Formalism and judgement in assurance cases

Lorenzo Strigini
Centre for Software Reliability, City University
Northampton Square, London EC1V OHB, UK
Tel: +44 (0)20 7040 8245 Fax: +44 (0)20 7040 8585
E-mail: L.Strigini@csr.city.ac.uk

Abstract

This position paper deals with the tension between the desire for sound and auditable assurance cases and the current ubiquitous reliance on expert judgement. I believe that the use of expert judgement, though inevitable, needs to be much more cautious and disciplined than it usually is.

The idea of assurance "cases" owes its appeal to an awareness that all too often critical decisions are made in ways that are difficult to justify or even to explain, leaving the doubt (for the decision makers as well as other interested parties) that the decision may be unsound. By building a well-structured "case" we would wish to allow proper scrutiny of the evidence and assumptions used, and of the arguments that link them to support a decision. An obstacle to achieving this goal is the important role that expert judgement plays in much decision making. The purpose of an assurance case is to a large extent to redirect dependence on judgement to issues on which we can trust this judgement; I doubt that this is done effectively in current practice. Making arguments explicit and if possible mathematically formal is one of the defences, yet formalism does not solve all problems and even creates some of its own. I believe that further progress must depend on better use of the knowledge produced by other disciplines about the cognitive, cultural and organizational factors that affect the production and use of assurance cases, and on studying the successes and failures of assurance cases.

1. "Cases" and judgement

Building a sound, well-structured "case" that collects facts about a system and arguments linking them to support decisions is an attractive idea. We would want decision makers to be able to see clearly the structure of arguments - whether built by themselves or by others - check that the evidence is used properly, the assumptions are acceptable, the reasoning is consistent, so that the claims made can then be seen to deserve sufficient confidence to drive decisions. In everyday practice, though, the various parts of a case tend not to be equally sound and well-structured. The part that is most often laid out clearly and rigorously is the raw evidence - e.g., results of tests and verification procedures, and implications that can be derived from them by straightforward algorithms. One could say that this is – in principle – the easy part, requiring only commitment and an adequate amount of effort. The arguments present greater difficulties. Some possible arguments are simple - e.g., straight statistical inference from realistic testing can be a textbook exercise - but they are often insufficient: they do not lend enough confidence in the claim on which a decisions should be based. A (somewhat extreme) example is given by applications with "ultra-high" dependability requirements: each kind of evidence available is usually insufficient to prove that the system is acceptable, even if no evidence points at it *not* having the required level of dependability attributes [6]. Statistical inference from test operation does not give sufficient confidence (within feasible testing duration), proofs do not cover certain aspects of systems, etc. A decision maker, e.g. a safety regulator, needs to consider these kinds of evidence and many others – often an unmanageable amount of evidence, in fact – and produce a simple decision: whether the system is safe enough to operate.

A sound case to help this regulator would detail what can be said about, for instance: how much confidence in a system being "safe enough" (as defined for the system and application in question) can we expect based on the production process used? How much does a certain period of successful test operation add to that confidence? How do the uncertainties about the future environment of the system – profile of use and threats – weaken it? And what about scenarios that have been assumed to have negligible probabilities?

All these are difficult questions. Quite often there is no sound scientific knowledge about the relationship between these disparate components of the evidence and the property that we are interested in assessing.

These difficulties are commonly solved by invoking "expert judgement", i.e., roughly, some subclaim (or the main claim) will be accepted without a detailed argument to support it, but on the basis of trust in the person (or consensus of persons) stating it. The role of judgement ranges from estimating individual variables – like the probability of a certain failure mode – to drawing conclusions from the whole complex collection of disparate evidence and partial claims. In the former case, the expert's opinion is substituted in place of empirical evidence that is not available. In the latter, it fills in for missing scientific knowledge and/or complex arguments. The experts' opinions are trusted even though their bases cannot be examined and audited: the rules the experts have applied, and their basis in reasoning and past experience, may not be clear even to the experts themselves, and anyway they are not spelled out. Indeed, if they were spelled out in full, the case would rest on these explicit rules and arguments (which other experts can examine and discuss), without a need to appeal to "judgement".

To avoid confusion, I must say that "judgement" is never absent from the acceptance of an argument, however scientifically rigorous. Scientific (or engineering) rigour largely consist in building an argument so that the "hard" parts – those that appear highly vulnerable to errors and misunderstandings – are argued explicitly and clearly. Direct reliance on "judgement" without explicit arguments is not eliminated, but it is relegated to less hard issues. For instance, we would not require an explicit argument in order to trust our manual verification of very simple logical or mathematical statements used in an argument. Accepting our direct judgement on these points prevents an infinite recursion in the building of a case. There will also be grey areas: statements for which some will think judgement is sufficient, and some will require more explicit arguments. However, when "expert judgement" or consensus is explicitly invoked in discussing system dependability, it is often, in my experience, to support statements about "hard" issues.

Even on "hard" issues, reliance on expert judgement is often necessary. Yet, the respect with which it is treated is not always supported by evidence of its accuracy. Actually, there is abundant anecdotal and scientific evidence of inaccurate expert judgement. Apart from the dangers of emotion and conflicts of interest, experimental psychologists have documented many types of failures of the heuristics applied by human minds to problems involving uncertainty and probability[1], as problems of assurance usually do. Even people trained in probabilistic reasoning have been shown to make serious mistakes if they tried to use "judgement" without applying explicit, formal probabilistic reasoning. It is not the case that experts are necessarily bad at dealing with uncertainty: some experts, and some categories of experts, have been shown to be routinely quite good. The problem is lack of evidence that we can trust expert judgement in general and a priori. Thus we cannot trust a priori a specific judgement by a specific expert, unless we have some explicit and convincing argument for doing so. Such an argument could be for instance that the expert has both the means for forming an accurate judgement (for instance, that his/her experience is pertinent to the problem at hand, and sufficiently extensive to support the kind of claim the expert is called to make) and a record of accurate judgement in similar circumstances. Unfortunately, such arguments are not commonly included in assurance cases, and I believe that in many important applications they would not stand. The experts' judgements may well be very accurate, but we would lack evidence for believing them to be so; just as very dependable systems are built for which – due to how the development is managed and documented – we could not build a convincing argument that they *will* be so dependable, before they turn out to be so in operation.

Luckily, researchers have also studied how circumstances, and especially the way tasks are set, affect the accuracy of judgements. For instance, it appears that people are much better at judging probabilities in settings that require them to count events than in settings that require reasoning directly in the more abstract language of probabilistic calculus. Some findings concern cognitive problems in individual judgements, some the formation of consensus, and so on. I believe that more use could be made of this kind of knowledge, and the community building "assurance cases" needs better contacts with the psychologists studying these phenomena.

2. Formalizing judgement

A possible defense, in the spirit of building assurance cases, is simply to reduce reliance on judgement by substituting instances of intuitive judgement with formal, explicit, auditable arguments. For drawing conclusions from disparate and largely "soft" evidence, as in the above example of the safety regulator, a way forward seems to be in giving the experts a way of describing the sound argument that their mental processes emulate or should emulate. If the description is formal and mathematical, it will require clear statements of its assumptions, the evidence and the general laws invoked to support a claim. Therefore, both the author and independent auditors will be better able to audit the strengths and weaknesses of the argument.

At CSR and with other colleagues, we have been studying for this purpose the use of the formalism of "Bayesian belief networks" [2, 3, 4, 5, 7]. These have appealing features recommending them for this task: their graphical appearance makes it reasonably easy to describe

[1] There is an abundant and growing body of research about these issues, with various accessible books on judgement under uncertainty and expert judgement in particular. I tried to summarise its implications from a dependability viewpoint [8] during the EU SHIP project (http://www.adelard.co.uk/research/ship.htm). The body of results has grown since, and I believe it has if anything reinforced the conclusion that expert judgement is now treated with insufficient caution.

even complex probabilistic models, they have a clear formal meaning, so that one can examine them for inconsistencies or factual fallacies, and they support probabilistic calculus for deduction and statistical inference (so that new evidence can be incorporated into an existing case) alike. In several projects we looked at the potential of BBNs for turning expert judgement from an obscure piece of magic into a logical process that the expert can describe in explicit terms, so as to be able to check and criticize it, and to communicate it to others for their verification and criticism.

I believe – on the limited basis of my own experience – that BBNs are indeed useful for these purposes. As is common experience with mathematical formalisms, being given a practical language for describing vague mental processes not only allows us to describe these processes, but encourages us to query and improve them. In the end, a sound argument is actually created through the process of describing it formally. I found that (small) BBNs would help me and others to clarify and debug probabilistic arguments.

Studying BBNs, however, led us to better recognize some of the practical problems that are not solved by more powerful mathematics or better notations. I believe that these problems include:

- limited comprehension: notations that appear easy to grasp intuitively may create false perceptions of their “true” formal meaning. This creates problems with the elicitation and validation of the BBN;
- excessive expectations: in part because of comprehension problems, people seem to tend to build more complex models than they can trust themselves to validate and often even to understand;
- undue reliance on results: even results that are known to be untrustworthy will affect decision-making.

All three problems are related to cognitive (and emotional), group and organizational phenomena, not directly attacked by conventional mathematical or physical-science research. About the first problem, my colleagues have been studying ways of showing experts alternative viewpoints of the models they create, to clarify what these models actually mean and imply, and whether this is consistent with the expert’s beliefs [2, 9]. Detecting an inconsistency would show the expert a need for changing the formal model or questioning the expert’s own informal beliefs. It is likely that tools with similar functions would be useful for many of the mathematical models we routinely use in dependability assessment, like complex fault trees or Markov chains. In all cases, relatively little attention has been paid so far to this aspect of the “ergonomics” of mathematical tools: how to help users not only to build and solve models, but also to comprehend and debug them. The enemy here is often complexity: although the mathematical formalism guarantees consistency, we may still be unable to grasp the meaning of the model we build in its entirety. For instance, when subdividing a complex model into more manageable and reasonably separate submodels, we may involuntarily introduce extra assumptions that alter the meaning of the model; or parameters that are in common between the submodels may end up being used in subtly inconsistent ways in the various submodels

New tools may enable us to comprehend somewhat more complex models, or better to notice when we have exceeded the levels of complexity that we can manage. On the other hand, other problems will remain, including unwillingness to heed the warnings that we may obtain from the tools.

3. The problems outside the mathematics

The techniques just mentioned try to reduce the risk of flaws in arguments by dealing with the technical and scientific difficulties in building cases, which would cause such flaws. But other factors that matter include those that affect the choice of the form of arguments to be pursued – since some forms will be more error-prone than others - and the likelihood that flawed arguments will be accepted and used.

Part of the problem seems to be a tendency to build over-complex cases, in order to include all the disparate evidence available about a system. It seems that, given a wealth of evidence about a system, we have a hard time accepting that the case we can actually build upon it may be no stronger that what we would have if we only had a subset of this evidence. Many kinds of evidence are clearly relevant, but we don’t know how much they should actually weigh. Argument that appeal to conformance to the various requirements of a standards (like IEC 61508) are an example. When an organization has spent much effort to collect evidence about a system, it will quite naturally want a return (some contribution to the claims about the system) on its investment. This unfortunately may push to make demands on the argument-builders to go beyond what is feasible. A possible way forward here, apart from empirical research to improve the ability to link observed evidence to claims, is to use subsets of the evidence to build a set of parallel and separate arguments [1]: the set of arguments may not allow stronger claims than each argument individually, but it will pay off in the form of protection against the consequences of hidden flaws in each individual one, and thus increase *confidence* in these claims.

Another problem is - I believe – the extent to which weak-based arguments are accepted, and used in decision making. Here, multiple factors are relevant. The argument-builders themselves are handling difficult problems to start with. The dependence on judgement reduces the chances for discovering errors while building the argument; it also probably reinforces the gap between technical experts and managers, by giving “official” status to the existence of forms of esoteric knowledge of the technical experts, that others cannot share or challenge. And I think we must not neglect the effects of even weakly based opinions on the development of consensus in situations of uncertainty. They may deepen an “uncertainty trough”, making decision makers more confident in analyses and predictions than the technical experts who originated them. And it is plausible that these opinions will produce a psychological “anchor”, an initial opinion that will be revised less rapidly and radically than it should be if new experience shows the revision to be necessary, and a natural “landmark” in the negotiations that surround an assurance case.

In recommending ways of building assurance cases, we have too limited an understanding of this tangle of technical, psychological and organizational factors.

4. Conclusions

This document is based on my experience in research concerning safety critical systems, but the difficulties and partial solutions seem common to many critical decisions in designing or accepting complex computer and computer-based systems.

The net conclusion is that there seem to be many ways of improving the way judgement is used. Explicitly building assurance cases certainly seems to be a step forward, but the communities building assurance cases can benefit from better communication both between themselves and with other disciplines, like psychology.

As in all engineering, the engineering of assurance cases must be a matter of tradeoffs. The push to use as much as possible of the available evidence to allow more optimistic claims leads to more complex arguments and/or more reliance on expert judgement, both factors that tend in turn to make the assurance case less trustworthy. The push towards more explicit and formal arguments, that help to manage complexity and reduce reliance on unaided judgement, may lead both to diminishing returns and to new problems. Studying – empirically – how these tradeoffs work out in the various application areas and organizational cultures would help much to assess the state of the art and identify factors of success and failure. We need more research focusing on assurance cases in themselves: how they are built, how often they prove accurate, what mistakes are made that matter, and – finally – which methods seem effective in preventing these mistakes.

Acknowledgment

This work was supported in part by the U.K. Engineering and Physical Sciences Research Council via the Interdisciplinary Collaboration on the Dependability of computer based systems, “DIRC”.

References

[1] R. Bloomfield and B. Littlewood, "Multi-legged arguments: the impact of diversity upon confidence in dependability arguments", in Proc. DSN 2003, International Conference on Dependable Systems and Networks, San Francisco, U.S.A., 2003, pp. 25-34.

[2] P.-J. Courtois, B. Littlewood, L. Strigini, D. Wright, N. Fenton and M. Neil, "Bayesian Belief Networks for Safety Assessment of Computer-based Systems", in E. Gelenbe (Ed.) "System Performance Evaluation: Methodologies and Applications", CRC Press, 2000, pp. 349-363.

[3] K. A. Delic, F. Mazzanti and L. Strigini, "Formalising a software safety case via belief networks", SHIP Project Technical Report SHIP/T/046, July, 1995.

[4] K. A. Delic, F. Mazzanti and L. Strigini, "Formalising Engineering Judgement on Software Dependability via Belief Networks", in Proc. DCCA-6, Sixth IFIP International Working Conference on Dependable Computing for Critical Applications, "Can We Rely on Computers?", Garmisch-Partenkirchen, Germany, 1997, pp. 291-305.

[5] N. E. Fenton, B. Littlewood, M. Neil, L. Strigini, D. R. Wright and P.-J. Courtois, "Bayesian Belief Network Model for the Safety Assessment of Nuclear Computer-based Systems", DeVa TR No. 52 , 1997.

[6] B. Littlewood and L. Strigini, "Validation of Ultra-High Dependability for Software-based Systems", Communications of the ACM, 36, pp. 69-80, 1993.

[7] B. Littlewood, L. Strigini, D. Wright and P.-J. Courtois, "Examination of Bayesian Belief Network for Safety Assessment of Nuclear Computer-based Systems", DeVa project Technical Report No. 70, 1998.

[8] L. Strigini, "Engineering judgement in reliability and safety and its limits: what can we learn from research in psychology?", SHIP Project Technical Report SHIP/T/030, July, 1994.

[9] D. Wright, "Elicitation and Validation of Graphical Dependability Models", in Proc. 22nd International Conference on Computer Safety, Reliability and Security, Edinburgh, United Kingdom, 2003, pp. 8-21.