Confidence: its role in dependability cases for risk assessment

Robin E Bloomfield, Bev Littlewood, David Wright
Centre for Software Reliability, City University, London
{reb, bl, dw}@csr.city.ac.uk

Abstract

Society is increasingly requiring quantitative assessment of risk and associated dependability cases. Informally, a dependability case comprises some reasoning, based on assumptions and evidence, that supports a dependability claim at a particular level of confidence. In this paper we argue that a quantitative assessment of claim confidence is necessary for proper assessment of risk. We discuss the way in which confidence depends upon uncertainty about the underpinnings of the dependability case (truth of assumptions, correctness of reasoning, strength of evidence), and propose that probability is the appropriate measure of uncertainty. We discuss some of the obstacles to quantitative assessment of confidence (issues of composability of subsystem claims; of the multi-dimensional, multi-attribute nature of dependability claims; of the difficult role played by dependence between different kinds of evidence, assumptions, etc). We show that, even in simple cases, the confidence in a claim arising from a dependability case can be surprisingly low.

1. Introduction: uncertainty, confidence

Risks associated with the use of computer-based systems are becoming increasingly important to society. Whilst the problems have been recognized for a long time in safety-critical industries, there is a new awareness in other industries, such as banking (see, e.g., the Basel II accords [6]). Assessing these risks, so that intelligent decisions can be made – e.g. about deployment, or about the cost-effectiveness of possible risk reduction procedures – is hard. Much of the difficulty stems from the fact that the fallibility of *software* plays such an important role as a source of risk.

Risk involves notions of *failure* and *consequence* of failure. Its assessment therefore requires an assessment of dependability; this might be expressed, for example, as probability of failure upon demand, rate of occurrence of failures, probability of mission failure, and so on. In this paper we shall address this dependability assessment problem only, and not further discuss the cost/consequence part of risk assessment.

There is now a huge literature on the assessment of the dependability of software-based systems, going back several decades. In recent years the assessment process has started to be formalized in *dependability cases*, most notably safety cases. A safety case has been defined as:

> *A documented body of evidence that provides a convincing and valid argument that a system is adequately safe for a given application in a given environment [7].*

In this paper we shall discuss the important role played by *uncertainty* in dependability cases. We believe that some aspects of uncertainty have been long neglected and we propose a formal quantitative treatment of 'confidence' to address this omission.

Computer scientists have long had an uneasy relationship with uncertainty, and with its most powerful calculus, probability. One of us can remember discussions of thirty years ago about software reliability. It was difficult then to persuade some software experts that there was inherent uncertainty in the failure processes of programs, and that *probability* was the appropriate way of capturing this uncertainty. Instead, it was asserted that software failed *systematically*, and thus that notions of 'reliability' were meaningless.

Over the years the position has changed. It is now widely agree that 'systematic failure' just means that a program that has failed in certain circumstances will *always* fail whenever those circumstances are exactly repeated. The uncertainty lies in our not knowing beforehand *which* circumstances (e.g. inputs to a program) will cause failure, and *when* these will arise during the operational execution of the program. It is this uncertainty that is represented in a probabilistic measure of dependability, such as reliability.[1]

[1] It is interesting that a similar reluctance to acknowledge uncertainty has occurred recently as attempts have been made to model *security* probabilistically. Here, notions of attacker

The uncertainty discussed above concerns system behaviour – it is 'uncertainty-in-the-world'. There is another form of uncertainty that has, we believe, been neglected: this is uncertainty in the dependability assessment process itself.

Consider, for example, a situation in which we want to claim that a software component has a probability of failure on demand (*pfd*) smaller than 10^{-3} (a figure that may have arisen from the requirements of a wider system safety case). Our evidence to support such a claim may be testing data, different types of static analysis, etc (it is a characteristic of dependability assessment, particularly for software-based systems, that the supporting evidence is usual disparate in nature).

The problem is that such evidence will never allow us to be *certain* that the claim is true: there is inherent uncertainty here. If we collect more supportive evidence, we might reasonably expect to increase our confidence in the truth of the claim, but it will rarely be possible to collect sufficient evidence to eliminate doubt completely.[2] As the cases often deal with critical systems there needs to be high confidence in the resulting judgement. Often there are areas where there is lack of understanding (e.g. due to deficiencies in the science, in experimental data) or there are problems that the engineering process has not been rigorously followed. There may be problems with uncertainties in the prediction due to the high level of human-computer interaction (e.g. in a cockpit) making quantified estimates of reliability problematic due, among other things, to the sensitivity to the exact context and variability of performance. Not that we argue against the use of humans, indeed it is often this inherent variability in human performance that allows systems to recover from unsafe situations, turning potential accidents into incidents or near misses. Rather that all these factors lead to uncertainties in our judgement.

This prompts questions such as: *How* confident are we that the claim is true? How do we express 'confidence' quantitatively? What effect does this 'assessment uncertainty' have upon decision-making?

Confidence in dependability cases stems from a multiplicity of judgements, some informal, some very formal, some from individuals and others from groups of experts. Confidence, like proof, is the product of a social process.

The greatest difficulty is to deal with the uncertainty that arises from weaknesses in the argument that supports a dependability claim. These might arise, for example, from uncertainty as to whether underlying assumptions are true. Although we are not aware of any formal treatments of this kind of uncertainty in the literature, the problem has been acknowledged implicitly. For example, in [9,10] it is recognized that there will be uncertainty arising from a single argument leg supporting a dependability claim, so that a second, different, leg should be added – a kind of 'argument fault-tolerance'. However, the reasoning here is informal and qualitative – there is no guidance about *how much* benefit will ensue, nor about *how* confident one would be in a claim after following this procedure. The recent reissue of the UK Defence Standards recognises the role of confidence and an earlier version of this paper provided some rationale behind the guidance in Part 2 [8]. Two of us were involved in a study of the use of computers in the UK nuclear industry [11]. One recommendation of this group was that the principle underlying much regulation in this area, ALARP (that the failure rate stated by a dependability claim should be As Low As Reasonably Practicable), ought to be accompanied by another principle, ACARP (that one should be As Confident As Reasonably Practicable in the truth of a claim). The recommendation has not yet been adopted.

In the rest of this paper we look at some of these problems in more detail, and provide some very tentative pointers to ways forward.

2. Judging the range of probability of failure

As an example of the interplay between confidence and failure rate (or *pfd*) we examine the judgement that a system has a certain classification safety integrity level (SIL): a measure of how safety critical a function is. While we have chosen the standard IEC 61508 [4] for this illustration the use of levels and the judgement of membership of "levels" is a pervasive issue (e.g. in aerospace [2], defence, security, nuclear, railways). IEC 61508 is a generic standard for the functional safety of computer-based systems and it defines safety integrity both in terms of a probability of dangerous failure on demand and the probability of dangerous failure per hour. SILs are defined by a range. For example a Safety Integrity Level n safety function with a low demand mode of operation has an average probability of failure to perform its design function on demand in the range $10^{-(n+1)}$ to 10^{-n}. In practice SILs are used in a variety of ways, not only to describe the judged probability of failure (whether qualitative or quantitative), but also to indicate confidence in the judgement being made. There is an interesting interplay between level and confidence: people seem to expect the higher SILs to be demonstrated to higher confidence.

---

intentionality have been used to claim that probabilistic measures of security are not appropriate.

[2] One exception might be exhaustive testing in some specialized situations. Such exceptions are, we believe, very rare.

Although we use judgement of SIL as an example, the issues this raises apply to many of the judgements made in safety and dependability cases. We should point out that while SIL applies to one important attribute of a safety critical system there are others such as robustness, security and maintainability that should be addressed in a full safety case.

Throughout the paper we shall interpret probabilities in the Bayesian sense of 'degrees of belief'. This has the practical advantage of providing a formalism that allows uncertainty to be treated in a consistent way, whether evidence comes from empirical data or the judgement of a human expert.

3 Modelling judgement of SIL

Deriving a SIL can be done in a number of interrelated ways. For example:

- Relying entirely on qualitative arguments to directly assess a SIL: the failure rate or SIL is not quantified and it may be denied that software reliability can be quantified at all.
- Using expert judgment based on standards compliance to assess the system. This approach suffers from lack of validation, calibration, and many influencing parameters some of which are ignored in the standard (e.g. size of the software).
- Using a best fit reliability growth model, assessing the accuracy of predictions, adding a margin for subjective assessment of assumption violation.
- Using a worst-case model of the failure process, taking into account uncertainty in parameters quantitatively, and using a subjective estimate of invalid model assumptions.
- Developing an argument of high confidence in zero defects. This may be credible for small highly analysed systems or hardware logic but is not developed further here.

What distinguishes these methods is the confidence that can be placed on the judged SIL but in every case there is some uncertainty that needs to be assessed. This essential uncertainty arises from a number of sources:

- uncertainty in the completeness and validity of data
- doubts about assumptions in the model and model validity
- doubts about the implementation of the model (this may be based on complex software itself)
- doubts about the application of the model and how to deal with conservatism in the model and the data

Confidence in SIL n can be expressed as the probability that the judged *pfd* (λ) is within the upper bound of the *pfd* for that SIL band:

$$P(\lambda < 10^{-n})$$

If the failure rate was normally distributed or symmetrical in linear space, changing the confidence in it by narrowing the distribution would not affect the mean value. However none of the examples would necessarily have a normal distribution and we think normality is unlikely to be the case for a number of reasons:

- The distribution cannot be negative yet we can have a large uncertainty in our judgement of a small *pfd*.
- We would expect the distribution's density function to tend to zero as the rate $\lambda \rightarrow 0$, although there may be special cases where there is belief in possible perfection of the system.[3]

We have undertaken some experimental work on how experts make judgements of the probability of failure on demand. The results are summarised in Section 3.3. We also note that in reactor safety studies log-normality is often chosen for model parameters (see the discussion in [3]). So in the analysis that follows we deploy a log normal distribution but the thrust of the results only require a non-symmetric distribution. We have repeated some of the results for a gamma distribution to illustrate the (low) sensitivity to the log-normal assumptions.

3.1 A log normal model

If we model our judgement of the dangerous failure rate or *pfd* with a log normal distribution, the mean will be

[3] At first glance it appears contradictory to allow the pfd to take the value 0, and at the same time assign almost zero probability to a very small non-zero value, say 10^{-10}. In fact, the reasoning to support such values would be very different. In the first case, the claim is one of perfection, and this might be supportable by non-probabilistic reasoning. In the second case, it is assumed that the system is imperfect, but it is claimed that the impact of its faults is vanishingly small.

different from the most likely "peak" value (the mode). The log-normal distribution is specified by two parameters: μ which controls the peak value and σ which controls the spread. Although the log-normal is generally difficult analytically, the mean and the mode can be calculated as:

$$a\_mean(\mu,\sigma) := \exp\left(\mu + \frac{1}{2}\cdot\sigma^2\right) \qquad log\_mean(\mu,\sigma) := \mu + \frac{1}{2}\cdot\sigma^2$$

and the peak value is at

$$a\_mode(\mu,\sigma) := \exp\left(\mu - \sigma^2\right) \qquad log\_mode(\mu,\sigma) := \mu - \sigma^2$$

where:

log_mean is the log of the mean failure rate
log_mode is the log of the peak failure rate

The probability density function is:

$$pdf\_\lambda l(\lambda, lmean, lmode) := \frac{1}{\sqrt{2\cdot\pi}\cdot\sqrt{\frac{2\cdot(lmean - lmode)}{3}}\cdot\lambda}\cdot\exp\left(\frac{-1}{2}\cdot\frac{\ln(\lambda) - \frac{2\cdot lmean + lmode}{3}}{\frac{2\cdot(lmean - lmode)}{3}}\right)^2$$

It can be shown that the ratio of the mean and the mode is: $\log_{10}$ (mean / mode)= 0.65 σ^2. So the difference between the mode and mean varies with the spread of the distribution (σ). As you might expect there is no difference when there is no spread ($\sigma = 0$), but for a broad spread there is a surprisingly big difference, e.g. the mean failure rate is one decade greater than the mode if $\sigma = 1.2$, and two decades greater if $\sigma = 1.7$: See examples in Figure 1. All judgements estimate the most likely failure rate to be 0.003 (i.e. in the middle of SIL2 range of 10^{-3} to 10^{-2} as defined in Table 1) but with varying degrees of confidence. The mean of the dashed curve is 0.004, which is quite close to the mode value of 0.003. By contrast, the solid curve has the widest spread and the mean is 0.01 putting the mean value in the SIL1 band rather than the SIL2 band.

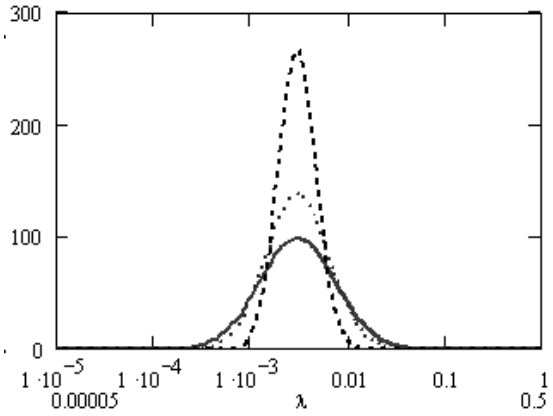


Figure 1: Density functions of the judgement of SIL

The impact of higher failure rates can be seen from plotting the probability density functions on a linear scale:

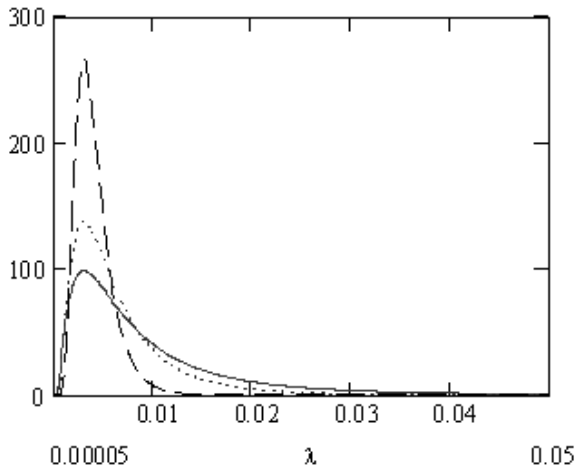


Figure 2: Log normal distribution functions on a linear scale

3.2 Variation of mean with confidence in SIL membership

We can calculate from the probability distributions how the mean SIL varies with the spread in the distribution. One measure of the spread in the distribution is to calculate the probability that we judge the system to be in the desired SIL or better. This is our one sided confidence in the system's SIL membership:

$$\text{confidence\_better\_x}(\text{bound}) := \int_0^{\text{bound}} \text{pdf\_}\lambda\text{l}(0, \lambda, -4.6, \ln(0.003))\, d\lambda$$

We model this by keeping the mode of the distribution constant. This is shown in Figure 3 where the mode has been kept at 0.003 (the middle of SIL2) and we see that if our confidence falls below about 67% that the system is SIL2 then the mean rate is actually in the SIL1 band.

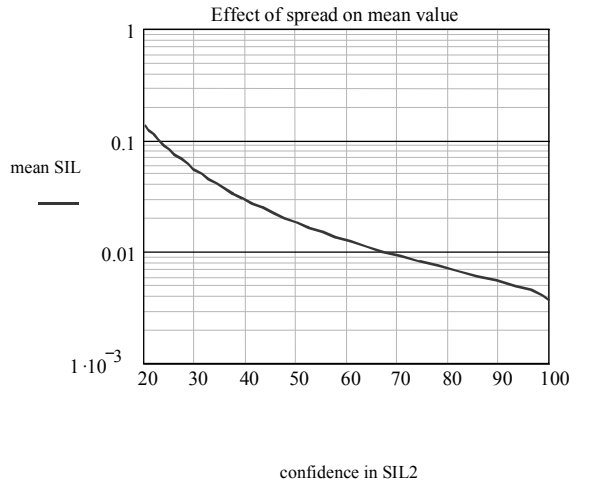


Figure 3: Relationship between confidence in a SIL and the mean value

Another of way of looking at the problem is, for a given mode and actual mean, to calculate the chances of the true system failure rate being in the different SIL bands. This is shown in Figure 4.

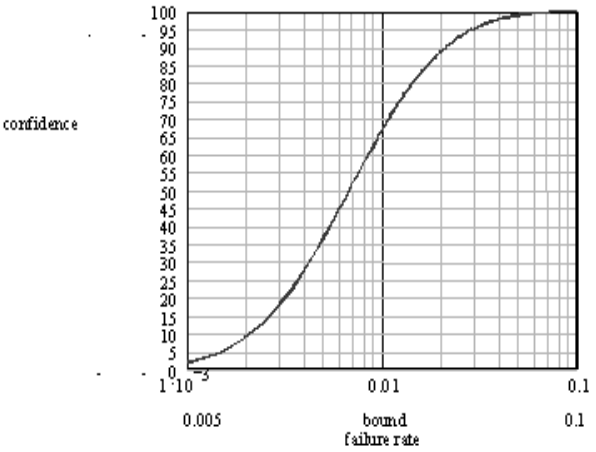


Figure 4: Confidence mean failure rate better than a bound

So for the widest distribution (corresponding to the solid line in Figure 1), the system has about a 67% chance of being in SIL2 or higher and a 99.9% chance of being SIL1 or higher.

3.3 Experimental results

We conducted an experiment with 12 experts from a variety of European countries and backgrounds. All were familiar with safety rated systems and some were experts with many years experience of the development and assessment of such systems. The experts were asked for judgments in four phases

1. After a 20 minute presentation describing a safety critical system and the implementation of a particular safety function. This was based on the Public Domain Case Study of the European nuclear R&D project Cemsis [5].
2. After a request for additional information, which (if available) was provided individually
3. After a group presentation of all items of additional information provided individually to the different participants in the previous phase;
4. After a Delphi phase where there was an opportunity to discuss decisions with the other participants;

Interestingly the assessors seem to fall into two groups: a minority of (3) doubters who expressed these doubts by giving the system a very high failure rate and another group that expressed their beliefs as shown below:

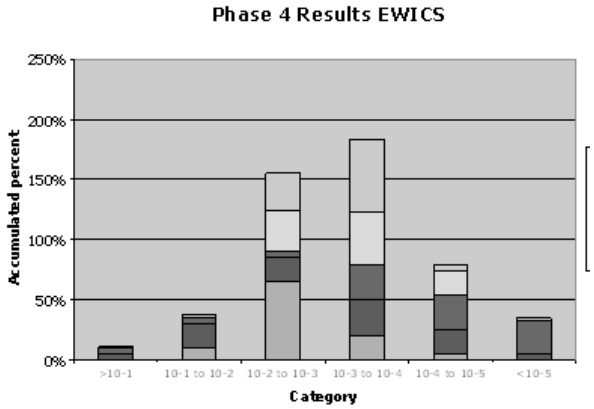


Figure 5 Experimental results

The group were about 90% confident that the system was in SIL2 or better yet the resulting *pfd* (0.01) is on the 2-1 boundary. However the main point of the experiment, as far this exercise is concerned, is to add plausibility to the use of an asymmetric distribution.

3.4 Assessment heuristics and reducing the claim

These modelling results would seem to confirm the heuristic used by safety assessors that although the evidence points to say a SIL2 system they consider it SIL1 because of the uncertainties.

It is therefore more likely that a better case can be made if the system is judged as most likely a SIL n+1 system and it could then be taken as a SIL n with high confidence. This can be seen in the safety justification for the Sizewell B Primary Protection System where doubts about the quality of the development process of the software led to an order of magnitude reduction in the judged probability of failure on demand (some background is provided in [1]). The situation is likely to be exacerbated when qualitative expert judgements are made on the SIL. It may be that the type of standards compliance argument that is often attempted should really lead to a greater than 1 reduction in the claimed SIL. This is discussed further in Section 3.2.

Because the judgement of the SIL is likely to be a combination of several sub-judgements, it would be useful to understand how the confidence in these contributing arguments can be combined. This is the subject of some on-going work.

It is unlikely that we will have precise estimates of the confidence of experts. To cope with this we have developed a conservative, worst case, way of distributing our doubt about the system. Consider the simple situation where the dependability case just has to support a claim about the *pfd* of a system.

We treat *pfd* as a random variable, with probability density function $f(p)$. This can be regarded as the (Bayesian subjective) belief of an expert that takes account of all his uncertainty, including assumption doubt.

In such a situation, for a randomly selected demand the expert's belief is:

$$P(\text{system fails on randomly selected demand}) = \int_0^1 \text{pf(p)dp} \qquad (4)$$

In fact it is well known, as we have stated earlier, that it is hard to elicit the beliefs of an expert in the form of a complete distribution like this. Indeed, some would argue that describing this as elicitation begs the question that the expert really does 'have' a complete distribution to be elicited. Rather he may only be prepared to express a belief of the kind $P(pfd < y) = 1 - x$. If this is all he is prepared to say, it is reasonable to ask what is the worst case $f(p)$ – i.e. the distribution that gives the most conservative result in (4).

Figure 6a shows a typical 'real' distribution, $f(p)$ that satisfies the expert's belief: $P(pfd < y) = 1 - x$. Of all the many such distributions that satisfy his belief, Figure 6b shows the most conservative: here all the probability mass for the interval $(0,y)$ is concentrated at y, and all the mass for the interval $(y,1)$ is concentrated at 1.

Figure 6a

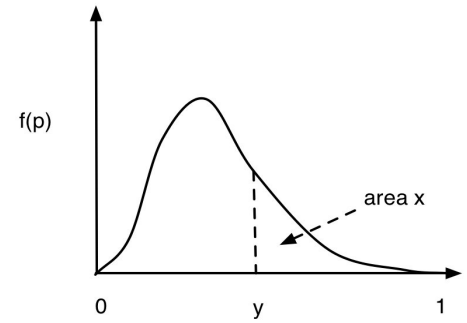
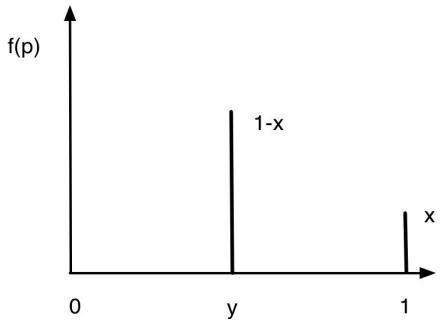


Figure 6b



It is easy to see that the maximum value of the probability of failure for a randomly selected demand occurs when $f(p)$ takes the form in Figure 6b. In other words, we have:

$$P(\text{system fails on randomly selected demand}) < (1-x)y + x = x + y - xy \qquad (5)$$

This should be interpreted as the expert being *certain* that the probability of failure on a randomly selected demand is smaller than $x+y-xy$.

The inequality, (5), can be used to give some support to the kind of informal reasoning that goes as follows: I need to claim that the *pfd* is less than 10^{-3}, but with my present dependability case I still have a small doubt that the *pfd* may be greater than this; so I strengthen my case (e.g. collect more supportive evidence) to make with high confidence the *stronger* claim that the *pfd* is smaller than 10^{-4}. This is the kind of reasoning we have seen applied in real safety cases.

The details of how to proceed are as follows. Suppose that the requirement imposed by a wider dependability case is that the *pfd* for the system is no greater than y (in our example, 10^{-3}). We thus wish this claim to be true. Suppose further that the expert believes sufficiently strongly, say with confidence (1-x^*), that the *pfd* is no greater than y^* ($<y$). Then, if $x^*+y^*-x^*y^*<y$, it follows from the result above that the expert believes the probability of failure on a randomly selected demand is less than y.

The point here is that the confidence (or doubt) about the *pfd* has been turned into a probability of the occurrence of an *event* (failure of the system on a randomly selected demand). This thus relates directly to the dependability requirement placed upon the system by the wider safety case.

It is instructive to consider some examples of (x^*,y^*) pairs, representing the expert's beliefs about the *pfd*, when $y=10^{-3}=x^*+y^*-x^*y^*$.

Example 1 At one extreme, we have $x^*=0$, $y^*=10^{-3}$. This is simply the expert believing *directly* that he is certain that the *pfd* is smaller than 10^{-3}, i.e. that his beliefs, represented by the probability density function $f(p)$ are such that there is zero probability mass to the right of 10^{-3}.

Example 2 At the other extreme, we have $x^*=10^{-3}$, $y^*=0$. This case represents the expert believing with 99.9% confidence that the system is 'perfect', i.e. has zero *pfd*. In the event that it is *not* perfect, the worst that can happen is that it is *certain* to fail – so for a randomly selected demand, there is a 10^{-3} chance of failure (i.e. that the system is not perfect).

Example 3 A more interesting example gives some supporting formalism to the way of proceeding that we have seen used, with informal justification, in real safety cases. The reasoning is as follows. The expert constructs an argument that allows him to have high confidence that the *pfd* is a whole decade better than the goal of 10^{-3}, i.e. he claims high confidence in the *pfd* being better than 10^{-4}. That is, $y^*=10^{-4}$. His (informal) reasoning is that if he believes strongly that the *pfd* is smaller than 10^{-4}, then he can be 'effectively certain' that it is smaller than (the more modest) 10^{-3}. In fact, since $x^*+y^*-x^*y^*=10^{-3}$, it follows that $x^*=10^{-3}-10^{-4}=0.0009$ (approximately – we can ignore the x^*y^* term here). So, for this reasoning to apply, he needs to have an argument sufficiently strong to be able to claim the *pfd* is smaller than 10^{-4} with confidence 99.91%.

More generally, if the expert wishes to claim that the probability of failure on a randomly selected demand is better than y, he needs to be able to claim with confidence 1-x^* that the *pfd* is smaller than y^*, where $x^*+y^*-x^*y^*=y$.

This last example shows how unforgiving this kind of reasoning can be. The bounds given above in (4) are conservative. The expert needs to have an argument that is sufficiently strong that his 'single point' elicited belief, (x^*,y^*) has the property that *both* x^* (doubt) *and* y^* (claim) are smaller than the required claim, y. The coupling here between claim and doubt suggest that there would be strict limitations to the use of this kind of reasoning. Imagine, for example, that the requirement is the more stringent $y=10^{-5}$. To use this kind of argument, the expert would need to believe the *pfd* is smaller than y^* (itself smaller than y) *with a confidence greater than 99.999%.* It seems unlikely that real experts would ever express confidence of this magnitude (and if they did they would not be believed by others).

Note that, if the expert believes there is a probability p_0 that the system is 'perfect' (i.e. *pfd*=0: $f(p)$ has probability mass p_0 at the origin), the upper bound in (5) becomes $x+y-(x+p_0)y$. It is simple to modify the reasoning of examples like those above.

This example has used the worst case assignment of the doubt to "1". If we could defend other approaches, for example that we were sure we were not wrong by more than a factor of 100, then other models along the same lines are possible – but harder to defend.

4 Discussion

There are a number of strategies that can be adopted in the dependability case to address the confidence issue:

- Reducing the claimed figure due to lack of confidence. See discussion Section 2.3 and Section 3.2
- Undertaking confidence building measures
- Reducing the required confidence by additional argument "legs".

The last two are now briefly discussed.

4.1 Confidence building from experience

The other side of reducing a judged *pfd* because of lack of confidence is undertaking assurance activities explicitly to increase confidence. In view of this in [11] we proposed a sister principle to ALARP that of As Confident as Reasonably Practicable (ACARP). In practice we often undertake analysis and verification activities that increase our confidence without actually changing the system and this is especially so for software. An alternative strategy to just reducing the SIL rating to give high confidence, is to use techniques that attack the high failure rate tail of the distribution. It is this tail that is causing the reduction of the SIL from n to n-1. Operating experience or statistical testing can "cut off" this tail so the distribution gets modified by the survival probability and renormalized. Later work will describe this in more detail.

Similarly we could analyse the growth in dangerous failure rate with failures (some safety systems such as air traffic control can fail several times a year and the overall system still be safe due to the large mitigations from others systems, providence etc). Preliminary results indicate that tests rapidly increase confidence and reduce the mean.. So one approach of tackling confidence might be to give a system a provisional SIL rating based on a broad distribution reflecting the initial uncertainties, and then increase this SIL rating after an operating period. The risk analysis would have to take into account the period of greater risk. This is similar to the organisational strategies for using COTS systems initially only in non-safety-related applications.

More work is needed to model how the worst-case confidence is impacted by subsequent testing. It may well be that there is an equivalent to the conservative bound on *mtbf* [13] for confidence.

4.2 Confidence building from legs

An alternative strategy to tackling the tail of the distribution is to find an alternative way of predicting the same result and so develop another argument "leg" that the system is in the required SIL band or higher. "Multi-legged" arguments are an informal concept and we see them used to mean both confidence building where a technique (e.g. testing) attacks the tail of the first judgement, and where a separate argument is made that does not tackle the tail but reduces the required confidence in the first argument.

These issues of interplay between adding assurance legs and confidence are subtle and the subject of continuing research (see [12]).

4.3 Standards issues

IEC 61508 [4] is an important, generic seven-part safety standard that sets out a detailed approach to the development of safety related computer based systems. In some ways our analysis is at odds with IEC 61508 in as much as the standard does not accept – or is at any rate inconsistent about – the use of statistics for systematic faults, and in the standard this includes software. Despite rejecting the use of quantified reliability for software in Part 1 the standard talks in Parts 3 and 7 about statistical testing of software and discusses statistical requirements for operating experience. Furthermore the quantified SILs implicitly require the software reliability (with respect to dangerous failures) to be quantified.

The definition of SIL in terms of the probability of failure on demand or per hour is technically useful as it allows for different distributions and requires the pdf to be integrated to arrive at the mean.

The confidence required in a SIL is not explicitly addressed in the standard. Part 3 does not mention "confidence" at all. However Part 2 clause 7.4.7.4 requires better than 70% confidence in hardware failure rate data and Part 2 Clause 7.4.7.9 requires 70% single side confidence for operating history. Higher confidence figures are used in Table B6 Part 2 which gives an example of 95% confidence as low effectiveness and 99.9% for high effectiveness, and Part 7 Table D1 provides examples for 95% and 99% confidence from operating experience.

If we were to apply the requirements for 70% confidence this would nearly push the mean failure rate of the system into the next SIL in the example in this paper, and in others with a broader spread it would have a bigger impact.

The more profound impact on the use of the standard might come if we can recommend adjustments to the SIL that can be claimed based on the rigour of the argument that is made and even link a claim limit for SIL to the argument. For example if a process-based qualitative argument was used SIL could be reduced by (at least) 2 levels. If we were to adopt the conservative approach outlined above then we would need at least 99% confidence in SIL2.

5 Conclusions

There is uncertainty in the judgement of the *pfd* of a system whether it is based on direct expert judgment, field experience or the case made from a wide range of test, analysis and experience based evidence. In this paper we have explored how the confidence in these judgments affects the overall judgement of a safety related *pfd* and have illustrated this with an example of SIL membership. It is plausible that judgement of SIL will not be a symmetric distribution: in this case increasing confidence will increase our belief in the integrity of the system. Increasing confidence has an effect on the mean failure rate in these common types of distributions, and this justifies the use of ACARP as a subset of ALARP and the use of confidence building verification activities.

We have modelled the relationship of confidence and mean formally for some particular distributions and have shown that it is more likely that a better safety justification case can be made that if the system is judged as most likely a SIL n+1 system it is taken as a SIL n with high confidence. We have presented some conservative modelling that shows how onerous the need for confidence is. More work is needed to establish quite how conservative this approach is as, in our experience, conservative values at one stage of the analysis do not

necessarily propagate through to other stages of the reasoning.

The application of standards should take into account the rigour of the arguments offered. Compliance with process and the predominance of expert judgement in the safety argument should lead to claims being heavily discounted (e.g. by 2 SILs) and a possible limit put on the claims that can be made.

6 Acknowledgements

This work was partially supported by the UK Engineering and Physical Sciences Research Council (EPSRC) under the DIRC and INDEED projects (Grant EP/E000517/1), and by British Energy Generation Ltd under the DISPO project (Contract No. P/40030532).

7 References

[1] D. M. Hunns and N. Wainwright, "Software-based protection for Sizewell B: the regulator's perspective", *Nuclear Engineering International*, September, pp.38-40, 1991.

[2] RTCA, "Software considerations in airborne systems and equipment certification, Requirements and Technical Concepts for Aeronautics", DO-178B, July 1992.

[3] G Apostolakis, "The concept of probability in safety assessments of technological systems", Science, vol 250, no. 4986, pp. 1359 – 1364, Dec 1990.

[4] IEC 61508, "Functional safety of electrical/electronic/programmable electronic safety-related systems", Parts 1–7, 1998

[5] Cemsis Public Domain Example, www.cemsis.org

[6] Bank of International Settlements. www.basel-ii-risk.com

[7] P G Bishop and R E Bloomfield , "A methodology for safety case development", *Safety-Critical Systems Symposium*, 1998, Birmingham, UK.

[8] Ministry of Defence, Interim Defence Standard 00-56 "Safety Management Requirements for Defence Systems, Part 2 Guidance on Establishing a Means of Complying with Part 1", Issue December 2004.

[9] "Requirements for Safety Related Software in Defence Equipment", Ministry of Defence, 1997.

[10] "Regulatory Objective for Software Safety Assurance in Air Traffic Service Equipment", Civil Aviation Authority, 2001.

[11] "The Use of Computers in Safety-Critical Applications", HSE Books, 1998.

[12] B Littlewood and D R Wright, "The Use of Multi-legged Arguments to Increase Confidence in Safety Claims for Software-based Systems", *IEEE Trans. Software Eng.*, 2007 (to appear).

[13] P G Bishop and R E Bloomfield. "A Conservative Theory for Long-Term Reliability Growth Prediction," *IEEE Trans. Reliability,* vol. 45, no. 4, pp 550-560, 1996.