely to cause the same fallacies.

3.1. Heuristics and common biases

Some heuristics that seem to predominate in people's application of "intuitive" inference are:

- ***Representativeness***. The perceived probability that a given object belongs to a certain class is highly affected by how well the object seems to "represent" the class. This heuristic comes into play whenever descriptive evidence is given about the object: it can apparently be "triggered" (made to prevail over other heuristics) simply by giving the experimental subject irrelevant but abundant evidence about

the object. This heuristic is of course insensitive to the *a priori* probability of the event of interest (proportion of the general population that belongs in the class of interest: this leads in particular to the common fallacy of non-regressive predictions, which will be discussed in more detail later), to the size of samples (if the task is to predict the outcome of some sampling process on the general population), and to the predictive value of the information provided about the object (i.e., both the probability of the information being true and its correlation with the factor to be predicted). It prompts people to put more trust in sets of information which appear to be more consistent, including cases where the consistent set is simply made of variables known *a priori* to be correlated. It causes the layman to believe that a heads-and-tails sequence (from flipping a coin) like HTHTTH is more likely (because it "looks more random") than HHHTTT.

- ***Availability***. The frequency or probability of an event is judged by the ease of imagining instances of it. This approximates "true" probability in many real-life cases, but often does not. Causes of bias may be the differences in ease of retrieving different instances from memory: salience (the probability of a car crash appears higher right after we have seen one), familiarity (after being told a list of names of celebrities, we will tend to base our estimate of how many were males only on the better-known among them, whom we can recall more easily); and the relative ease of different search modes (we tend to believe that a given consonant is more likely to appear in the first than in the third position in a random English word, for *any* consonant, including those for which the reverse is true, because it's easier to search for words by their initial than by their third letter). Other biases may come from the ease of imagining representative cases: naive subjects estimate that there are more combinations of 2 items out of 10 than of 8 items out of the same 10, and, in general, scenarios which are difficult to construct may be neglected (and scenarios that are easy to imagine can be overestimated) in predicting probabilities. Yet another effect tends to confirm the subjects' own pre-conceived theories about correlations of factors, as the cases in which the supposedly correlated factors did coexist are easier to recall than the others.
- ***Adjustment*** and ***anchoring***. In producing estimates of numerical values, people often produce first an initial estimate, based on some piece of the evidence available, and then adjust it using the remaining evidence. However, this adjustment process seems to be over-conservative: people are unwilling to change the initial estimate by much. So, procedures with different starting points yield different estimates, each biased towards its initial estimate. Among the effects of this heuristics are the fact that people tend to overestimate the probability of events of the form "A **and** B", and underestimate that of "A **or** B", when they start from the probability of A and then correct to take into account that of B. Another interesting effect is observed in the elicitation of subjective probability distributions. Asking a subject to state the values of given percentiles of a distribution usually produces a narrower distribution (as subjects operate by corrections from their perceived median or mean values) than asking for the probabilities that given values of the random variable are exceeded in an experiment, although the two sets of questions are equivalent in theory.

The above observations apply mostly to "lay" people, the subjects of most controlled studies[2]. Experiments on experts are obviously more difficult and expensive. However, this body of research provides conjectures on the functioning of experts' minds as well, insofar as it indicates shortcuts which the human mind uses to "approximate" those tasks that would be too taxing if performed rigorously. Furthermore, there is disquieting evidence of fallacies in the reasoning of real-world experts:

---

[2] There is a common joke to the effect that scientists have by now reached a thorough understanding of the operation of the minds of Psychology undergraduate students.

- in terms of *results* (predictions). A wide body of research (summarised for instance in [Goldberg 1986]) has shown that physicians' clinical judgement produced results which were in many studies inconsistent between physicians, often invalid, not improving with the physicians' experience nor with the amount of information provided to them. Moving from medicine and psychology to the more rigorously based disciplines of engineering or physics, one would expect good judgement to be easier. Yet, scattered reports and anecdotes include: a preposterous estimate of the safety of the Therac-25 cancer-treatment machine, which killed a few patients due to unsound design [Leveson and Turner 1992]; in the history of modern physics, the values attributed to physical constants have oscillated, with corrections repeatedly exceeding the confidence bounds previously believed to apply to the "best current" estimates [Henrion and Fischhoff 1986]; and in the experiment [Hynes and Vanmarcke 1976] in which a number of expert engineers were asked to predict how high an embankment could be built before it collapsed, the predictions had a bimodal distribution, with the actual collapse occurring at a height somewhere in between, and outside the 95% confidence limits estimated by the two groups; in an experiment on software engineers, the subjects consistently believed they were more effective at finding software bugs by testing the software than by inspecting it, while the experimental log showed the opposite to be true [Basili and Green 1994];
- in terms of the *methods used* to produce predictions. There is evidence of a "belief in the law of small numbers" [Tversky and Kahneman 1982]: behaviour which *would be* rational *if* small samples could be trusted to represent faithfully the statistical characteristics of the whole population. This tendency is kept in check in all cases where standard statistical tests are applied, but not in others. For instance, researchers (who had published in psychology journals) were observed to decide the sample size for an experiment without appropriate consideration of the likelihood that the experiment would produce insignificant results: a "believer in the law of small numbers" would "gamble his research hypotheses on small samples, without realising that the odds against him are unreasonably high" [Tversky and Kahneman 1982]. When an independent experiment supported the results of a previous one, but with lower confidence, it was seen by many as a failure to replicate the result, rather than as confirmation (as would be the case if the data from the two experiments were pooled together). Likewise, there was an excessive tendency, when confronted with two contrasting experimental results, to look for causes of the difference, even when they were quite likely to be due solely to sampling variations. Further, investigations among clinicians have shown trust in discredited tests, based on a "normatively wrong" interpretation of their personal experience. In other studies, clinical decisions following test results or about administering tests were shown to violate any rational decision theory.

A general conclusion is that numerous statistical fallacies come naturally to people, including experts, when they are not consciously applying the rules of statistics. This may even be true when experts are reasoning informally about data which are themselves the result of controlled experiments or of formal statistical analyses!

In more detail, some observed phenomena are:

- people are quite good at building theories to explain their observations, but not as good at refuting or improving them; we tend to have overconfidence in our theories, and these then affect our interpretation of new data so as to become self-reinforcing;
- the propensity to theory-building, and other factors, lead us not only to predicting more than is warranted by the data but also to misdiagnosing new situations on the basis of our theories;

- we tend to attribute events to special characteristics of the involved individuals (or other factors in individual cases) rather than systematic, randomly operating influences;
- we suffer from "hindsight bias", so that we believe past events to have been more predictable than they were;
- we often reason about rare events via a "simulation" heuristic, i.e., by building scenarios for the rare events, causing significant errors in judgement;
- our judgement is affected (either through our recall of information or our interpretation of it) by the focus of our attention; e.g., especially vivid evidence is given more weight than is appropriate.

Much of the experimental evidence is still subject to different interpretations. Deviations from "normative", "rational" inference behaviour in laboratory experiments may often be attributed to lack of understanding, by the experimenters, of which task the subjects were really performing (e.g., due to the subjects simply not being familiar with statistical jargon or with problems of one-shot, optimal judgement on a limited set of evidence). The phase in research when many results were published showing very poor human judgement were followed by a stage of "revisionist" research trying to better bound the resulting pessimism, in view of the apparent general success of humans in many tasks. ([Jungermann 1986] contains a thorough discussion of the state of the debate at the time of its publication. [McClelland and Bolger 1994] surveys some of the models of intuitive probabilistic reasoning which have evolved to account for the existing body of diverse experimental results).

To understand the effects of this uncertainty of scientific opinion on the problems addressed here, I will discuss briefly one of the strongest "optimistic" views [Gigerenzer 1994], which maintains that humans are actually fairly well equipped for dealing with statistics in a *frequentist* fashion: the fallacies discovered in the "heuristics and biases" line of research would be mostly due to the experimenters' attempts to force the problems and the subjects' answers into a Bayesian, single-event-oriented view of probability, so as to determine both misunderstandings by the subjects and fallacies in the researchers' interpretation of results. This argument also points out that collecting statistics (counting events) is a natural task, so natural selection would have prepared us for it (though not for applying Bayesian probabilistic calculus). The experimental results show that many people will be perfectly able to reason about frequencies of an event over a (real or hypothetical) population, and yet be totally inept at solving equivalent problems stated in terms of probability. The archetypal problem in which this applies is that of judging the probability that a given patient has a disease, after the patient tested positive on a given test, knowing the false positive and negative rates for the test and the base rate of the disease over the population. People may ignore the base rate, and thus give a completely wrong answer, when these data are presented as probabilities, but answer correctly when they are given as numbers of events (e.g., the number of people with the disease who test negative) over a population. This observation points to a way for helping people to reason probabilistically. However, many of the dangers of which we should be wary when using the results of engineering judgement do not seem to fall into this category of misunderstanding probabilistic language. For instance, nature may well have equipped us with a fairly good event-counting mechanism, reasonably effective in most everyday situations, and yet the availability bias may often affect this mechanism, as shown by experiments.

Another problem with existing research is that very little has been observed first-hand about the behaviour of practitioners of different disciplines, except that wide variations have been observed between the few categories that have been studied. However, knowing which problems *have* been observed is obviously useful.

The following sections consider these problems in more detail. For each identified problem, I have added examples of tasks in dependability assessment that it can be conjectured to affect.

3.2. Building, improving, refuting theories

In interpreting our observations, we often seem to be overly eager to build explanatory theories whereby every detail in the observed data is the effect of a natural law rather than of chance. So, we can draw strong beliefs about correlation and causation from insignificant observed samples. Another effect is an inability to consider the "regression" effect whereby, in any series of observations, any extreme value is likely to be followed by another which is closer to the mean. In an example study of real-life experts, experienced instructors (in a flight school) were shown to believe that praising a student for a good performance leads to poorer performance the next time. In reality, the investigators were able to explain the variation in performance as "noise" around a slow learning curve, where each outstanding performance would naturally tend to be followed by less good ones (which were then falsely blamed on the praise that followed the better performance). This effect may well explain widespread beliefs in the value of punishment towards improving people's performance [Kahneman and Tversky 1982b].

There is also some evidence that, besides building theories on shaky bases, we tend to stick to them against evidence. Clinicians who had been observed to believe in "illusory correlations" between some test results and some clinical conditions (that is, they failed to take into account those observations that did not support the theory of a certain correlation) showed great difficulty in refuting the theory when prompted to re-examine the data and even when given faked data showing *negative* correlation.

A hypothetical story of unwarranted theories could be as follows.

1. We observe that a certain design, obtained by using a specific design method M, contains the defect D. We conjecture that the use of M may make it more likely for designers to err producing defect D. This conjecture is perfectly legitimate. To become a respectable theory it would need either a causal explanation, or an analysis of a sample of the four categories of designs (those obtained using M and containing D, those not obtained with M and showing D, etc.) showing a significant correlation. However, we are likely to start applying the theory as soon as we have seen a few cases in the "M and D" category, without considering, e.g., whether the "non M and D" case is frequent.

2. As we examine new designs, cases of "M and D" or "not M and not D" naturally reinforce our belief in the theory. For cases of "not M and D" or "M and not D" we may be able to: i) consider that method M was applied with some variation, or the observed defect does not really belong in class D; or ii) observe that the design problem (or the design team, or any other accompanying circumstance) had a certain peculiarity, explaining why the general law did not apply in this case[3]; or iii) think that this information represents a chance effect. All these procedures can be legitimately applied, if subject to explicit scrutiny, but if they are instead applied semi-consciously by our built-in mental mechanisms for "learning from experience", they are likely to reinforce baseless theories.

Furthermore, in many real-life decision problems we do not have the luxury of collecting uncensored samples. "Self-fulfilling prophecies" are but one example. If we decide not to adopt M because it may cause D, and then observe that the prevalence of D in new

---

3 There is actually some evidence [Goldberg 1986] that asking a subject to formulate such "exception rules" (which would explain an observed departure from the theory) reinforces the subject's belief in the theory, without further critical analysis.

projects does not increase, we have no data to confirm or refute our theory (though this lack of data may well strengthen our belief in the theory).[4]

When we do use new evidence to update a theory, and do take it into account, we seem often to be over-conservative: the corrections we make, e.g., in estimating the probability of an event, are smaller than prescribed by Bayes' theorem. Last, although one might expect experts in a discipline to be especially able to avoid fallacies in judgement in their field, there is also reason to expect that in some cases they may be more subject than "lay-persons" to the "confirmation bias" discussed above [Ayton 1992].

3.3. Overconfidence

Another important phenomenon is *overconfidence*. Generally speaking, people's confidence in their judgement tends to be excessive: they will describe their beliefs in terms of distributions that are too narrow. Such predictions are not "well calibrated". If a "well-calibrated" person utters statements like "I am x % confident in prediction X", it should turn out that in the class of all predictions for which the person's confidence was x, precisely x % of these are true[5].

Known experimental evidence is impressive in terms of the overconfidence bias that it has demonstrated, but it is difficult to judge its representativeness of expert performance in real-world problems. More importantly, it is quite difficult to infer expectations about any given expert's performance from research results. However, overconfidence *has* been observed in predictions of failure rates [Chhibber et al. 1992]. So, an expert's prediction of a narrowly distributed time to first failure, for instance, may follow from evidence which would warrant a much flatter, less satisfactory subjective distribution with the same mean.

3.4. Conditional probabilities in causal and diagnostic roles

In estimating a conditional probability, P(X|D) (event X conditioned on data D), people have been observed to be much more confident in predictions that follow the cause-effect chain than in inference from effect to cause ("causal inference" is more natural than "diagnostic inference"). For instance, people are more confident in inferring a son's height from his father's than vice versa, although both heights are correctly perceived to have the same distribution (and if P(A)=P(B), then P(A|B)=P(B|A)). Between two indicators of a third variable, people seem to predict more confidently on the basis of the indicator that is perceived as affecting the variable more strongly in a causal sense.

---

4 Another known phenomenon of theory-building is the difficulty of "taking a fresh look" at data after one has first interpreted them. Outside the statistical domain, a striking example is the likelihood that if a plant operator, confronted with an unexpected emergency situation, initially forms a wrong diagnosis (a wrong mental scenario of what is happening), he may be unable to revise the basic assumptions of this diagnosis on the basis of new evidence, choosing instead to revise details (e.g., by assuming that the new puzzling evidence comes from faulty sensors). A "fresh view" is needed to produce a new diagnosis that better fits the whole set of data, but may only come from a person who did not form the first diagnosis, e.g. an operator of the next shift.

5 In a variation of this experiment, one would ask many subjects to answer "yes or no" questions (the answers to which are known to the experimenter), each subject adding an estimate of his/her degree of belief in the correctness of his/her answer. By then calculating the fraction of respondents who were correct in answers for which they had stated a same degree of confidence, and comparing this fraction with the stated degree of confidence one can evaluate whether the sample of subjects is collectively well calibrated.

3.5. The illusion of control

It has been found that most people tend to rate their probabilities of incurring many types of accidents (e.g. driving accidents) as lower than average. Of course some of them may be right, but on the whole this is a badly uncalibrated prediction.

This can be attributed to a number of causes: one's apparent immunity so far; one's apparent prowess in avoiding accidents; the ability to identify the mistakes that led to other people's accidents, which in hindsight seem easy to avoid; the general fact, in the end, that we are in control, and this fact outweighs what we know about statistics for the general population. As anecdotal evidence that these mechanisms also operate when we evaluate design reliability, we can consider how frequently, after finding and fixing a program bug, one is (wrongly) convinced that the program is now correct.

Evidence like this invites us to be wary of accepting a developer's perception that the "excellence" of his development process guarantees a certain level of accomplishment [Hannaford et al. 1993]. Of course we do not know a priori that an individual statement of this kind is overoptimistic (*some* drivers are certainly better than average!), but we have no *a priori* reason to trust it at face value. Likewise, if we feel that a safety-critical program should be coded in assembly language, we should probably double-check whether we are overestimating the reliability advantage given by direct control on the low level code.

3.6. Causality

If we are naturally over-eager theory-builders, we should be wary of the way we proceed from data to cause-effect chains. Many software engineering experts believe that it is important to collect data about the software production process, so that organisations may learn how to improve these processes. However, we may be prone to learning too much from the data we have. Assume, for instance, that we observe the reliability growth exhibited by a product during debugging. If we mark on the time axis the time some change occurred in the development organisation, and we perceive some change in the pattern of failures after that point, we are likely to conjecture that a) the perceived change in the pattern is an actual change, and b) the change in the organisation caused the change in the failure pattern. Moreover, we are all too likely to treat this conjecture as a valid theory, unless we explicitly submit it to rigorous tests.

3.7. Hindsight biases

When judging past events, people indeed behave according to the folk-psychology law that "hindsight is 20/20". However, this is not necessarily due to an ego-serving bias. Rather, it may be ascribed to "creeping determinism" [Fischhoff 1982b]: the tendency to see a series of events as a linear cause-and-effect chain rather than an accidental sequence. When reviewing a sequence of events and decisions which ended in failure, we build a theory that predicts what we already know to have been the final outcome; then, the decisions which preceded it appear to have been wrong: we no longer recognise the dearth of information, or the ambiguity of the information available, at the time decisions were made.

This fallacy is seen by [Fischhoff 1982b], e.g., in professional historians, as well as in "lay" people. In the field of dependability, it may contribute to an excessive tendency to blame accidents on "human error" when operators misdiagnose a situation, rather than questioning whether the design of systems or procedures was likely to cause a wrong diagnosis [Reason 1990]; the same "fundamental attribution problem", of imputing errors to individual human defects rather than to error-prone situations and tasks, may lead to wrong estimates of the likelihood of human error both in designing systems and in operating them. Last, there is the problem [Reason 1990] of ad hoc solutions for perceived dangerous scenarios and neglect of those scenarios that are not so easily imagined. If a specific sequence of events is found (by analysis or by observing an incident/accident) that may cause an accident, the reaction of a designer or decision maker

may be to devise specific "patches" to prevent or tolerate that specific scenario. If this scenario was seen as an exception in a safety analysis indicating satisfactory safety, the patch restores trust in the analysis. This response may be appropriate in a simple system where the accident scenario in question is clearly one of the more probable ones. However, if it is only one of many, individually very unlikely accident scenarios, eliminating it may be irrelevant or counterproductive (through side-effects of the design patch on other unlikely scenarios).

3.8. Simulation heuristic for rare events

It seems likely that when intuitively evaluating the probability of a rare event, we do so by building mental scenarios, that is, plausible chains of events that would lead to the event of interest. Our estimate of probability will grow with the ease of conjuring such scenarios and with their number.

A dependability-related example may be the following. Checklists for discrete control systems may include: check that the system's outputs vary as specified while we systematically set to TRUE one input at a time, with the others kept at FALSE. This procedure gives an illusion of completeness, but is obviously insufficient to determine that the controller is defect-free. However, such a testing strategy is sometimes included among the evidence of reliability without an estimate of how much it really proves, and may be expected to lead to overly optimistic conclusions.

It may be noticed that building hypothetical scenarios is an indispensable mental tool for exploring the space of possibilities, finding counterexamples for one's conjectures, and building robust strategies. Once more, the problems arise from misuse of a useful tool.

3.9. Tricks of attention

The vividness of evidence has a relevant role in determining its effect on intuitive judgement, possibly due to the "availability heuristic". Paired with the difficulty of drawing statistical inference without explicitly applying the rules, this should probably discourage the practice of presenting the raw results of software engineering experiments. Such results should probably always be accompanied, and overshadowed, by explicit indications of the conclusions that can justifiably be drawn from these results: practitioners may otherwise be overly influenced by statistically insignificant data, like some extreme case observed in the sample. A similar warning probably applies to the use of coverage indicators in software testing: 100 % success on a sample (test set) with 100 % coverage of program structure (however defined) may make us forget that a sample (of the population of possible test cases) satisfying such coverage criteria is biased in an unknown way.

Another problem which may be expected is similar to the observed phenomenon that, e.g., if a discussion group includes only one woman among many men, she is perceived by observers as doing more of the discussion than she actually does. So, the parts of a system which are most innovative, of most interest to the expert, or most subject to controversy may be given an excessive weight in intuitive judgement. A common discussion is whether the "most critical" part for the assessment of a complex system is the software, or the actuators, or whatever. While "most critical" could be given a rigorous meaning and then the discussion could be led in scientific terms, one should guard against the possibility that its intuitive perception biases the weights given by an expert to the probabilities of different events.

3.10. Biases from emotion rather than from heuristics

It is worth mentioning that other factors, besides those internal to the reasoning mechanisms of the experts, may cause biases. An expert's prestige may be damaged by admitting uncertainty, and this would lead to overconfident statements. For a medical

doctor, high confidence in a positive prognosis implies higher risk of a malpractice suit than "correct" confidence. Such external causes of bias ought to be reduced.

Here it seems reasonable to consider that decisions regarding risks are made difficult by their emotional overtones. If "engineering judgement" is required about the reliability of a subsystem, the experts about the subsystem may know that this judgement is crucial in deciding whether a larger system will kill people. They may then succumb to the common reaction of denial of hazards, by neatly separating hazards in two well-separated classes: some that are too probable, and hence should be eliminated or neutralised, and some that are vanishingly improbable and can safely be ignored. They will be likely to believe that the hazards in the two classes coincide with those that have been respectively avoided and ignored in the actual design; otherwise, they would be in a severe conflict situation. So, this defensive bias in judgement will automatically classify any system as practically hazard-free.

Such considerations are relevant when expert judgement is used, and have been cited here for completeness, but discussing them in detail is outside the scope of this paper.

4. Assessing expert judgement

4.1. Predictions from past performance

To assess the trustworthiness of an expert's judgement (an individual statement of prediction about a system of interest), one could start from the past predictions of the same expert and how many of them turned out to be right. This method is available, e.g., for checking how good a meteorologist is, since his/her predictions can be checked every day against reality[6]. This "black-box" measurement of performance is more difficult when dealing with predictions of very low probabilities. By considering the predictions of many experts about many events, one could estimate some global goodness for all experts; but nothing could be said about an individual prediction about an individual event, unless the prediction is seen to fail dramatically. The controlled experiments which showed experts to fail dramatically in prediction are a reason for caution, but not a clear assessment. There is not , and there cannot be, any strong evidence of *good* judgement for small sets of predictions about very rare events. And, of course, we are most interested in those individual experts on whom we depend in each individual case, rather than in broad categories. However, there we have even less hope. Suppose that an expert in the safety of products of a certain class has analysed, during his/her career, twenty such products, and judged that they all had a probability lower than ε of causing an accident in the next 50 years. No one of these products has produced an accident yet. If some of them had, we would probably trust the expert less than previously, of course (though how much less? Analysts still disagree as to whether, or to what extent, the Three-Mile Island incident refuted or not the conclusions of the Reactor Safety Study). Unfortunately, the fact that no accident has yet occurred is no great validation of the accuracy of this one expert.

Even when comparing predicted and actual outcomes gives too little information, experts can at least be assessed for (presumed) necessary conditions of good judgement, like consistency (or "reliability", as it is often called by psychologists). For instance, a good judge should presumably judge consistently every time he/she is presented with the same evidence, irrespective of when this is done and of which additional irrelevant evidence

---

6 Notice, however, that this is not the easiest task when predictions are allowed to be in terms of probabilities ("40 % chance of rain tomorrow") rather than deterministic ("rain tomorrow"). The old philosophical problem arises of defining the "true probability" of an individual, non-repeatable event. A proposed indicator of [a likely necessary condition for] proficiency is "calibration" (cfr. footnote 5); but a way a meteorologist could achieve "perfect" calibration is to predict the same probability of rain every day, irrespective of observations, with the precaution of using, for the prediction, the average probability over all the days in the year.

accompanies it [Einhorn 1986]. Likewise, consistency with other judges is often considered a necessary condition. Predictions in terms of probabilities should presumably satisfy the axioms of probabilities (although the language of probabilities may be unnatural for many experts, so that alternative languages could be necessary, as discussed in 5.2.2. One might even find judges who cannot state their beliefs consistently in terms of any formalism for representing uncertainty, and yet prove good at predicting the actual events - such experts may be found and their predictions used, if the events of interest are common enough that the experts' performance can be statistically measured). Last, one might presume that unless an expert proves good at simple predictions in his/her area of expertise, he/she cannot be trusted for difficult predictions. So, in theory, judges could be tested for all these necessary conditions of good judgement, on real but comparatively easy tasks (like statistical inference about frequent dependability-related events) as well as on fictitious tasks related to their area of expertise.

4.2. Studying the judgement process

Rather than assessing an expert as a black box, one can examine and challenge the mental processes which produced the current prediction, to correct errors and reach a correct prediction. Of course, the word "challenge" here does not imply any preconceived hostility, but just the systematic scepticism which is inherent in the scientific attitude (the experts themselves may be the "challengers"). So, we need to describe or model the expert's reasoning. To be more precise, we can describe an expert's judgement process as a function from the multi-dimensional space of evidence about the system (different measures on the situation to be judged) to a dependability score for the system[7].

Models of experts can, however, be built at different "depths". Any discussion of the issues involved is bound to repeat the debates within the Artificial Intelligence research community, so I will only summarise the essential choices. At the two extremes, we may try to reproduce:

- just the externally observed *behaviour* of a human expert. Such a "behavioural" model is an input-output function from cues (evidence) to opinions, and we shall judge it based on how closely and reliably it reproduces (or predicts) the behaviour of the expert. A behavioural model is rather "trained" (e.g., by linear regression, or like a neural network) to mimic a human (experts who are not consistent with themselves, of course, pose problems both in training the model and using its outputs), than "designed" to be a mechanical expert. For human experts who have proven good at their tasks, a model like this could be both a cheap substitute (in unimportant tasks!) and a synthetic "challenger", telling them when they seem to depart from their usual behaviour and prompting them to re-examine their criteria. When no strong evidence exists that the experts are good, reproducing their behaviour without understanding it may be risky; however, building a behavioural model would allow one to identify the important cues used by experts and study whether they are appropriate for guiding judgement, evaluate whether the weights used are correct, etc. All this knowledge could then help in building a more "correct" model of how experts *should* behave;
- the behaviour of the expert when reasoning *correctly* (or, how experts *should* behave). Such a "rational" model describes a formally defensible chain of inference and deduction. When a rational model disagrees with an expert, one can, in theory, check the model's argument, and i) find logical flaws (bugs in the

---

7 We could also model an expert who is not consistent with himself - whose reactions to the same set of inputs vary - by a function from the space of evidence to a distribution of scores; but when we are interested in synthesising a "good" expert, who is supposedly consistent, we do not need this complication.

model) or ii) disagree on the premises used in the argument, or iii) if neither of these previous cases occurs, conclude that the human expert is wrong.

Intermediate levels of "depth" are possible, where some part of the algorithms in the model emulate logical processes and some simply mimic observable behaviour. And, of course, one can instead aim at modelling the "real" operation of the mind or brain, with its heuristics and biases [Reason 1990], an endeavour which, if successful, should offer a model comparable for trustworthiness to a behavioural model.

Linear models (where the mapping from the multi-dimensional evidence to the judgement is a linear function of the various measures used as evidence) are rather popular, and thus deserve a brief comment. Of course, many rational decision algorithms are non-linear. A well-behaved function can often be approximated by a linear function *only in the immediate vicinity of a given point* in the domain space. So, a linear model with weights that are adjusted depending on the subset of the input space where it is applied may be appropriate. In simpler terms, the algorithm becomes: first check that you are in the subspace X, then apply the linear model M_X. To know whether this is a reliable procedure, we should first find a true model of the algorithm we want, and only then we can look for suitable approximations (in practice, of course, the discovery of a linear statistical relationship may also be a stage on the route towards a true model). It has, however, been observed that linear models tend, in many fields, to outperform the individual experts who "trained" them, presumably because they capture the essence of the expert's behaviour but apply it more consistently than humans. A more complete explanation [Dawes 1982] is that in general linear models, used in problems of prediction with great inherent uncertainty (that is, where predictions are often wrong, but it is difficult to do any better), are very robust with respect to the weights used, provided that their signs are right. In essence, these models simply capture which cues reinforce and which weaken the belief of the expert.

4.3. Reasonableness checks and diversity

It is often possible to apply different methods of reasoning to a problem and compare their results, or check whether the consequences of a stated opinion are all reasonable if compared with independent evidence. Such checks may also allow one to spot errors and improve previous conclusions, and will be considered again in the next section. A very common form of diversity is that of employing more than one expert. When several opinions are available, consistency between them is often considered a necessary condition for correctness [Einhorn 1986]; the problem with this is that, with difficult and controversial issues, disagreement is normal, and there is no simple method (say, majority vote) for deciding who is right. Furthermore, consensus in the conclusions is hardly a guarantee of correctness in difficult problems, unless the methods used are also scrutinised. So, disagreement can be used at least as, and probably just as, an indication that a thorough revision of the evidence and inferences used is necessary.

5. Remedies

5.1. Generalities

If one has to use an expert's informal judgement as a basis for a safety-related decision, at least two questions are appropriate: how good (trustworthy) is this person's judgement, and what precautions can be taken to make it as good as possible. The former question seems very difficult to answer, as discussed above. We have very little reason for trusting our individual expert as an intuitive judge, and we have good evidence that other experts (or experts in general) are prone to well-known fallacies.

This prompts us to be cautious. Moreover, we know that the current beliefs about how the mind works lend little support to the hope that it mimics a perfect scientist's conscious

thought processes. This, at least, is a clear indication: the tool of intuitive judgement is not perfectly fit for its purpose. Unfortunately, there are tasks for which we have no other tool, and we must use all possible precautions to use it at its best. We will not necessarily, through these precautions, acquire a knowledge of *how good* the expert is; but we will obtain a better judgement than we would without those precautions. One may observe the similarity with the problem of safety itself: there are ways to improve it, but beyond a certain level of safety we no longer know how much (or even whether) we gain by applying these methods.

The first remedy is of course to substitute, whenever possible, formal scientific reasoning for the expert's intuitive assessment. This usually calls for the experts to be able to list the facts they know, the inferences they draw from them and the deductive rules they use based on the known laws governing the behaviour of the system to be judged, and the way they then build their conclusions. An independent assessor (or the expert himself, of course) can then double-check all these individual items, represent the expert's reasoning (or the way the expert *should* have reasoned) in a rigorous form and subject it to formal verification and if necessary to corrections.

All this should be done *when possible*, or rather when feasible, given the time, money and personnel available to the decision maker, and the limits to the complexity that any human mind can master. In any case, reducing the area where intuitive judgement is needed, so that it is less critical and/or more reliable, and improving intuitive judgement itself, are all useful steps. The literature suggests means towards this end. A decision maker who is conscious of the problem of experts' fallibility can seek means to:

- change the experts' tasks to make them less error-prone;
- change the experts' tasks to make them more amenable to analysis;
- help the experts in detailing their evidence, deductions and inferences;
- help the experts in finding and correcting fallacies in their reasoning;
- make the best use of the availability of *multiple* experts.

In the rest of this section I enumerate some plausible means for improving the results of engineering judgement. Table 1 is an attempt to summarise some known risks and remedies. As will be noticed, and will appear from the following discussion, the remedies overlap both in terms of which problems they may attenuate and in which cases they should be applied. A few of these remedies are simple prescriptions against specific fallacies, easy to apply mechanically (e.g., "use odds rather than probabilities"), which are, however, derived mostly from laboratory experiments, and might well be ineffective in a specific case of interest. Towards the bottom of the table I have collected the broader-scope precautions, amounting to principles of good scientific reasoning. These should always be used, but what their application amounts to in practice is determined by the details of each case, and their effectiveness depends on the skill and competence of the people involved.

Task	Origin of mistakes	Possible precautions or remedies
Producing from own experience a statement of probability about an event	Availability bias, effects of attention	Separate the task of enumerating relevant events from that of extracting statistical statements
Producing from experience a statements of correlation	Illusory correlations	Formalise task, tests of significance
Stating subjective probability distributions	Overconfidence	Ask about probabilities of ranges of values rather than percentiles; use frequentist, not Bayesian framework
Updating own beliefs with new evidence	Conservative updating "anchoring"	Use odds rather than probabilities (but see Note 8)
Predicting probability of event by combining case-specific evidence with information about the general population	Excessive weight to case-specific evidence (neglect of base rate)	Formalise procedure; make the subject sample the population ("experience the base rate") rather than being told what the base rate is
Deriving a statement of probability by combining probabilistic statements	Difficulties with the calculus of probability	Formalise application of probability calculus; state problem in terms of frequencies of events
Producing a statement of probability for an unlikely or implausible event	Use of scenario-building heuristics	Look for alternative scenarios; formalise difference between counting scenarios and stating probabilities
Any tasks	Problem complexity: expert uses heuristics instead of probabilis-tic reasoning	Decompose task into simpler subtasks
	Misunderstanding of question	Clarify questions; provide alternate formulations
	Expert has difficulty expressing knowledge in terms of probabilities	Assist expert in stating knowledge in terms perceived as more appropriate, then re-state it into rigorous (probabilistic or other) terms
	Any	Reasonableness checks: show the consequences of the expert's state-ment in different terms or on diverse aspects of problem; make the reasoning of the expert explicit and formal
	Any shortcoming of individual expert	Provide assistance, forewarning of problems, feedback, training; change the experts; evaluate error and recalibrate judgement; make decisions robust with respect to judgement errors

Table 1

5.2. Elicitation of judgement

5.2.1. Formalism, asking the right questions

In debates in dependability assessment, one can observe some simple but common problems which cannot but detract from the reliability of judgement.

Simply defining the questions asked with sufficient rigour may make a big difference in the ability of people to judge properly. [Chhibber et al. 1992] lists common mistakes, like reasoning about failure rates without specifying for which operating conditions, so that questions and statements are interpreted inconsistently by different experts or at different stages of a study. In the software field, it is still common to have arguments about software testing in which the purposes of detecting faults and of estimating reliability, and the meaning of counts of faults in the product and of failures over time, are confused.

An important issue seems to be, simply, whether the question asked is appropriate. For instance, a specialist may not be qualified to state that a system has a certain probability of failure, but may be able to argue rigorously that the system is to be considered more reliable than another system, based on a sound model of the structures of the two systems and known reliability data. Now, if the structure of a safety case requires, to fill it, statements that the experts cannot reliably produce, the experts may be unable to recognise the problem. It seems that the safety analyst must explicitly investigate which questions the available engineers can answer more reliably, before settling on the final structure of a safety case.

5.2.2. Asking questions in the right terms; changing the formalism for representing uncertainty

Intuitive judgement is affected by how a question is posed (different forms of the question elicit inconsistent answers). Many instances of poor judgement by experimental subjects seem to arise from the fact that the chosen notations for the statements or measures of interest are not familiar or intuitive for the subjects. For instance, there is some evidence that:

- asking people explicitly for subjective percentile values of a distribution makes them more prone to the "anchoring" problem (leading to a narrow distribution around the median or mean, used as an "anchor") than asking for the probability that each in a series of values of the random variable will be exceeded (although of course the two sets of questions are formally equivalent);
- changing the question posed from an absolute evaluation of some measure to ranking of the measures of appropriate different objects can improve the consistency of the answers [Anderson 1986];
- asking questions in frequentist terms ("how many times would the event happen in a hypothetical sample of 100 similar situations?") rather than in Bayesian, single-event probability terms ("what is the probability of this event in the situation at hand?") may avoid those mistakes which are due to the unfamiliar nature of the latter formulation [Gigerenzer 1994][8].

More considerations are found in the literature about formalisms for representing uncertainty (Bayesian vs., e.g., Shafer-Dempster or fuzzy logic. Surveys are found e.g. in [Hollnagel 1989; Wright and Cai 1994 Ng and Abramson 1990; Saffiotti et al. 1992]). However, all the recommendations above share the property that they allow the expert to

---

[8] However, which form of questions are best at eliciting correct answers may well very between groups of people. form instance, [Bolger and Wright 1994] points out that asking for statements in terms of odds rather than probabilities ("4 to 1" rather than "0.8" or "80 %") has been shown in some experiment to reduce certain biases, while other experiment showed the opposite to be true.

produce statements that make as much sense as possible to him/her in intuitive terms, and still have a clear and non-misleading formal meaning in a probabilistic context.

Another, related consideration is that many experimental subjects (lay or expert) may "fail" a test for probabilistic reasoning for the "legitimate" reason that they are not in fact solving the given problem by a probabilistic strategy but rather by a "knowledge-based" strategy [Beach and Braun 1994]. It seems that the kind of strategy chosen is affected by cues in the presentation of the problem (evident elements of chance and repeatability would cause one to favour probabilistic reasoning). [Curlo and Strudler 1993] claims (based on experimental results) that people use "causal" reasoning not only to integrate but also to override statistical reasoning[9]. We should derive two consequences from these observations. First, if we wish an expert to reason probabilistically, we should present the problem so as to prompt that style of reasoning. More importantly, we may fear that we may thus cause the experts to neglect some of the knowledge which they would use in "non-probabilistic" reasoning: we should then strive to obtain this knowledge, in whichever form the experts can state it, and use it. However, this cannot amount to accepting the experts' opinion without scrutiny: we also wish the conclusions derived from this knowledge to be sound. We need then to re-express the derivation process (and thus the knowledge itself) in a rigorous formalism which can be subject to proof or confutation, for instance (though not necessarily only) the language of probabilities.

5.2.3. Challenging the expert's opinion

It is recommended that an analyst interviewing an expert should manipulate and vary the questions so as to highlight any inconsistencies in the answers, so that the expert can try and correct errors in reasoning and express his/her "true" belief.

Although I have found no specific reference to this effect, it would seem that asking an expert simply to justify a conclusion may well be counterproductive, as the conclusion would tend to dominate his/her new exam of the evidence. It would be better to separate the decomposition of the argument into individual inference steps, and then consider each step in isolation; or to derive and represent the consequences of the expert's reasoning in a form different enough from that of the expert's own statement that he/she could scrutinise them without bias.

As people are often conservative in revising their judgement on the basis of new evidence, it seems possible that, even if the questioning makes the experts realise that they neglected some evidence he knew, they may not be able to change their previous answers as much as they should. Presumably, making the revision of the conclusion explicit (in Bayesian terms) would help.

[Fischhoff 1982a] lists a long series of methods for "debiasing" and discusses their efficacy. For instance, "hindsight bias" appears to be quite "robust" with respect to how the problem is posed, and quite impervious to most attempts to restructure the task. A useful technique is to ask the experts how they would explain the *non*-occurrence of the event. However, it is not known how much this procedure would improve predictions, and whether it might be self-defeating in making experts over-confident that they have overcome their hindsight bias.

---

9 The experiment presented there is reminiscent of a common situation in dependability assessment: the subjects have to choose a bicycling helmet using information about the accident statistics, design details, and manufacturing standards of various brands. The authors' conclusions seem stronger than those of [Beach and Braun 1994]: "While the most important factor in deciding whether an event is evaluated probabilistically or causally is the availability of appropriate information, other factors contribute to alter the appropriateness of probabilistic reasoning, as perceived by an individual [...].. ; highly specific and precise probabilistic information may in fact encourage causal reasoning when evidence of a causal process is available".

A partially effective remedy against overconfidence is asking people to look for reasons why they might be wrong. In general, observed calibration varies widely between categories of experts (e.g., it is good in weather forecasters, bad in doctors; interestingly good in reporters specialising in horse races). A habit of thinking in probabilistic terms is probably important. Recalibration methods exist, whereby one would correct the experts' own confidence statements. However, they may cause experts to alter their habits; and some recalibration methods lead to recalibrated "probabilities" which are no longer true probabilities (they violate the axioms of probability).

5.2.4. Cross-checking

One may often detect errors in intuitive reasoning by simple calculations and comparisons with other available knowledge, as mentioned before under the heading "reasonableness checks and diversity". Examples are "back of the envelope" calculations using different methods ("how does this probability of operator error vary if I decompose the operator action in a different way?"), comparison with situations different from that under consideration and where more knowledge is available ("how does this prediction compare with the observed behaviour of other systems? If it differs markedly, does the knowledge available about this system warrant so strong a departure from the average of the population?"), checking that parts of the assessment performed separately did not rely on incompatible assumptions, etc. Such checks could be proposed both by the expert engineers themselves, if they depend on special properties of the system under consideration, or by a less specialised collaborator or analyst, who has a better chance of "seeing the forest despite the trees".

5.2.5. Checking intuitive statistical reasoning through causal reasoning

The use of scenario-building as a way to build intuitive estimates of probability would not be a fallacy if two conditions were satisfied: i) the scenarios evoked by the expert were the whole set of possible scenarios, and ii) the expert were able to assess and sum the probabilities of all these scenarios. Making the scenario-building activity explicit may be sufficient to eliminate the illusion of completeness. For instance, testing strategies for complex systems often aim at being "complete" in some intuitive sense which does not necessarily warrant trust that all defects can thus be found. Simply finding the classes of defects that a testing strategy would not detect may be sufficient to avoid excessive overoptimism based on the results of a testing campaign.

5.2.6. Systematisation of tasks

Changing intuitive statistical tasks into more explicit ones has an important role. For instance, [Kahneman and Tversky 1982a] suggests the following procedure to correct for people's tendency to "non-regressive" prediction, i.e., to excessive reliance on information about the individual case about which prediction is sought, compared to information about the population to which it belongs. The expert is guided through a sequence of steps:

- election of a reference class;
- assessment of the distribution for the reference class;
- intuitive estimation for the individual case, based on available information;
- assessment of predictability: the expert is guided to assess the predictive power of the information available about the individual case. This may still be based on the expert's own judgement, applied to questions about different hypothetical situations: e.g., how often would the expert expect, if confronted with two specific cases, to correctly predict at least in which of the two the unknown variable would have the greater value?

- correction of the intuitive estimate, where the expert is shown how the intuitive estimate can be made more regressive, using the expert's own assessment of predictability, and can then choose to correct one or the other judgement to improve the prediction.

5.3. Organisation of decision-making

5.3.1. Separation of roles

A common enough problem is that the knowledge that is needed to reach a decision is divided among different experts. An expert engineer, with much experience in the problem of interest, may be untrustworthy as an intuitive statistician. Rather than asking such engineers to produce a judgement, it would be useful to obtain from them the knowledge upon which they would base this judgement, and let an expert statistician do the inference. A common recommendation is, therefore, to separate properly the roles between the engineer expert and the statistical analyst. Another dangerous confusion is between the tasks of producing evidence for a decision and of producing the "values" or goals on which the decision will be based, or between the roles of expert and decision maker. Examples abound in public policy decisions, where experts are asked to suggest solutions before the public goals have been spelled out, but can also be found in industrial contexts, where an expert's decision on operability may be solicited. The expert may not know the goals of the management or of the regulator, or the expert may end up being confused by the added complication and pushed into neglecting evidence. Although in many cases a single person may have to fill more than one role, the advice to be aware of the necessary division of these tasks seems appropriate (if one really needs to "take off one's engineer's hat and put on one's managerial hat", this switch should at least be conscious and explicit).

5.3.2. Multiple experts

There is much literature in risk assessment about the use of multiple experts. A common approach is to try and "combine" the experts' conclusions. For instance, after asking experts for their subjective distributions for a variable of interest, one can repeatedly "update" (in the Bayesian sense) one of these subjective distributions, using as evidence the next expert's distribution with likelihood functions which appear appropriate based on knowledge about the experts [Wright and Cai 1994]. The results of procedures for combining expert opinions are reported e.g. in [Van Steen and Cooke 1989].

Such procedures can be made comparably simple, but they make no attempt to improve the experts' opinions to start with: a shared bias would survive the combination process without being revealed. Some researchers therefore argue (quite rightly in my opinion) that the multiple experts available should be used to criticise and improve one another's reasoning; they can point out fallacies related to their technical knowledge (e.g., omissions in a fault tree, neglect of some relevant past evidence), and, interacting with a professional statistical analyst, in intuitive inference steps. In other words, the intention is to move as much of the process as possible from intuition to reasoning (including reasoning about uncertainty): to quote [Kaplan 1992], "Weigh evidence, not experts!". The expected result is either a consensus, or a clearer understanding of where disagreement really exists among the experts, what degree of uncertainty it introduces in a final decision, and what could be done to reduce it. Of course, there is some risk of undesired psychological effects from group interaction, in the form, e.g., of irrational tendencies to unwarranted consensus or dissension. Detailed procedures for organising such sessions are indicated in [Kaplan 1992; Ortiz et al. 1991]. The aid of specialists is considered necessary to help the experts to limit the effects of both individual psychological biases (as seen above) and of undesired group effects.

Among the suggestions made in [Fischhoff and Whipple 1982] for public health policy decisions, there is that of using "quasi-experts" to facilitate exchange and/or collate and cross-check the experts' arguments.

5.4. Training

5.4.1. Probabilistic thinking

There is some consensus that training people in probabilistic thinking improves (as should be hoped) their performance even in intuitive inference, e.g., by reducing overconfidence. It seems obvious that this should improve communication between engineering experts and decision makers, or make the engineers better decision makers, when they have both roles. For instance, [Keeney and von Winterfeldt 1991] reports favourably on an attempt to improve the elicitation of probability judgements to be used in nuclear safety assessment: the engineering experts received training and assistance by experts in probabilistic evaluation, and on the basis of this experience a new elicitation procedure was subsequently specified.

5.4.2. Learning from experience, feedback

Overconfidence (the tendency to produce subjective distributions which are too narrow), seems to be reduced (in controlled experiments) by training the subjects, with feedback about their own performance, and coaching about the relationship between feelings of certainty and numerical expressions. Training seems also to be effective in real-world professional settings, while real expertise in the problem is, by itself, no defence (e.g., in samples of bankers, clinical psychologists, civil engineers,...) [Fischhoff 1982a].

The often-quoted, very good calibration of professional meteorologists is explained by [Edwards and von Winterfeldt 1986] in terms of very favourable conditions: frequent forecasts, feedback from them, and systematic scoring of their performance, known to them and tied, to some extent, to their wage and promotions. Similar considerations seem to apply to the less often quoted, good calibration of horse-racing betting specialists [Lichtenstein et al. 1982]. Reproducing these conditions for other experts may be difficult. However, if engineers are to be used as expert probabilistic predictors, one could attempt to systematically give feedback about predictions, if possible, and even elicit more frequent predictions to improve judgement.

Techniques have also been proposed which allow an analyst to correct the experts' overconfidence; however, the right corrective factors are a function of the problem and the techniques require, therefore, a knowledge of the difficulty of the problem or of the performance of the expert in a comparable problem.

6. Conclusions

The first conclusion which stems from these considerations is, of course, a healthy scepticism about the trustworthiness of engineering judgement as a basis for answering difficult questions. It would be easy to describe the research available as just confirming that we must expect intuitive, non-formal reasoning to be easily flawed by mostly well-known human weaknesses, and that common-sense remedies may help a little. More optimistically, one can state that the existing literature can offer an improved understanding of how bias is built into judgement, and evidence as to which "common-sense" remedies are indeed useful.

There are wide variations in the performance of those categories of experts who have been studied, so that there is usually no direct evidence that a certain professional category is as unreliable as one might infer from this survey. However, given these very wide variations, a decision maker should be aware of these problems which may affect the judgement of

experts. For disciplines where engineering or expert judgement is critical for decision-making, more experimental research is desirable about the trustworthiness of the specific categories of experts involved, and about which conditions tend to improve it.

Two defences seem to exist, and these need to be used together, for reducing the criticality of this problem: the first one (and the best remedy, as far as it goes) is to make intuitive judgement unnecessary, as often as possible, by turning it into systematic, scientific reasoning. This requires the experts to be able to spell out the evidence and the procedures leading to their conclusions. So, precautions that help the experts in this task, attempting to formulate clearly the questions asked, and to lay out the judgement process in a rational form, are useful. However, proper descriptions of very complex arguments will still be unfeasible: in the end, we are confronted with the limited resources of the human mind.

Despite these limits, it is clear that every judgement process may be improved. No one can muster the resources needed for a perfect judgement; yet, in many cases, it is possible to rely on something more sophisticated than the "gut feelings" of the experts. The second defence is, therefore, to modify or aid the task of the experts in ways that have been shown to improve their performance: ways exist for decreasing the risk of error from the part of the process that is left in its intuitive state. Among these are attempts to double-check facts, rules and conclusions, to shelter the expert from known error-causing factors, and so on. A general result of research in decision-making is that presenting the same question in different forms tends to elicit widely different answers, and that some of the possible forms are less error-inducing than others. Ways to improve decision-making have been studied in some depth for such critical and high-visibility areas as risk assessment for nuclear power, or environmental policy-making. Although there is no definitive consensus about this problem, decision analysts have developed "tricks" for eliciting "better" answers, which can usefully be applied in solving difficult dependability assessment problems. When reduced to using "engineering judgement" to produce an assessment of very high dependability, it seems that a proper checklist about the quality of this assessment should include at least:

- was an attempt made to formalise the reasoning used?
- was the remembered evidence checked against the records?
- were the expert's assumed correlations and conditional probabilities properly elicited, cross-checked with facts, and challenged to prompt the expert's criticism?
- were the known defences employed against the inherent biases of intuitive reasoning?
- was appropriate computer support made available to reduce the problem of complexity in the reasoning?
- was evidence about the quality of the expert used, if available?
- were different experts or quasi-experts asked to debate their respective arguments?

This checklist includes methods for making judgement more scientific and hence trustworthy, not, evidently, for making it always correct. Even if a really complete analysis were possible for a real-world situation, it would not remove the basic limitation that no amount of empirical information would allow one to predict the future with certainty. Any physical "law" may be refuted by a single new experiment. The "scientific" character of an analysis or argument is a matter of degree, rather than of kind, and the practical question is whether an analysis is "scientific enough", given the weight of the decisions that must be based on it. This paper has argued that, in view of current knowledge, the way engineering judgement is commonly used is not "scientific enough", and that there are ways for improving its use.

A separate issue - dealing with improving the general quality of judgement rather than specific instances of it - seems to be that judgement can be improved by specific training and by providing as much feedback as possible. It would seem that the common (or

increasingly common) practices of revising the safety case for an installation every so many years, of collecting reliability data and of providing fixes for "bugs" found in designs during operation, could be made the basis for a more formal process of improving not only the judgement about individual systems, but also the ability of engineers as judges and the knowledge of this ability.

Last, I must mention that the literature about practical applications of these methods belongs mostly to narrow (and "wealthy") fields, like risk assessment for nuclear power, or environmental policy-making [Keeney and von Winterfeldt 1991; Ortiz et al. 1991; Thorne 1993;]. The techniques suggested often require a costly selection of engineering experts, elicitation sessions with the help of professional analysts, etc. This does not make these considerations and techniques inappropriate for the wider field of dependability assessment (and decision-making based on it). On the one hand, recognising a need is the first step towards procuring the resources for satisfying it; and there are fields where dependability assessment already costs large amounts of money (viz. aircraft certification), or are going in that direction (cf. the large investments being made in ISO-9000-related activities). In these sectors of industry, redirecting some of this investment to improve its effectiveness would not be a large problem. On the other hand, most of the suggested safeguards amount to injecting some scientific discipline into otherwise obscure processes. These safeguards, and the knowledge itself of the problems, can be useful to anyone performing or using engineering judgement, even in less structured environments where all the roles in the process (expert, statistical analyst and decision maker) have to be performed by the same person.

Acknowledgements

This work was supported in part by the ESPRIT Basic Research Action 6362 "Predictably Dependable Computing Systems" and by project SHIP, "Assessment of the Safety of Hazardous Industrial Processes in the Presence of Design Faults" (Contract EV5V-CT92-0103 under the EU Environment Programme). Previous versions of this report appeared as SHIP and PDCS Technical Reports. Several colleagues have provided useful input and criticism, and I would especially like to thank David Wright, Bev Littlewood, Robin Bloomfield and Antonia Bertolino.

References

[Anderson 1986] N.H. Anderson. “Algebraic rules in psychological measurement,” in *Judgment and decision making: An interdisciplinary reader, ed. H. R. Arkes and K. R. Hammond*, pp. 77-93, Cambridge University Press, 1986.

[Ayton 1993] P. Ayton. “On the Competence and Incompetence of Experts,” in *Expertise and Decision Support, ed. G. Wright and F. Bolger*, pp. 77-105, Plenum Press, 1993.

[Basili and Green 1994] V. Basili and S. Green, “Software process Evolution at the SEL”, *IEEE Software*, 11 (4), pp.58-66, July 1994.

[Beach and Braun 1994] L.R. Beach and G.P. Braun. “Laboratory Studies of Subjective Probability: a Status Report,” in *Subjective Probability, ed. G. Wright and P. Ayton*, pp. 107-128, John Wiley & Sons, 1994.

[Bolger and Wright 1994] F. Bolger and G. Wright, “The quality of expert probability judgement: issues and analysis,” *Expert Systems*, vol. 11, no. 3, pp.149-158, 1994.

[Chhibber et al. 1992] S. Chhibber, G. Apostolakis and D. Okrent, “A taxonomy of issues related to the use of expert judgments in probabilistic safety studies,” *Reliability Engineering & System Safety*, vol. 38, no. 1-2, pp.27-45, 1992.

[Curlo and Strudler 1993] E. Curlo and A. Strudler, "Causal inference as a cognitive strategy," *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 5, no. 1, pp.57-71, 1993.

[Dawes 1982] R.M. Dawes. "The robust beauty of improper linear models in decision making," in *Judgment under uncertainty: heuristics and biases, ed. D. Kahnemann, P. Slovic and A. Tversky*, pp. 391-407, Cambridge University Press, 1982.

[Edwards and von Winterfeldt 1986] W. Edwards and D. von Winterfeldt. "On cognitive illusions and their implications," in *Judgment and decision making: An interdisciplinary reader, ed. H. R. Arkes and K. R. Hammond*, pp. 642-679, Cambridge University Press, 1986.

[Einhorn 1986] H.J. Einhorn. "Expert judgment: Some necessary conditions and an example," in *Judgment and decision making: An interdisciplinary reader, ed. H. R. Arkes and K. R. Hammond*, pp. 480-492, Cambridge University Press, 1986.

[Fenton et al. 1994] N. Fenton, S. Pfleeger and R. Glass, "Science and Substance: A Challenge to Software Engineers," *IEEE Software*, pp.86-95, July 1994.

[Fischhoff 1982a] B. Fischhoff. "Debiasing," in *Judgment under uncertainty: heuristics and biases, ed. D. Kahnemann, P. Slovic and A. Tversky*, pp. 422-444, Cambridge University Press, 1982a.

[Fischhoff 1982b] B. Fischhoff. "For those condemned to study the past: Heuristics and biases in hindsight," in *Judgment under uncertainty: heuristics and biases, ed. D. Kahnemann, P. Slovic and A. Tversky*, pp. 335-351, Cambridge University Press, 1982b.

[Fischhoff and Whipple 1982] B. Fischhoff and C. Whipple, "Risk Assessment: Evaluating Error in Subjective Estimates," *The environmental professional*, vol. 3, pp.277-291, 1982.

[Gigerenzer 1994] G. Gigerenzer. "Why the Distinction between Single-event Probabilities and Frequencies is Important for Psychology (and vice versa)," in *Subjective Probability, ed. G. Wright and P. Ayton*, pp. 129-161, John Wiley & Sons, 1994.

[Goldberg 1986] L.R. Goldberg. "Some research on clinical judgments," in *Judgment and decision making: An interdisciplinary reader, ed. H. R. Arkes and K. R. Hammond*, pp. 335-353, Cambridge University Press, 1986.

[Hannaford et al. 1993] J. Hannaford, D.M. Hunns, M.R. Sayers, N. Wainwright and R.L. Yates. *The Sizewell B Protection System: Status Report on NII's Assessment of the Primary Protection System Software,* ACSNI(93) P10, Health & Safety Commission Advisory Committee on the Safety of Nuclear Installations, 1993.

[Henrion and Fischhoff 1986] M. Henrion and B. Fischhoff, "Assessing uncertainty in physical constants," *American Journal of Physics*, vol. 54, no. 9, pp.791-798, 1986.

[Hollnagel 1989] E. Hollnagel. *The Reliability of Expert Systems,* Ellis Horwood Limited, 1989.

[Hynes and Vanmarcke 1976] M. Hynes and E. Vanmarcke. "Reliability of embankment performance predictions," in *ASCE Engineering mechanics Division Specialty Conference,* pp. 31-33, Waterloo, Ontario, Canada, University of Waterloo Press, 1976.

[Jungermann 1986] H. Jungermann. "The two camps on rationality," in *Judgment and decision making: An interdisciplinary reader, ed. H. R. Arkes and K. R. Hammond*, pp. 627-642, Cambridge University Press, 1986.

[Kahneman and Tversky 1982a] D. Kahneman and A. Tversky. "Intuitive prediction: Biases and corrective procedures," in *Judgment under uncertainty: heuristics and biases, ed. D. Kahnemann, P. Slovic and A. Tversky*, pp. 414-421, Cambridge University Press, 1982a.

[Kahneman and Tversky 1982b] D. Kahneman and A. Tversky. "On the psychology of prediction," in *Judgment under uncertainty: heuristics and biases, ed. D. Kahnemann, P. Slovic and A. Tversky*, pp. 48-68, Cambridge University Press, 1982b.

[Kahnemann et al. 1982] D. Kahnemann, P. Slovic and A. Tversky, (Ed.). *Judgment under uncertainty: heuristics and biases,* Cambridge University Press, 1982.

[Kaplan 1992] S. Kaplan, "Expert information" versus "expert opinion": Another approach to the problem of eliciting/combining expert knowledge in PRA," *Reliability engineering and System Safety*, vol. 35, no. 1, pp.61-72, 1992.

[Keeney and von Winterfeldt 1991] R.L. Keeney and D. von Winterfeldt, "Eliciting probabilities from experts in complex technical problems," *IEEE Transactions on Engineering Management*, vol. 38, no. 3, pp.191-201, 1991.

[Leveson and Turner 1992] N.G. Leveson and C.S. Turner. *An Investigation of the Therac-25 Accidents,* Technical Report 92-108, University of California Irvine, 1992.

[Lichtenstein et al. 1982] S. Lichtenstein, B. Fischhotf and L.D. Phillips. "Calibration of probabilities: The state of the art to 1980," in *Judgment under uncertainty: heuristics and biases, ed. D. Kahnemann, P. Slovic and A. Tversky*, pp. 306-334, Cambridge University Press, 1982.

[Littlewood and Strigini 1993] B. Littlewood and L. Strigini, "Validation of Ultra-High Dependability for Software-based Systems," *CACM*, vol. 36, no. 11, pp.69-80, 1993.

[McClelland and Bolger 1994] A.G.R. McClelland and F. Bolger. "The Calibration of Subjective Probabilities: Theories and Models 1980-93," in *Subjective Probability, ed. G. Wright and P. Ayton*, pp. 453-482, John Wiley & Sons, 1994.

[Ng and Abramson 1990] K.-C. Ng and B. Abramson, "Uncertainty Management in Expert Systems," *IEEE Expert*, vol. April, pp.129-42, 1990.

[Ortiz et al. 1991] N.R. Ortiz, T.A. Wheeler, R.J. Breeding, S. Hora, M.A. Meyer and R.L. Keeney, "Use of expert judgment in NUREG-1150," *Nuclear Engineering and Design*, vol. 126, no. 3, pp.313-331, 1991.

[Reason 1990] J. Reason. *Human Error,* Cambridge University Press, 1990.

[Slovic et al. 1982] P. Slovic, B. Fischhoff and S. Lichtenstein. "Facts versus fears: Understanding perceived risk," in *Judgment under uncertainty: heuristics and biases, ed. D. Kahnemann, P. Slovic and A. Tversky*, pp. 463-489, Cambridge University Press, 1982.

[Saffiotti et al. 1992] A. Saffiotti, S. Parsons and E. Umkehrer. *A Case Study in Comparing Uncertainty Management Techniques,* TR/IRIDIA/92-28, Universite Libre de Bruxelles, 1992.

[Strigini 1994] L. Strigini, "Considerations on current research issues in software safety," *Reliability Engineering and System Safety*, vol. 43, no. 2, pp.177-188, 1994.

[Thorne 1993] M.C. Thorne, "The use of expert opinion in formulating conceptual models of underground disposal systems and the treatment of associated bias," *Reliability Engineering and System Safety*, vol. 42, pp.161, 1993.

[Tversky and Kahneman 1982] A. Tversky and D. Kahneman. "Belief in the law of small numbers," in *Judgment under uncertainty: heuristics and biases, ed. D. Kahnemann, P. Slovic and A. Tversky*, pp. 23-31, Cambridge University Press, 1982.

[Van Steen and Cooke 1989] J.F.J. Van Steen and R.M. Cooke. "Expert opinions as data source: methods and experiences," in *Reliability Data Collection and Use in Risk and Availability Assessment. Proceedings of the 6th EuReDatA Conference,* pp. 262-285, Siena, Italy, 1989.

[Wright and Cai 1994] D. Wright and K.-Y. Cai. *Representing Uncertainty for Safety Critical Systems,* Technical report T/002, SHIP Project, 1994.