City Research Online

City, University of London Institutional Repository

Citation: Littlewood, B. & Bloomfield, R. E. (2004). On the use of diverse arguments to increase confidence in dependability claims. Paper presented at the International Conference on Control and Instrumentation in Nuclear Installations, 2004, Liverpool, UK.

This is the unspecified version of the paper.

This version of the publication may differ from the final published version.

Permanent repository link: https://openaccess.city.ac.uk/id/eprint/1601/

Link to published version:

Copyright: City Research Online aims to make research outputs of City, University of London available to a wider audience. Copyright and Moral Rights remain with the author(s) and/or copyright holders. URLs from City Research Online may be freely distributed and linked to.

Reuse: Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge. Provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way.

City Research Online: http://openaccess.city.ac.uk/ publications@city.ac.uk

On the use of diverse arguments to increase confidence in dependability claims

Robin Bloomfield, Bev Littlewood
Adelard and Centre for Software Reliability, City University, London
reb@adelard.com, b.littlewood@csr.ac.uk

Abstract

Intellectual diversity – 'two heads are better than one', 'don't put all your eggs in one basket' – has long been used to reduce the impact of human error. More formally, in recent years engineered design diversity has been used extensively to achieve dependability in software-based systems. This use of design diversity prompted the first formal quantitative studies of the efficacy of intellectual diversity. In this paper we examine how diversity might be used to increase confidence in dependability claims (reliability, safety) about systems. Our purpose is to see whether the probabilistic approach that has been successful in design diversity can be applied to diversity in arguments.

1. Introduction

We are all familiar with informal ways in which diversity is used to increase confidence. For example, if you ask me to check your arithmetic you are tacitly assuming that the use of a different person (intellectual diversity) is more likely to pick up your mistakes than if you simply checked your own work. This idea of 'a different pair of eyes' is widespread in science and engineering. Indeed, it could be said that the whole scientific method, with its checks and balances based on review by colleagues (and rivals), is crucially dependent on the efficacy of diversity. More formally, deliberate diversity in design is often used to gain protection from common faults that may be present in merely *redundant* systems. In particular, in recent years design diversity has been proposed as a means of protecting against *software* faults (see [1] for a recent review of work in this area).

In this paper we consider the use of diversity in *arguments* – so-called multi-legged arguments – that are used to make claims about the dependability (safety, reliability, etc) of a system. The use of diversity here is aimed at increasing the confidence we can place in such claims. Whilst multi-legged arguments to support safety claims have been used for years, there appears to be no formal theory to support them: the work reported here is a tentative beginning to such a theory.

The *need* for better ways of justifying dependability claims is clear. In critical systems, particularly, the costs of justifying safety claims can be enormous, and are likely to grow. In some cases, the costs involved in supporting safety claims can be greater then the costs involved in building the systems. For example, the French RER railway in Paris contains a safety-critical system with about 20,000 lines of code [2]; 100 person years was spent on safety assurance. NASA quoted several years ago a figure of more than $1,000 per line for production and assurance of Space Shuttle code, and it is known that this contains many faults.

Diverse arguments have been applied in real safety cases. For example, in the safety case for the UK's Sizewell B nuclear power plant, an issue concerned what could be claimed about the probability of failure on demand (*pfd*) of the software in the Primary Protection System (PPS). A two-legged argument was used [3] based on the 'special case' procedure of the UK's Safety Assessment Principles for Nuclear Power Plants [4] (also available at http://www.hse.gov.uk/nsd/saps.htm).

Several standards and codes of practice suggest the use of diverse arguments. In UK Def Stan 00-55 [5], for example, one leg is based upon logical proof, the other upon a probabilistic claim based upon statistical analysis. The legs are sometimes quite asymmetric: for example, in [6] the first leg is potentially complex, whereas the second leg is deliberately simple[1]. Occasionally, the only difference between the legs lies in the people involved, e.g. in certain kinds of IV&V.

At an informal level, diversity seems plausibly to be 'a good thing', but there is no theoretical underpinning to such an assertion. In particular, we do not know how much confidence we can justifiably place in a dependability claim when this is supported by diverse arguments. This situation contrasts with the use of

[1] The language here – and indeed the underlying structure – is very similar to a type of *design* diversity used in certain types of safety system, where a complex, highly functional primary version is backed up by a much more simple secondary system. This is just one of many examples of the duality between design diversity and argument diversity, an issue we intend to address in more detail in future work.

diversity in *design*, where probability models have provided significant new insights[2].

It is worth saying here that we should not expect diversity to be a panacea for problems of building dependability cases. Even in areas where we might expect diversity to be particularly successful, it is possible to find surprising evidence for its limitations. We have mentioned earlier the importance diversity seems to play in the 'hard' sciences, e.g. in notions of repeatability of experiments, of independent peer review, etc. There has been some study of this kind of diversity by social scientists. In a fascinating paper [7], it was shown that the diversity present in the world-wide community of physicists did not prevent over-confidence in claims about the accuracy of assessments of numerous physical constants (speed of light, charge on the electron, etc) over more than a century. Such observations are somewhat chastening for anyone involved in the application of diversity to computer-based systems: most of us would not expect our understanding of our domain to rival that of physicists of theirs…

2. How have multi-legged arguments been built in practice? – informal discussion

It is notable that multi-legged arguments show very different structures, and use very different types of content in the different legs. Examples, for two-legged arguments, include:

1. A leg based upon logical proof and a leg based upon statistical evidence from operational testing: e.g. in the case of an argument about software dependability, the first could involve a claim of complete freedom from (a class of) faults, the second a claim for a particular probability of failure upon demand (*pfd*) [5, 8]
2. A leg based upon indirect evidence, such as design process quality, and a leg based upon direct evaluation of the built system, each leg involving assessment via expert judgement: e.g. in the case of software, the first could involve evidence such as CMM (Capability Maturity Model) level and types of procedures used, the second could involve evidence of static analysis and of operational testing [3];
3. A primary argument leg involving extensive evidence, and a simpler secondary leg whose purpose is just to compensate for potentially serious weaknesses in the primary [6];
4. Legs that are based upon exactly the same evidence, but use different, non-communicating teams of human expert analysts.

It is striking how some of these arguments mimic the use of diversity in fault-tolerant design. Thus 4 captures the common design diversity notion of unforced diversity between 'functionally identical' channels: it is the only one in which the evidence (input) and claim (output) are the same between the two legs. Example 3 is similar to the system architecture in which a primary channel has extensive functionality (at the price of complexity), whilst a second 'get-you-home-safely' channel is deliberately kept simple – often the first is implemented in software, the second in hard-wired logic.

Whilst these examples all involve diversity, this diversity is used in quite different ways, and the legs differ widely both in content and type of claim. In example 1, the first leg involves a claim for complete perfection of operational behaviour (at least with respect to a subclass of failures) based upon logical reasoning, whereas the second leg would only allow a probabilistic claim based upon statistical evidence. If the overall argument is intended to support a claim of (better than) 10^{-4} *pfd*, only the statistical leg addresses this directly. Nevertheless, it is easy to see how the logical leg can support the other: if the statistical evidence alone gives 99% confidence that the *pfd* is smaller than 10^{-4}, then the additional 'proof leg' might allow this level of confidence in 10^{-4} to be increased.

Note, however, the dependence here: the observation of a failure in the testing leg would completely refute the perfection claim of the first leg. A similar potential dependence can be seen in the second example: direct evidence arising from the examination of the built system in the second argument leg could cast doubt upon the claims for process quality coming from the first leg. Just as claims for independence between system failures are generally not believable, so it seems we might expect there to be dependence between different argument legs.

The differences between these examples suggest that there is no agreement about the best ways to structure diverse arguments. Informally, we might say that our goal is to make claims at as high a 'level' as we can, with 'confidence' as high as we can make it. But it is not clear how we should go about achieving these goals, nor even how we should formally express them (e.g. how we should resolve the trade-off between claim level and confidence).

For example, what are the relative merits of arguments that mimic the symmetric structure of 1-out-of-2 protection systems, and those unsymmetric arguments where a second leg is designed to compensate for the expected weaknesses of a primary leg? Perhaps the former are more appropriate for those situations where we have little knowledge of the precise ways in which arguments might fail – here we would be depending upon a claim for *general* efficacy of diversity.

What are the relative merits of 'forced' and 'natural' diversity? Again, this may come down to how much knowledge we have about potential weaknesses in

[2] It has to be admitted that these insights have been mainly conceptual ones – there remain great difficulties in estimating the parameters of the models in particular instances to allow their use in safety cases. Nevertheless, the models have produced better understanding – for example by warning of the perils of unwarranted assumptions of independence of failure behaviour between versions.

arguments – we need to know that, if we have forced diversity in a particular way, this is appropriate for our problem. Issues of this kind also arise, of course, when diversity is used to achieve dependability in systems; here also it is necessary to be confident that 'functional diversity' fits the problem. For example, in the design of a 1-out-of-2 protection system the choice to monitor temperature and pressure in the different channels, rather than (say) temperature and flow rate, requires knowledge on the part of the designers.

The answers to questions like these depend on understanding better what it is we are trying to protect ourselves from by using diverse arguments. What are the kinds of weaknesses present in single arguments that might be addressed by diversity? These seem to fall into two kinds: weaknesses in modelling assumptions, and weaknesses in evidence[3].

Any argument to support a dependability claim about a system will depend upon some assumptions. For example, a claim of 'perfection' about some software, based upon a formal verification that the program correctly implements its specification, assumes that this formal specification is an accurate representation of some higher level informal engineering requirements. If we had any doubt about the truth of this assumption (and how could we not?), we might require a second argument leg in the form of appropriately extensive operational testing. Seeing no failures in the testing (judged against the engineering requirements) would make us more confident in the truth of the assumption (in spite of what Edsger Dijkstra said[4]…).

Weaknesses in the *evidence* for a single argument leg can similarly be a reason to require a second leg. For example, here we might reverse the reasoning of the previous paragraph. If the first leg involves a statistical claim using operational testing, but it is infeasible to test for sufficiently long to substantiate the claim at the required level of confidence, we might require a second argument leg involving extensive static analysis.

Whatever the reason for using multiple argument legs, the intention is always the same: we want to be able to have more (justifiable) confidence in the dependability of our system from the two (or more) legs than is provided by either one alone. This issue of composability seems fundamental: *how much* benefit do we get from this approach in a particular instance? In the following section we present a somewhat tentative formalism to address this question for a simple example of a two-legged argument.

3. Uncertainty, confidence and diversity – a tentative formalism

Uncertainty is ubiquitous in dependability studies. Most obviously, we are uncertain about when systems will fail. In the face of such uncertainty we use probability models as a means of describing the failure behaviour. So measures of dependability are probabilistic: e.g. probability of failure on demand, failure rate, etc.

All this is well understood and widely accepted,[5] and there is a comprehensive probabilistic theory and methodology for systems dependability. Less well understood is the uncertainty associated with the *assessment* of dependability. We have argued elsewhere [10] that a formal theory of uncertainty, based upon probability, is needed here to capture the notion of *confidence* in dependability claims. In particular, we need such a formalism to be able to analyse the efficacy of argument diversity as a means of increasing confidence in dependability claims: cf. the probabilistic modelling of design diversity, which allows its benefits in increasing systems reliability to be analysed [1].

The interpretation of probabilities in these two situations is different of course. Probabilistic measures of dependability concern 'natural' uncertainty in the world; probabilistic measures of confidence in dependability claims concern beliefs about the world, e.g. about whether some assumptions underpinning the reasoning in an argument are true. It seems inevitable that there will be a subjective element in the latter, and thus probabilities – confidence – will need to be interpreted in a Bayesian, subjective way.

In a diverse *system* – e.g. a 1-out-of-2 system – the reliability is determined by the reliabilities of the individual versions and the dependence between the two version failure processes. It seems possible that similar considerations apply in the case of diverse *arguments*.

Thus 'confidence' could play the same role for an argument as 'reliability' does for a system. Similarly, argument 'dependence' might also be defined in terms of confidence. Thus we could say that two arguments are *independent* if they allow simple multiplicative manipulation of confidence, as in the case of reliability. For example, we would say that two arguments *A* and *B* are independent if each *individually* gives 90% confidence that the *pfd* is smaller than 10^{-3} and *together* they give 99% confidence in the same claim.

In practice, it seems unlikely that arguments will be independent in this way, just as claims for independence between the failure behaviour of design-diverse systems

[3] For simplicity we shall ignore here a third possibility, that the *reasoning* used by the expert to make the claim, based on the assumptions and the evidence, is flawed. This may need to be considered in some cases, but we believe that it can be treated as a simple extension of the general approach described here.

[4] 'Program testing can be used to show the presence of bugs, but never to show their absence!' [9]E. W. Dijkstra, "Notes on structured programming," in *Structured Programming*, O.-J. Dahl, E. W. Dijkstra, and C. A. R. Hoare, Eds. London and New york: Academic, 1972, pp. 1-82.

[5] Although there have been attempts to eliminate uncertainty by using formal verification to support claims for complete perfection in software-based systems, these have been largely unsuccessful. There is now a consensus that for systems of even modest complexity such claims will be unbelievable.

are rarely believable. Notice, however, that dependence need not be 'a bad thing': as in design diversity [11], there may be the (theoretical) possibility for a kind of negative covariance in argument diversity, resulting in confidence even greater than would come from independence. In fact, we may have some knowledge about the weaknesses of one argument, and be able to construct a second one to avoid these (albeit at the expense of containing other, novel, weaknesses).

We now proceed to develop these ideas somewhat more formally. Each argument leg can be thought of as a triple, comprising some *evidence* and modelling *assumptions* that together support a *claim* or *assertion* at a certain level of confidence.

Consider the situation of an argument leg *A* in which a safety goal is expressed as an assertion, G_A. For example:

- G_A: 'the probability of failure on demand of the protection system software is less than 10^{-3}'.

Argument *A* is based upon some assumptions that we call Ass_A. An example might be:

- Ass_A: 'the statistical testing, from which we shall obtain a quantitative estimate of the reliability, is truly representative of the distribution of demands that will be met in real system operation, and these demands are statistically independent'.

We now go out and collect evidence for the leg. That is we conduct a statistical test, e.g. we observe 4603 demands under conditions that satisfy ass_A (this number of demands is chosen so that, if they exhibit no failures, we shall be able to claim that the system has, with 99% confidence, a better than 10^{-3} *pfd* [12]). The evidence then tells us, subject to the truth of our assumptions, whether that leg stands or falls (i.e. whether we *succeed* in executing the 4603 demands without observing a failure, or we *fail*). Denote by E_A the event that the evidence for *A* does in fact turn out to support the assertion G_A:

- E_A: 4603 demands are executed without failure.

Then, for argument leg *A* we can say

$$P(\overline{G_A} \mid E_A, ass_A) \leq \alpha \text{ or } P(G_A \mid E_A, ass_A) \geq 1-\alpha \qquad (1)$$

where α is 0.01. That is, we can say that the argument leg *A*, represented by the triple (G_A, E_A, ass_A), supports the claim at the (1-α) level of confidence, given the truth of the assumptions and the support of the evidence. More precisely, the probability of the claim being false, given the evidence is supportive and the assumptions are true, is no greater than α.

The uncertainty in this kind of argument, which gives rise to the doubt in the truth of the claim expressed in the probabilistic confidence level, comes from the statistical nature of the evidence. Other kinds of argument may be completely deterministic, so that a claim can be expressed with certainty assuming the truth of the assumptions. For example, consider argument leg *B*:

- G_B: 'the protection system software contains no faults'.
- Ass_B: 'the formal specification correctly captures the informal engineering requirements of the system';
- E_B: the mathematical verification that the program implements the specification is successful;

Then, for argument leg *B*, represented by the triple (G_A, E_A, ass_A) we are certain that the assertion is true. That is, we can say

$$P(\overline{G_B} \mid E_B, ass_B) = 0 \text{ or } P(G_B \mid E_B, ass_B) = 1$$

In both these examples, the reasoning assumes the truth of the assumptions. If, as seems likely, there is some doubt about the truth of the assumptions of an argument leg, this will change (reduce) the confidence that we have in the claim. For the deterministic argument, *B*, we have

$$P(\overline{G_B} \mid E_B) = P(\overline{G_B} \mid E_B, ass_B)P(ass_B)$$
$$+P(\overline{G_B} \mid E_B, \overline{ass_B})P(\overline{ass_B})$$

i.e.

$$P(\overline{G_B} \mid E_B) \leq P(\overline{ass_B})$$

if we conservatively assume that the claim is false when the argument is based upon a false assumption.

For the statistical argument leg *A*, similarly, we have

$$P(\overline{G_A} \mid E_A) = P(\overline{G_A} \mid E_A, ass_A)P(ass_A)$$
$$+P(\overline{G_A} \mid E_A, \overline{ass_A})P(\overline{ass_A})$$

If we conservatively assume

$$P(\overline{G_A} \mid E_A, \overline{ass_A}) = 1$$

this becomes

$$P(\overline{G_A} \mid E_A) \leq \alpha P(ass_A) + P(\overline{ass_A}) \qquad (2)$$

As an example, if we were 99% confident in the truth of the assumptions underpinning *A*, then (2) shows that our confidence in the claim G_A has about halved compared with the case when we are completely certain of the assumptions: the bound (2) is approximately 0.02 compared with 0.01 from (1).[6]

Expression (2) shows the different roles played in this example by *extensiveness of evidence* and *assumption confidence* in arriving at the confidence level that can be placed in the safety claim. By collecting more evidence of a supportive nature (i.e. observing more failure-free demands of the protection system) we can reduce the value of □ and thus increase our confidence in the claim. However, our scope for doing this is restricted by the level of our confidence in the assumptions, represented by $P(\overline{ass_A})$. There is clearly a limit to the value of collecting more evidence of the same kind in the face of such assumption uncertainty.

One interpretation of the reasoning behind requiring a multi-legged argument is that it is a way to overcome, or at least minimise, this problem. It is a means of increasing our confidence in *G*, given the evidence, *when we have doubts about the truth of the assumptions underpinning an individual argument.*

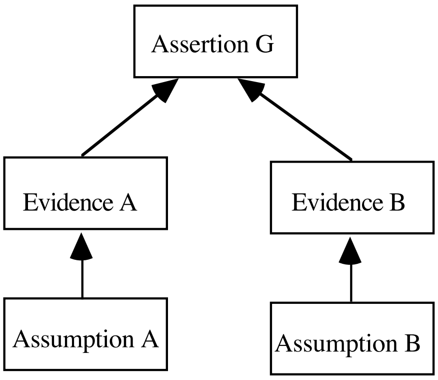

Figure 1. The structure of a two-legged argument in support of a claim *G*

Consider the case of a two-legged argument like that of Fig 1. This is very similar to a 1-out-of-2 *system* structure: a claim is rejected if either argument leg rejects it. Just as we would claim that the probability of failure of a 1-o-o-2 system was less than the probability of failure of either channel, so here we can expect that the chance of accepting a false claim will be less with the two-legged structure than it would be with either leg on its own. The detailed way in which confidence in the claim depends upon the evidence and assumptions can be seen in the following expression:

$$\begin{aligned}
P(\overline{G} \mid E_A, E_B) &= P(\overline{G} \mid E_A, E_B, ass_A, ass_B)P(ass_A, ass_B) \\
&+ P(\overline{G} \mid E_A, E_B, ass_A, \overline{ass_B})P(ass_A, \overline{ass_B}) \\
&+ P(\overline{G} \mid E_A, E_B, \overline{ass_A}, ass_B)P(\overline{ass_A}, ass_B) \\
&+ P(\overline{G} \mid E_A, E_B, \overline{ass_A}, \overline{ass_B})P(\overline{ass_A}, \overline{ass_B})
\end{aligned} \tag{3}$$

We shall now consider some simple examples to illustrate how this works.

Example 1

Consider the case where each of the arguments is of the deterministic, logical type, supporting a claim for impossibility of failure, like *B* above: call them B_1, B_2. We have

$$\begin{aligned}
&P(\overline{G} \mid E_{B_1}, E_{B_2}) \\
&= P(\overline{G} \mid E_{B_1}, E_{B_2}, \overline{ass_{B_1}}, \overline{ass_{B_2}})P(\overline{ass_{B_1}}, \overline{ass_{B_2}}) \\
&\Box\ P(\overline{ass_{B_1}}, \overline{ass_{B_2}})
\end{aligned} \tag{4}$$

Here we have used the fact that in this type of argument, if either of the assumptions is true, and the evidence supports the claim, then the claim is true with certainty. Thus the first three terms of the right hand side of (3) are zero. In addition, the bounding value assumes conservatively that the claim is false with certainty if both assumptions are false.

Clearly, there will be greater confidence in the claim, *G*, using the two-legged argument compared with using only one argument leg. The benefit will depend upon how much smaller is the probability of both assumptions being false, compared with the probabilities of single assumptions being false.

Example 2

Consider now an argument involving two statistical legs A_1, A_2, similar to *A* above. We have, from (3):

$$\begin{aligned}
&P(\overline{G} \mid E_{A_1}, E_{A_2}) \\
&\Box\ P(\overline{G} \mid E_{A_1}, E_{A_2}, ass_{A_1}, ass_{A_2})P(ass_{A_1}, ass_{A_2}) \\
&+ P(\overline{G} \mid E_{A_1}, ass_{A_1})P(ass_{A_1}, \overline{ass_{A_2}}) \\
&+ P(\overline{G} \mid E_{A_2}, ass_{A_2})P(\overline{ass_{A_1}}, ass_{A_2}) \\
&+ P(\overline{ass_{A_1}}, \overline{ass_{A_2}})
\end{aligned} \tag{5}$$

Here we have assumed that, when the assumption underpinning a leg is false, the argument conservatively

[6] When we say that our confidence is halved, we mean that our belief in the *falsity* of the claim has doubled. This is similar to 'reliability' studies, where probabilities almost always relate to *unreliabilities*. Whilst our discussion in the paper will use words like confidence, we shall almost always be dealing with such probabilities of falsity or failure.

reduces to a single leg, i.e. confidence in the claim depends only upon the leg that is based on a true assumption. The bounding value, again, conservatively assumes that when both legs are based on false assumptions the claim is certain to be false.

Assuming further that our beliefs about the arguments are 'symmetric' in the following way (this can be thought of, informally, as a kind of indifference between the arguments), and extending the earlier notation for the single argument, we have:

$$P(\overline{ass_{A_1}}) = P(\overline{ass_{A_2}}) = p$$
$$P(\overline{ass_{A_1}}, ass_{A_2}) = P(ass_{A_1}, \overline{ass_{A_2}}) = q \;\; (< p)$$
$$P(\overline{ass_{A_1}}, \overline{ass_{A_2}}) = r$$
$$P(\overline{G} \mid E_{A_1}, ass_{A_1}) = P(\overline{G} \mid E_{A_2}, ass_{A_2}) = \alpha$$

A special case is the one where the argument assumptions can be seen as 'mutually exclusive', i.e.

$$P(ass_{A_1}, ass_{A_2}) = 0$$

Then, since

$$P(ass_{A_1}, ass_{A_2}) + P(ass_{A_1}, \overline{ass_{A_2}}) + P(\overline{ass_{A_1}}, ass_{A_2}) + P(\overline{ass_{A_1}}, \overline{ass_{A_2}}) = 1$$

we have

$$2q + r = 1$$

and so from (5) we get

$$P(\overline{G} \mid E_{A_1}, E_{A_2}) \le \alpha(1 - r) + r \qquad (6)$$

which contrasts with the single argument result, (2):

$$P(\overline{G} \mid E_{A}) \le \alpha(1 - p) + p \qquad (7)$$

To get a feel for the benefits of a two-legged argument over a single one in these rather specialised circumstances, consider the following numerical examples. Let α=0.1, p=0.1, and r=0.05 (suggesting a plausible positive dependence between the assumptions). The bound in (7) is 0.19, that in (6) is 0.145. Thus confidence in the claim has risen from 81% (based on a single argument) to 85.5% as a result of using two argument legs.

Consider now the special case where α=0.1, p=0.1, and r=0.01. Here $r=p^2$, i.e. the same value for the probability of simultaneous assumption falsity as would occur if there were independence between the assumptions (but note that in fact there is dependence here). The two-legged argument increases the confidence in the claim from 81% to 89.1%.

If we were able to reduce the chance of simultaneous failure (falsity) of the arguments further, in the spirit of the 'forced design diversity' idea of [11], we would have even greater confidence in the claim. Let

$$P(\overline{ass_{A_1}}, \overline{ass_{A_2}}) = (1 - \beta) p^2 \text{ where } \beta > 0$$

and it is easy to show that

$$P(\overline{G} \mid E_{A_1}, E_{A_2}) \le \alpha(1 - p^2) + p^2 - (1 - \alpha)\beta p^2$$

where the last term is the improvement over the previous case of $r=p^2$. If α=0.1, p=0.1, and β=0.5, our confidence in the claim now becomes 89.55%

Remember that in this example the highest confidence in the claim from a single argument is 90%, which occurs when we are certain the assumption is true (the uncertainty then centres entirely upon the evidence, i.e. upon the value of α). Thus the use of two legs has almost restored the confidence to the highest level possible, i.e. almost completely eliminated the problem of assumption doubt. In fact, if we could be *certain* that the two arguments were not both false, i.e. β=1, we get exactly this maximum confidence – all the uncertainty in the claim comes from the evidence.

The *worst* case in this example occurs when r=0.1 (i.e. $=p$): there is then no benefit from the two-legged argument, and the confidence in the claim is exactly the same as for a single argument, i.e. 81%.

It is worth briefly considering the case of independence. That is, instead of the 'mutually exclusive' arguments used above, consider the case where

$$P(\overline{ass_{A_1}}) = P(\overline{ass_{A_2}}) = p$$
$$P(\overline{ass_{A_1}}, ass_{A_2}) = P(ass_{A_1}, \overline{ass_{A_2}}) = p(1 - p)$$
$$P(ass_{A_1}, ass_{A_2}) = (1 - p)^2$$
$$P(\overline{G} \mid E_{A_1}, ass_{A_1}) = P(\overline{G} \mid E_{A_2}, ass_{A_2}) = \alpha$$

Then

$$\begin{aligned} &P(\overline{G} \mid E_{A_1}, E_{A_2}) \\ &\le P(\overline{G} \mid E_{A_1}, E_{A_2}, ass_{A_1}, ass_{A_2})(1 - p)^2 + 2\alpha p(1 - p) + p^2 \\ &\le \alpha(1 - p)^2 + 2\alpha p(1 - p) + p^2 \\ &= \alpha(1 - p^2) + p^2 \end{aligned} \qquad (8)$$

where it is conservatively assumed that confidence in the claim G based on two valid legs is only as great as that based on one (i.e. 1-α). As above, when α=0.1, p=0.1 the two-legged argument raises confidence in the claim from 81% to 89.1%.

So far in this example we have made very conservative assumptions that ensure we can never do better than a single 'perfect' argument leg, i.e. one where the assumption is known to be true. Then, for the numerical values we have used for illustration, interest has centred upon how close the use of diversity can bring us to this 'perfect' argument confidence level of 90% (=1-□). That is, we have been solely concerned with the doubt in the claim that arises from our doubt in the assumptions, rather than in the evidence. A less conservative approach (in fact, perhaps unrealistically optimistic) would be to assume that the truth of G is certain if it is supported by two valid arguments. We could then, for example, ignore the first term on the right of (8). In this case, the bound on confidence in G from the two-legged argument rises to 97.2%, which exceeds the highest confidence we could ever get from a single argument.

Example 3

Consider now a 'mixed' two-legged argument in which one leg is of the 'logical certainty' type (see B above), and the other is statistical (see A). We shall conservatively take the claim G to be the same as G_A – since G_B implies G_A. In (3)

$$\begin{aligned}
P(\overline{G} \mid E_A, E_B) &= P(\overline{G} \mid E_A, E_B, ass_A, ass_B) P(ass_A, ass_B) \\
&+ P(\overline{G} \mid E_A, E_B, ass_A, \overline{ass_B}) P(ass_A, \overline{ass_B}) \\
&+ P(\overline{G} \mid E_A, E_B, \overline{ass_A}, ass_B) P(\overline{ass_A}, ass_B) \\
&+ P(\overline{G} \mid E_A, E_B, \overline{ass_A}, \overline{ass_B}) P(\overline{ass_A}, \overline{ass_B})
\end{aligned}$$

the first and third terms on the right hand side are zero, because G is true with certainty, from leg B, if the verification, based on valid assumptions, supports G. So the probability of incorrectly deciding that G is true is:

$$\begin{aligned}
P(\overline{G} \mid E_A, E_B) &= P(\overline{G} \mid E_A, E_B, ass_A, \overline{ass_B}) P(ass_A, \overline{ass_B}) \\
&+ P(\overline{G} \mid E_A, E_B, \overline{ass_A}, \overline{ass_B}) P(\overline{ass_A}, \overline{ass_B}) \\
&\square\, P(\overline{G} \mid E_A, ass_A) P(ass_A, \overline{ass_B}) + P(\overline{ass_A}, \overline{ass_B}) \\
&= \square P(ass_A, \overline{ass_B}) + P(\overline{ass_A}, \overline{ass_B})
\end{aligned}$$

Here we have assumed, conservatively, that G is false if the two sets of assumptions are false; and that if just ass_B is false, then confidence in G depends only upon leg A.

This bound contrasts with the single argument cases (in an obvious extension of the earlier notation):

$$P(\overline{G} \mid E_A) \,\square\, \square(1 \square p_A) + p_A$$

$$P(\overline{G} \mid E_B) \,\square\, p_B$$

If we assume independence here, the bound for the two-legged argument becomes:

$$\square(1 \square p_A) p_B + p_A p_B \qquad (9)$$

Letting p_A=p_B=0.1, confidence in the claim G is 98.1%, an increase from 81% from the A argument alone, or 90% from the B argument alone. Note, however, that this 98.1% falls short, as would be expected, of the 100% confidence we have when the B assumption is known to be true with certainty. On the other hand, it is better than could be attained by A alone *even if we knew ass_A were true*.

4. Discussion and conclusion

We have only considered in this paper some quite special examples of the use of diverse argument legs. Although further work is needed, we think that they give us some insight into the way that diverse argument legs work, and the benefits of extra confidence in dependability claims that they can bring.

The examples show that – not surprisingly – there is an increase in confidence about a dependability claim, when using a two-legged argument, compared with the confidence to be gained from either of the legs alone. On the other hand, it is not easy to quantify this increase in confidence without making many simplifying assumptions, such as independence between different argument assumptions. Another huge difficulty is to assign numerical values to the many different parameters in expressions like (3) and its successors.

An important issue that needs further exploration is the interplay between 'evidence' and 'assumptions' in arguments. In particular, can this kind of modelling help in constructing arguments that make intelligent trade-offs between the *extensiveness* of evidence and *doubt* in argument validity?

Not surprisingly, issues of dependence (and independence) play an important role in determining the levels of confidence that come from multi-legged arguments. A naïve claim of independence in the confidence we place in the truth of the two different sets of argument assumptions seems unreasonable here for exactly the same reasons that Eckhardt and Lee [13] first proposed in the case of design diversity. Specifically, it seems likely that if we were to discover that ass_A were false, we might decrease our confidence that ass_B were true. The reasoning here is that the evidence of ass_A's falsity suggests that we 'do not understand things well' in a general sense. Thus if we found out that our statistical testing was not an accurate representation of operational use, this might make us doubt whether we had correctly captured the engineering requirements in *other* ways - in particular in writing a formal specification against which to conduct a verification for argument B.

On the other hand, continuing the analogy with design diversity, there is a *possibility* in certain circumstances of deploying arguments of forced diversity, as in [11]. That is, assumptions might be devised such that

$$P(\overline{ass_A}, \overline{ass_B}) < P(\overline{ass_A})P(\overline{ass_B})$$

In design diversity, such claims would generally be treated with justifiable suspicion. Is there any reason to be less sceptical in the case of diverse argument legs? The optimistic view would be that we might have a better understanding of the potential weaknesses of arguments – and so be able to build ones that are complementary to one another with respect to these weaknesses – than is the case in systems design.

Even if this *could* be done, there would presumably be a price to be paid in the amounts of evidence needed in the individual legs: the new assumptions would be of necessity 'weaker'. We might need, for example, to circumscribe ourselves strongly in each case as to what could be assumed, in order to seek this 'negative dependence'. And it seems reasonable to expect that to support a claim at a particular level of confidence with weaker assumptions would require stronger (e.g. more) evidence.

Example 3 shows an interesting aspect of dependence between legs when the evidence from the testing leg, *A*, includes at least one failure of the system. In this case the testing leg completely refutes the proof leg, *B*: if a fault is found in testing, of a type that the proof leg claimed was completely absent, the confidence in the proof leg is immediately reduced to zero.[7] The result is that the multi-legged argument is reduced to the single testing leg, which may or may not have sufficiently strong evidence to support the claim at the required confidence (e.g. if there is only one failure, but 6635 failure-free demands seen in a test, then argument *A* will support a claim for a *pfd* of 10^{-3} at 99% confidence [12])

Examples like the ones above may be somewhat special, inasmuch as each argument alone allows the same claim to be made with a certain confidence – namely that the *pfd* is smaller than 10^{-3} (even *B* does this), which is the top-level claim for the overall two-legged argument. Such examples are thus analogous to the use of design diversity in a 1-out-of-2 system in which each subsystem has *similar* functionality. It was this special structure that allowed us to discuss 'dependency' above simply via confidence. Not all multi-legged arguments have this useful symmetry, just as not all applications of diversity in system design are of the 1-out-of-*n* type. It might be interesting to consider other types of diverse system design and see whether there are analogies for diverse arguments.

This paper has tentatively addressed only a small part of what seems to be a large and difficult problem. Obviously, there is much research work to be done before we have a formal model that supports the effective use of diversity in dependability arguments. Nevertheless, the approach does seem promising, and such research will have been worthwhile if it eventually allows us to say *how much* our confidence in dependability claims can be increased by the use of diversity.

Finally, it may be worth indicating how this work might eventually be useful in practice. There are several different approaches to dependability cases, across different sectors and different standards and guidelines. In principle, a better understanding of 'confidence' and argument diversity will help us to understand the strengths and weaknesses of these, and perhaps aid in selecting the best for particular applications.

5. Acknowledgements

This work was partially supported by the DISPO-2 (DIverse Software PrOject) Project, funded by British Energy Generation Ltd and BNFL Magnox Generation under the IMC (Industry Management Committee) Nuclear Research Programme under Contract No. PP/40030532, and by the project DIRC ('Interdisciplinary Research Collaboration in Dependability of Computer-Based Systems') funded by UK Engineering and Physical Sciences Research Council (EPSRC). An earlier version of this paper was presented at the International Conference on Dependable Systems and Networks (DSN2003) in San Francisco, 2003.

6. References

[1] B. Littlewood, P. Popov, and L. Strigini, "Modelling software design diversity - a review," *ACM Computing Surveys*, vol. 33, pp. 177-208, 2002.

[2] G. Guiho and C. Hennebert, "SACEM software validation," presented at 12th International Conference on Software Engineering, 1990.

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

[4] HSE, "Safety Assessment Principles for Nuclear Plants," Health and Safety Executive ISBN 011 882043 5, 1992.

[5] MoD, "The Procurement of Safety Critical Software in Defence Equipment," Ministry of Defence Def-Stan 00-55, Issue 2, August, 1997 1997.

[7] We are not, here, considering what might be claimed after the supposed removal of this fault. It may be that the previous (flawed) proof can be used, together with evidence of the efficacy of the fault removal, to support a non-zero confidence that the software is *now* completely free of this class of faults.

[6] CAA, "Regulatory Objective for Software Safety Assurance in Air Traffic Service Equipment," Civil Aviation Authority SW01, 2001.

[7] M. Henrion and B. Fischhoff, "Assessing uncertainty in physical constants," *Americal J. of Physics*, vol. 54, pp. 791-798, 1986.

[8] B. Littlewood, "The use of proofs in diversity arguments," *IEEE Trans Software Engineering*, vol. 26, pp. 1022-1023, 2000.

[9] E. W. Dijkstra, "Notes on structured programming," in *Structured Programming*, O.-J. Dahl, E. W. Dijkstra, and C. A. R. Hoare, Eds. London and New york: Academic, 1972, pp. 1-82.

[10] R. E. Bloomfield and B. Littlewood, "Confidence in safety claims," Centre for Software Reliability, City University http://bscw.cs.ncl.ac.uk/bscw/bscw.cgi/0/43989, 2003.

[11] B. Littlewood and D. R. Miller, "Conceptual Modelling of Coincident Failures in Multi-Version Software," *IEEE Trans on Software Engineering*, vol. 15, pp. 1596-1614, 1989.

[12] B. Littlewood and D. Wright, "Some conservative stopping rules for the operational testing of safety-critical software," *IEEE Trans Software Engineering*, vol. 23, pp. 673-683, 1997.

[13] D. E. Eckhardt and L. D. Lee, "A Theoretical Basis of Multiversion Software Subject to Coincident Errors," *IEEE Trans. on Software Engineering*, vol. 11, pp. 1511-1517, 1985.