

# The Use of Multi-legged Arguments to Increase Confidence in Safety-Claims for Software-Based-Systems: a Study Based on a BBN Analysis of an Idealised Example

Bev Littlewood<sup>a,\*</sup> David Wright<sup>a</sup>

<sup>a</sup>*CSR, City University, Northampton Square, London, EC1V 0HB*

---

## Abstract

The work described here concerns the use of so-called multi-legged arguments to support dependability claims about software-based systems. The informal justification for the use of multi-legged arguments is similar to that used to support the use of multi-version software in pursuit of high reliability or safety. Just as a diverse, 1-out-of-2 *system* might be expected to be more reliable than each of its two component versions, so a two-legged *argument* might be expected to give greater confidence in the correctness of a dependability claim (e.g. a safety claim) than would either of the argument legs alone.

Our intention here is to treat these argument structures formally, in particular by presenting a formal probabilistic treatment of ‘confidence’. This will enable claims for the efficacy of the multi-legged approach to be made quantitatively, answering questions such as ‘How much extra confidence about a system’s safety will I have if I add a verification argument leg to an argument leg based upon statistical testing?’

For this initial study, we concentrate on a simplified and idealized example of a safety system in which interest centres upon a claim about the probability of failure on demand. Our approach is to build a BBN model of a two-legged argument, and manipulate this analytically via parameters that define its node probability tables. The aim here is to obtain greater insight than is afforded by the more usual BBN treatment, which involves merely numerical manipulation.

We show that the addition of a diverse second argument leg can, indeed, increase confidence in a dependability claim: in a reasonably plausible example the doubt in the claim is reduced to one third of the doubt present in the original single leg. However, we also show that there can be some unexpected subtleties here; for example an entirely supportive second leg can sometimes undermine an original argument, resulting in less confidence than came from this original argument. Our results are neutral on the issue of whether such difficulties will arise in real life - i.e. when real experts judge real systems.

## 1 Introduction

Assessment of dependability of software-based systems has long been acknowledged to be difficult. There are several reasons for this. Software is often novel, so that claims can rarely be based upon previous experience. Much of the evidence available concerns the software process – how it was built – and not the built product itself. There is a great reliance upon expert judgement – e.g. in how claims about the quality of the build process can be turned into claims about the delivered product’s dependability.

Such uncertainty about dependability claims is particularly important when the systems involved are safety critical. One approach that has been proposed to try to limit and control this uncertainty is the use of diverse arguments to support dependability claims: the idea is analogous to the use of fault tolerance to make systems reliable.

In recent years, some standards and codes of practice have suggested the use of diverse arguments. In UK Def Stan 00-55 [1], for example, it is suggested that one leg be based upon logical proof of correctness, the other upon statistical analysis to support a probabilistic claim. Argument legs are sometimes quite asymmetric: for example, in [2] the first leg is potentially complex, whereas the second leg is deliberately simple. Occasionally, the only difference between the legs lies in the people involved, e.g. in independent V&V. This last can be plausible when there is a paucity of hard empirical evidence upon which to base the arguments, and thus necessarily a large element of expert judgement is used – different teams might provide some protection against identical human mistakes.

The differences shown in these examples reflect, we believe, the need for better understanding about the use of diversity in arguments. At an informal level, diversity seems plausibly to be ‘a good thing’, but there is no theoretical underpinning to such an assertion. For example, we do not know what are the ‘best’ ways to use diversity (nor even exactly what ‘best’ means here); we do not know how much we can claim for the use of diversity in a particular case.

Our aim in this paper is to provide the beginnings of a formalism to answer some of these questions, and provide support for the ‘diversity approach’. We shall look at the combination of multiple legs in a part of a safety case for a critical system. The difficult questions here concern how to combine the disparate evidence and assumptions that form the different legs. There are several such questions that might be of interest. For example, we might want to know whether multi-legged arguments are efficient in the sense of being cost-effective: for a given outlay, would it be better to divide this between a

---

\* Corresponding author.

proof and a testing leg, or to spend it all on a larger test?

It seems to us inevitable that questions of this kind should take account of both the *claim level* and the *confidence* in that claim level which arise from an argument. Thus an argument that allows us justifiably to place 99% confidence in a claim that a system's probability of failure on demand is less than  $10^{-3}$ , is clearly superior to an argument that only allows us to place 90% confidence in the same claim. In the following work, we shall assume that the claim is a given – e.g. it arises as a requirement from the safety case of some wider system. Thus comparisons between arguments can be made by comparing the confidence that they engender in this fixed claim. Clearly, this approach is not the only way that the problem could be tackled, but it will suffice for the purposes of the present paper.

The work reported here can be seen as a more formal treatment of the example examined in [3, Example 1, p27]. In this example a 2-legged argument was proposed. Firstly, a leg based upon statistical evidence from operational testing and the use of an oracle produces a claim for a particular probability of failure upon demand (pfd). It is reasoned that this pfd represents a sufficiently small risk during the expected operational life of the system. To this part of the argument is added a second leg based instead on logical reasoning which is assumed to produce a claim for complete perfection of operational behaviour (at least with respect to a subclass of failures). Here, the second leg produces a claim of complete freedom from (a class of) faults. If the overall argument is intended to support a claim of (better than)  $10^{-3}$  pfd, then only the statistical *testing leg* addresses this directly. Nevertheless, it is easy to see how the logical leg can provide additional support: if the statistical evidence alone gives 99% confidence that the pfd is smaller than  $10^{-3}$  then the additional *verification leg* might allow this level of confidence in  $10^{-3}$  to be increased.

Note, however, that in this example the observation of a failure in the testing leg would completely refute the perfection claim of the second leg. This is a particular instance of *dependence* between different argument legs. If we recall that argument diversity is a form of fault tolerance, similar to the use of *version* diversity in fault tolerant *system* design, we might expect that such dependence between argument legs will have an effect upon the confidence that we can justify when they are used in a multi-legged argument. One of the lessons from *system* diversity is that claims for independence are generally not believable. If this is also true of arguments, as seems intuitively likely, then we would need to be sceptical of simplistic claims, e.g. that 99% confidence could be justified from a two-legged argument based on two legs each of which alone only allow 90% confidence. Proper understanding of argument *dependence* is therefore going to be an important goal of this research. It turns out that issues of dependence between legs can be subtle and counter-intuitive.

We realise that much of what we say here applies to dependability assessment of systems in general, but the issues discussed are often particularly acute for software based systems, where there may be great complexity and novelty in the system design, and there is typically a large reliance in the safety assessment on expert judgement. We shall conduct a formal analysis of arguments to support reliability claims about a simple, idealised demand-based system like the one above. This means that the operational use takes the form of a series of executions over discrete time, and our interest centres on the probability of failure on demand. Our 2-legged argument consists of: (i) the statistical *testing leg*, using an oracle to test the output of the system during a series of simulated demands that are assumed to be representative of operational use; and (ii) the code *verification leg* consisting of a formal investigation of the correctness of the built system as judged against its specification. Note that an important source of potential ‘argument failure’ is the possibility of incorrectly assuming the truth of assumptions such as ‘the oracle is perfect’, or ‘the specification is correct’. Note also that beliefs about such assumptions are likely to be dependent. Such dependence will clearly induce dependence in the argument legs - most likely resulting in less confidence in the dependability claim.

The most important argument failure from the point of view of a safety case is, of course, the acceptance of an untrue safety claim. The other kind of argument failure - rejecting a claim when it is true - will also be important (e.g. it may have severe economic consequences), but will not be considered in the current paper.

We begin by describing a 6-variable BBN representing the structure of this 2-legged argument supporting a safety claim. A first pass BBN topology for this system assessment is first presented, with a comprehensive enumeration of the *independency model*<sup>1</sup> which it represents.

Also contained here are two different proposals for the content of those parts of the node probability tables that would be required in order to deal with the observation case which is of greatest practical importance for the application. This is the ‘complete success’ case, which we shall term the *ideal observation* case for this 2-legged argument. For this, we suppose that the execution testing discovers *no failures at all* (*testing leg* success); and furthermore, the system is formally verified correct against its specification (*verification leg* success). This case is of practical importance in some safety-critical industries, where it is the *only* case which would allow acceptance of a system for operational use. In the UK nuclear industry, for example, failures in test would be unacceptable regardless of what could be inferred statistically from the test result.

---

<sup>1</sup> The terms *dependency model* or *Markov model* are sometimes used, synonymously.

Our two allocations of node probability tables used to analyse this ideal observations case are all parameterized, i.e. they specify the conditional probabilities of each node, given its parents' values numerically except for unspecified values of a set of independent model parameters. The simplest case here assumes that the verification process is infallible; the more complex (and more realistic) cases allows the verification to be fallible. In the first case there are 7 of these parameters. That is, there are only 7 parameters that are significant for the analysis of the ideal observation case. In the second case there are 12 independent model parameters. The first 7-parameter case is, in fact, a special case of the second 12-parameter case, which is presented as a relaxation of some key assumptions used to develop the former, simpler parameterisation.

This model is presented as a tentative, first experiment with BBN modelling of a 2-legged argument. We explore some of the model consequences graphically, concentrating on the key question of the probability, as modelled by this BBN structure, that the 2-legged argument will 'fail' dangerously: That is, the probability that an insufficiently safe system will be incorrectly accepted as adequately safe by this 2-legged argument. We compare the efficacy of this two-legged argument, on this criterion of 'dangerous failure', with that of a single argument leg.

## 2 BBN Topology and Model Variable Definitions

The construction of our model proceeds in the usual stages of: model variable identification, definition and respective state-space construction; BBN topology construction, to represent graphically assumed conditional independencies (CIs) among the model variables; and 'finally' local node conditional probability table definition. In this ordered presentation of the process, we should normally expect that the discoveries and difficulties encountered during later stages may well feed back into adjustments and refinements to earlier stages. There is a validation activity progressing throughout and in parallel with this model construction: We expect the consequences resulting from combined model assumptions often to cause us to revisit and question those assumptions, and perhaps to replace them with better assumptions, in successive iterations of the process here outlined.

The first stage in building a BBN is the identification of model variables. Our model variables are defined, with a certain amount of deliberate flexibility, as follows:

$S$  - the built *system* itself, or rather some abstract representation or property of it (whose value can never be known with complete confidence). In this paper,  $S$  is just the system's unknown, true probability of failure on demand

(pfd).

$Z$  - system *specification*. A system verification is performed directly against this specification, to form one leg of the system dependability argument.

$V$  - conclusion from the *verification* of the system against its specification.

For the purposes of the current investigation we shall be interested only in the *ideal observation* outcome that the system is verified correct.

$O$  - *oracle* used in system testing (by execution of the system). In this current BBN, we have for simplicity made the unrealistic assumption that the *operational profile* used to simulate the test inputs, is perfectly representative of the statistical pattern occurring during real use.

$T$  - system *test results*. We shall be interested here only in the *ideal observation* case of no failures.

$C$  - final *claim* as to whether or not the system is fit for use. For the purposes of illustration we shall take the claim here to be that the probability of failure on demand is better than  $10^{-3}$ . Of course,  $C$ 's two parents,  $T$  and  $V$ , are both *observable* nodes. So in one usage scenario we have in mind of this model,  $C$  itself will be a 'deterministic' node, in the sense that its realised value will be a chosen deterministic function of its parents' values.

The ultimate purpose of the following BBN model is to derive posterior distributions of random variables,  $S, C$  of *practical importance* (goal variables), following observation of other variables  $T, V$  whose values are *directly measurable* (or amenable to direct human assessment). The model, like most other BBN models, also includes essential model-structural 'mediating variables',  $Z, O$ , in this case, which fall into neither of these two categories. Formally, the BBN topology encodes a system of conditional independence assumptions, collectively termed a *Markov model*<sup>2</sup>, *(in)dependency relation*, *(in)dependency model*, or *Conditional Independence (CI) relation*, which enable the grand multivariate distribution of all model variables to be composed of several node conditional probability distributions of lower dimensionality. Consequently the required posterior distributions, of goal variables given evidence, can likewise be expressed, in §3 below, in terms of these node conditional distributions. The CI assumptions encoded in the topology justify the derivations involved, and the pictorial representation of this topology, properly understood, efficiently communicates these CI assumptions (some more evidently than others). So probabilistic CI assumptions and the BBN topologies that encode them are devices for constructing, communicating, and reasoning with multivariate probability models.

The theoretical study of how a probabilistic CI model may be precisely represented by a graph is based on an analogy of probabilistic CI relations with various notions of the separation of two sets of graph nodes by a third 'separator'

---

<sup>2</sup> An exhaustive enumeration of precisely for which triples  $\langle \mathcal{A}, \mathcal{B}, \mathcal{S} \rangle$  equation (1) below holds

set of nodes. The formal notions of ‘d-separation’, ‘graphoid’, ‘semi-graphoid’, and ‘I-map’ are central to this theory. See [4–7] for further details.

Our BBN topology is shown in Figure 1, where the left-hand figure is the

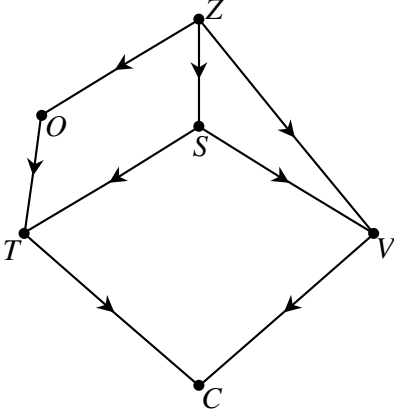


Fig. 1. BBN model topology

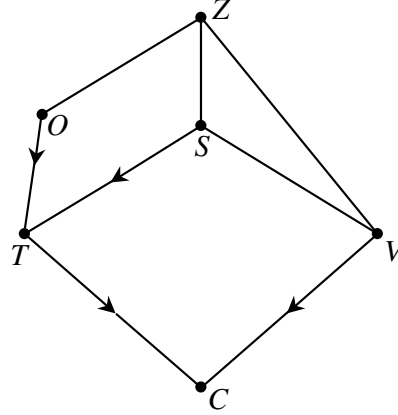


Fig. 2. Markov equivalent LCG

model as we originally formulated it, and the right-hand figure is the *largest chain graph* (LCG) canonical representation of this model’s Markov equivalence class (of distinct but equivalent graph topologies). The latter, Figure 2, is an alternative graphical representation of exactly the same ‘2-legged argument’ Markov model represented by Figure 1 and can serve as an aid to the Markov model’s understanding and verification. For example, all alternative BBN topologies that are ‘Markov equivalent’<sup>3</sup> to the BBN in Figure 1 may be derived from the LCG by a simple algorithm [7]. The LCG topology also exposes any Markov model symmetries<sup>4</sup> that are not necessarily retained graphically in the individual topology of every Markov equivalent BBN representation. (In the current case, Figure 2 simply confirms that our dependency model—like the topology, Figure 1, in which form we originally constructed it—possesses no symmetries<sup>5</sup>.)

Expressed algebraically, the ‘CI statement’ (or just ‘CI’, for short), “ $\mathcal{A}$  is conditionally independent of  $\mathcal{B}$ , given  $\mathcal{S}$ ”, can be thought of as a factorization property of the joint distribution function:

$$\mathcal{A} \perp\!\!\!\perp \mathcal{B} | \mathcal{S} \quad \text{means} \quad p(\mathcal{A}, \mathcal{B}, \mathcal{S}) = p(\mathcal{A} | \mathcal{S}) p(\mathcal{B} | \mathcal{S}) p(\mathcal{S}), \quad (1)$$

<sup>3</sup> i.e., that define exactly the same Markov model in the sense of footnote 2

<sup>4</sup> i.e. symmetries of the *dependency model* alone, which if present would not generally extend to a corresponding symmetry of *numerical probabilities* subsequently assigned; c.f. fuller comments in footnote 21 on p25.

<sup>5</sup> This is not always the case: For a very simple counterexample, see our asymmetric ‘verification leg’ BBN on p24 whose topology conceals a symmetry property of its Markov model. However in the *LCG*-form, the *automorphism group* of a *topology* (as nodes and edges) and of its *Markov model* (as a set of CIs) are always *equal*.

where  $\mathcal{A}, \mathcal{B}, \mathcal{S}$ , here represent *sets* of model *random variables*<sup>6</sup>. Thus, each CI assumption asserts that joint (or “joint conditional”) probability distributions will *factorize* – in precisely stated ways – into products of other *conditional* joint distributions, *each involving fewer variables*<sup>7</sup>. The present Markov model was constructed in the BBN form of Figure 1 by explicitly making the CI assumptions that each graph node is conditionally independent, given its parents, of its other non-descendants, although other CI statements are logical consequences of these. (See e.g. [4,8] for precise definitions and theory of how to use the net structure to determine these logical CI consequences entailed by the topology.) The term *conditional dependence* denotes simply the absence of a specified CI factorization.

The graph topology of our BBN can be thought of as an embodiment of a Markov model, i.e. a complete and consistent<sup>8</sup> set of conditional independence beliefs of an expert concerning the model random variables. There are many ways of specifying the dependency model which this graph topology represents. E.g. one economical, logically independent set of CI assumptions which specify the model is:

$$\begin{aligned} O \perp\!\!\!\perp SV \mid Z \\ T \perp\!\!\!\perp ZV \mid OS \\ C \perp\!\!\!\perp OSZ \mid VT \end{aligned}$$

Expressed in terms of its 26 *elementary CI statements*<sup>9</sup> as in [7], the same dependency model can be written

$$\begin{array}{ll} O \perp\!\!\!\perp S \mid Z(V) & \\ O \perp\!\!\!\perp V \mid Z(S) & O \perp\!\!\!\perp V \mid SZT(C) \\ Z \perp\!\!\!\perp T \mid OS & Z \perp\!\!\!\perp T \mid OSV(C) \\ V \perp\!\!\!\perp T \mid OS(Z) & V \perp\!\!\!\perp T \mid SZ \\ Z \perp\!\!\!\perp C \mid VT(OS) & Z \perp\!\!\!\perp C \mid OSV \\ O \perp\!\!\!\perp C \mid VT(SZ) & O \perp\!\!\!\perp C \mid SZT \\ S \perp\!\!\!\perp C \mid VT(OZ) & \end{array}$$

<sup>6</sup> Other forms, such as  $p(\mathcal{A}, \mathcal{B} \mid \mathcal{S}) = p(\mathcal{A} \mid \mathcal{S})p(\mathcal{B} \mid \mathcal{S})$ , are for most practical purposes equivalent; although there may be occasional exceptions relating to conditioning on zero-probability  $\mathcal{S}$ -events. Some authors use the very inclusive definition, avoiding potential zero divisions,  $p(\mathcal{A}, \mathcal{B}, \mathcal{S})p(\mathcal{S}) = p(\mathcal{A}, \mathcal{S})p(\mathcal{B}, \mathcal{S})$ .

<sup>7</sup> *Conditional* independence means distinct factors may contain common variables.

<sup>8</sup> No set of CI assertions can in itself exhibit logical inconsistency. The associated Markov model includes all the logical CI consequences of the explicitly asserted CI statements. We need all of these to be consistent with any conditional *dependence* beliefs expressed by the same expert.

<sup>9</sup> [7] show that *every* probabilistic dependency model is completely characterised by listing its elementary CI statements.

where a string of symbols in brackets to the RHS of the “|” can be replaced by any subset of those symbols (including both the entire set, and the empty set). E.g.  $S \perp\!\!\!\perp C | VT(OZ)$  means that all four of the CI statements

$$S \perp\!\!\!\perp C | VT \quad S \perp\!\!\!\perp C | VTO \quad S \perp\!\!\!\perp C | VTZ \quad S \perp\!\!\!\perp C | VTOZ$$

hold.

In later sections of the paper, we have obtained a set of initial computations of the consequences of this net using the following simple, initial, state-spaces for the 6 variables. Apart from variable  $S$ , the other 5 proposed state-spaces are essentially Boolean:

- $S$  - Here,  $S$  has the only continuous state-space, being the unit interval  $0 \leq S \leq 1$ , interpreted as the true pfd of the system<sup>10</sup>.
- $Z$  - correct; incorrect.
- $V$  - verified; not verified.
- $O$  - correct; incorrect
- $T$  - failures; no failures.
- $C$  - accepted; rejected. We can interpret ‘accepted’ as the claim that pfd  $S \leq 10^{-3}$  in operational use.

The idea captured by the ternary relation  $\mathcal{A} \perp\!\!\!\perp \mathcal{B} | \mathcal{S}$  can be expressed less formally by the statement: ‘*Observation of  $\mathcal{S}$  renders  $\mathcal{A}$  irrelevant to  $\mathcal{B}$* ’, [4]. A probabilistic uncertainty model implies an  $\mathcal{A} \leftrightarrow \mathcal{B}$ -symmetry of this statement. Care is required with its interpretation: The term “observation” denotes *complete* observation, in the sense that the values of all variables in  $\mathcal{S}$  should be made *exactly* known before a person interested (solely) in the values of  $\mathcal{B}$  will lose interest in information about variables  $\mathcal{A}$  [4,9,8,10–13], [6, §1.2.1]. For example, our model contains the assumption  $V \perp\!\!\!\perp T | SZ$ . A factorization such as

$$P(VT | S > 10^{-3}, Z = \text{correct}) = P(V | S > 10^{-3}, Z = \text{correct})P(T | S > 10^{-3}, Z = \text{correct})$$

does *not* follow; whereas

$$P(VT | S = 10^{-3}, Z = \text{correct}) = P(V | S = 10^{-3}, Z = \text{correct})P(T | S = 10^{-3}, Z = \text{correct})$$

is logically entailed within our model. This particular CI assumption also illustrates another important point about the above informal interpretation

---

<sup>10</sup>We must bear in mind that creating ‘bins’ (e.g.  $\text{pfd} > 10^{-3}$ ;  $0 < \text{pfd} \leq 10^{-3}$ ;  $\text{pfd} = 0$ .) to discretise a variable’s state-space, could have the effect of undermining the accuracy of some conditional independence assumptions. E.g. conditioning given an event “ $S > 10^{-3}$ ” may leave sufficient residual uncertainty about the exact value of  $S$  to induce a degree of conditional association between other variables which each have in common an association with the system pfd  $S$ , and are conditionally independent given *exact* knowledge of the value of  $S$ .

of CI assumptions. In practice we may still incorporate in our model such a CI assumption, even though we may not expect to observe variable(s)  $\mathcal{S}$ , or where  $\mathcal{S}$  may be in principle impossible to observe. Then it is *hypothetical* exact knowledge of  $\mathcal{S}$  that is the conceptual device used to interpret the above CI assumption. I.e., the CI assumption then asserts that the structure of our beliefs is such that, were we somehow to become completely assured of the exact value (*any* exact value) of  $\mathcal{S}$ , then such knowledge would, for us, ‘render  $\mathcal{A}$  irrelevant to  $\mathcal{B}$ ’—We can still hold the CI assumption even in cases where such certainty about  $\mathcal{S}$  could never arise. In practical model building, many CI assumptions may be of this form, in which the precise values of conditioning variables of some CI assumptions are never expected to be known, as with the assumption  $V \perp\!\!\!\perp T \mid SZ$  in our model. To state one last clarification about the interpretation of CI assumption  $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{S}$ , note that the conditioning knowledge-state, assumed to produce the irrelevance of  $\mathcal{A}$  to  $\mathcal{B}$ , consists of knowing *only* the exact value of  $\mathcal{S}$ . The irrelevance can later be destroyed by subsequently acquired knowledge, exact or approximate, of other variables. So strictly, assumption  $\mathcal{A} \perp\!\!\!\perp \mathcal{B} \mid \mathcal{S}$  says that when  $\mathcal{S}$  is exactly known,  $\mathcal{A}$  is irrelevant to  $\mathcal{B}$  for only just so long as the state of all other model variables remains completely unknown, i.e. while we do not discover *anything else than* the value of  $\mathcal{S}$ .

### 3 Computations from this BBN Topology

We are primarily interested in the updated joint probability distribution  $P(C, S \mid \text{observations})$ , particularly the value  $P(C = \textit{accepted}, S > 10^{-3} \mid \text{observations})$  representing an unsafe failure of the entire, two-legged assessment activity. Starting from this BBN model, the observations available will typically consist of values for the pair  $V, T$  of model variables. Under the conditional independence assumptions comprising this dependency model, the joint distribution of these four variables has a representation

$$P(C, S, V, T) = P(C \mid V, T) \left\{ \sum_Z P(V \mid SZ) P(S \mid Z) \sum_O P(T \mid OS) P(O, Z) \right\} \quad (2)$$

For any pair of observed values  $(V, T)$ , we obtain the desired, updated distribution for  $(C, S)$  by normalising

$$\begin{aligned}
P(C, S|V, T) &= \frac{P(C|V, T) \left\{ \sum_Z P(V|SZ)P(S|Z) \sum_O P(T|OS)P(O, Z) \right\}}{\sum_C \int_S P(C|V, T) \left\{ \sum_Z P(V|SZ)P(S|Z) \sum_O P(T|OS)P(O, Z) \right\} dS} \\
&= \frac{P(C|V, T) \left\{ \sum_Z P(V|SZ)P(S|Z) \sum_O P(T|OS)P(O, Z) \right\}}{\int_S \left\{ \sum_Z P(V|SZ)P(S|Z) \sum_O P(T|OS)P(O, Z) \right\} dS} \quad (3)
\end{aligned}$$

We have used integration over our single continuous model variable  $S$  here, interpreting  $P(S|Z)$  as a density function. In fact, notice that, if we are to accept that perfection ( $S=0$ ) is possible, then mixed distributions (including joint and conditional distributions) of  $S$  must be allowed. In this notation, this means thinking of integrands over  $S$  as potentially exhibiting ‘delta-function-like’ behaviour, at  $S=0$ <sup>11</sup>.

## 4 Effect of Some Strong Simplifying Assumptions

### 4.1 Node Probability Tables

We start with some assumptions that will simplify the mathematics. Some of these assumptions are quite strong. In later sections we consider first some parametric refinements (starting in §4.3), and then some relaxations (starting in §5) of some of these starting assumptions.

- An assumption, for our *ideal observations* case of success on both argument legs,  $(V, T) = (\text{verified}, \text{no failures})$ , which simplifies the above considerably, is the conditional probability table entry

$$P(C=\text{accepted} | (V, T)=(\text{verified}, \text{no failures})) = 1. \quad (4)$$

That is to say, the value of the claim  $C$  is fully determined by these observed values of  $(V, T)$  so that  $P(C=\text{accepted}, S | (V, T)=(\text{verified}, \text{no failures})) = P(S | (V, T)=(\text{verified}, \text{no failures}))$ . For this observed  $(V, T)$ , (2) and (3) then both become zero for  $C=\text{rejected}$ , while, for  $C=\text{accepted}$ , (2) loses the term to the left of the braces, as does the numerator of (3). Thus, substituting

<sup>11</sup> Or, more rigourously, we can represent (conditional) probabilities and expectations by the appropriate Lebesgue-Stieltjes integrals [14].

into (3) leaves us with the expression

$$P\left((C, S)=(\text{accepted}, s) \mid \text{ideal obs.}\right) = P\left(S=s \mid \text{ideal obs.}\right) = \frac{\sum_Z P(V=\text{verified} \mid S=s, Z)P(S=s \mid Z) \sum_O P(T=\text{no failures} \mid O, S=s)P(O, Z)}{\int_S \left\{ \sum_Z P(V=\text{verified} \mid S, Z)P(S \mid Z) \sum_O P(T=\text{no failures} \mid O, S)P(O, Z) \right\} dS} \quad (5)$$

for the updated probability density of  $S$ , given the *ideal observations* (for both argument legs). Here, the condition “ $\mid \text{ideal obs.}$ ” is a shorthand for “ $\mid (V, T)=(\text{verified}, \text{no failures})$ ”. Keep in mind that we can interpret the numerator and denominator of the conditional probability (5) as, respectively, an unconditional probability density, and an unconditional probability, in the usual way. See e.g. (11) on p13 below.

• **Infallible Verification Assumption:** Against a correct specification, the verification will be successful precisely if the true pfd is zero: Provided the specification is correct, any positive system pfd, however small, will always have been caused by a fault, which will certainly show up as a failure to verify the system. Conversely, all systems which fail the verification have non-zero true pfd:

$$P(V=\text{verified} \mid S=s, Z=\text{correct}) = \begin{cases} 0 & \text{if } 0 < s \leq 1, \text{ or} \\ 1 & \text{if } s = 0. \end{cases} \quad (6)$$

• **Conservative Assumption for Incorrect Specification:** Against an incorrect specification, we make the conservative (i.e. pessimistic, taking the perspective of the overriding undesirability of accepting a bad system) assumption that any system will always *pass* a verification against this specification.

$$P(V=\text{verified} \mid S=s, Z=\text{incorrect}) = 1. \quad (7)$$

• For the probability table of variable  $T$ , we assume that we can use the conclusions from existing modelling approaches to produce the probability that a period of testing uncovers zero failures, when the oracle is correct. This probability will be a known, function  $f(S)$  of the true system pfd,  $S$ , decreasing from  $f(0) = 1$  to  $f(1) = 0$ :

$$P(T=\text{no failures} \mid S=s, O=\text{correct}) = f(s). \quad (8)$$

In practice,  $f$  will depend on the duration of the testing, which we denote later in this paper by  $n$ , the number of discrete demands applied to the system

during the testing. The obvious candidate is the assumption of  $n$  independent trials, yielding the geometric distribution  $f(s) = (1 - s)^n$ .

• **Conservative Assumption for Incorrect Oracle:** To address the case of an incorrect oracle, we again adopt a conservative assumption, similar to that used above for the case of verification against an incorrect specification:

$$P(T=no\ failures \mid S=s, O=incorrect) = 1. \quad (9)$$

• We propose for the conditional distribution of  $S$  given  $Z$  a requirement for the following kind of stochastic ordering as a function of  $Z$

$$P\left(S > s \mid Z=correct\right) < P\left(S > s \mid Z=incorrect\right), \quad \text{for all } 0 \leq s < 1. \quad (10)$$

where it is allowed that either or both of these distributions can have mass concentrated at  $s=0$ , subject to this inequality.

#### 4.2 Effect of Node Probability Table Assumptions on Equation (5)

The above assumptions about the node probability tables can now be substituted in the *numerator of (5)*. This numerator is in fact simply the prior probability density of “ $S$  and *ideal observation on both argument legs*”. We will denote it  $P(S\&ideal\ obs.)$ , meaning, more precisely

$$P(s\&ideal\ obs.) = P\left((C, S, V, T)=(accepted, s, verified, no\ failures)\right). \quad (11)$$

Note that the case  $s > 0$  is much simplified by assumption (6)—the two 0’s in the body of Table 1 on p44. The table shows how the value of (11) is a sum of four terms corresponding to the different configurations of  $ZO$  — dealing separately with the case  $S=0$ . The four terms summed are each a product of three entries in the top part of a vertical column of the table<sup>12</sup>, weighted also by the prior probability of the value of the pair  $ZO$  which selects the column. The second to last row of the table gives the value of (11), for  $S=0$  on the left, and for  $0 < S \leq 1$  on the right. In the latter case, this probability is actually a density in its  $S$ -argument. We can think of the value “*ideal obs.*” as the vector of observed values of  $CVT$ , or effectively of just  $VT$ , since  $C$  is determined by  $VT$  in this ideal case (4).

<sup>12</sup>—in each single column, the three entries in the three rows immediately under the double line that separates the headers.

Notice also that, for these *ideal observations*  $VT = (\text{verified, no failures})$ , the  $S=0$ -cases of the four equations (6-9) produce the eight<sup>13</sup> 1's that occur in the left half of Table 1. These assumptions therefore mean simply that  $P(\text{ideal obs.} | S=0) = 1$ , i.e. that  $P(s \& \text{ideal obs.}) = P(s)$ , at  $s=0$ .

The last row of Table 1 on p44 gives two alternative representations for the *denominator of (5)*. This normaliser is just  $P(\text{ideal obs.}) = P((V, T) = (\text{verified, no failures}))$ . Substituting the entries of Table 1 into (5), gives the conditional probability  $P(S=0 \mid \text{ideal obs.})$  and also, for  $S>0$ , the conditional probability density pdf( $S \mid \text{ideal obs.}$ ), in each case as the ratio of expressions in the last two rows of the table. Care is necessary in use of notation for discontinuities and points of concentrated mass: Note the comments under the table.

#### 4.3 *First-Pass Attempt at Some Further Distributional Assumptions for the Conditional Probabilities*

First some shorter notation for those probabilities which will not now be substituted by a parametric distribution:

We see in Table 1 on p44 that some of the probabilities coming out of the net topology involve quite long expressions (even with only 6 variables in our model). There is commonly a conflict between making symbols long enough to be self-explanatory (less strain on memory), and making them very short (less requirement to read, parse, and typeset long expressions). We will use the symbol  $\pi$  for the unconditional joint distribution of variables  $ZO$ , taken in that order, with a first index to represent the  $Z$  value, and a second index to represent  $O$ . That is, we name four unconditional probabilities,  $\pi_{cc} + \pi_{ci} + \pi_{ic} + \pi_{ii} = 1$ , with  $c$  for correct,  $i$  incorrect. We also use a “wildcard notation”,  $*$ , for the marginal distributions of  $Z$  and of  $O$  so, for example,  $\pi_{c*} = \pi_{cc} + \pi_{ci} = P(Z = \text{correct})$ . Later, we will sometimes display these prior probabilities of

---

<sup>13</sup> Note how just four equations produce eight 1s here essentially because the topology says that  $T \perp\!\!\!\perp Z \mid OS$  and  $V \perp\!\!\!\perp O \mid ZS$ . So, reading across the first row, the conditional probabilities of  $T$  alternate, in each half of the table; and reading across the third row, the conditional probabilities of  $V$  occur in adjacent pairs.

variables  $ZO$  using a  $2 \times 2$  matrix layout

$$\begin{array}{cc}
 & O \\
 & \text{correct} \quad \text{incorrect} \\
 Z \text{ correct} & \pi_{cc} \quad \pi_{ci} \quad | \quad \pi_{c*} \\
 \text{incorrect} & \pi_{ic} \quad \pi_{ii} \quad | \quad \pi_{i*} \\
 \hline
 & \pi_{*c} \quad \pi_{*i} \quad | \quad 1
 \end{array} \tag{12}$$

For the conditional distribution  $P(S|Z)$ , we have the complication that it may be a mixed distribution. In our parametric examples we assume this to be continuous on  $S \in [0, 1]$  except for a possible concentrated mass at  $S=0$ . Denote the two concentrated masses  $p_{0c}$  and  $p_{0i}$ , where the  $c$  or  $i$  indicates the conditioning value of  $Z$ . (We will require  $p_{0c} > p_{0i}$  in compliance with assumption (10).) For the sake of statistical conjugacy [15, Ch. 9] with the discrete time model  $f(s) = (1-s)^n$  in (8), we will try a beta distribution for the continuous component. So for  $0 < S \leq 1$ , use

$$\text{pdf}(S|Z=\textit{incorrect}) = \frac{(1 - p_{0i})}{\beta(a, b)} S^{a-1} (1 - S)^{b-1}, \quad \text{for some } a, b > 0. \tag{13}$$

To analyse our ‘ideal observations’ case, we do not at present need a notation to distinguish the conditional pdf (13) from the corresponding quantity conditioned on “ $Z=\textit{correct}$ ” because this latter conditional pdf is ‘masked out’, by assumption (6), from the distributions (3,5,11) derived from this BBN topology<sup>14</sup>. See columns 5 and 6 (of 8) in Table 1 on p44. With this  $f(s)$ , one interesting figure to use for the number  $n$  of test-cases in (8)

$$P(T=\textit{no failures} | S=s, O=\textit{correct}) = (1 - s)^n, \tag{14}$$

is  $n = 4,602$ . This is the number of tests required to give a Bayesian 99% upper confidence bound on the pfd  $S$  that is less than  $10^{-3}$  when no failures

<sup>14</sup>This masking from model consequences of any assumption for the “positive- $S$  component” of the distribution  $P(S | Z=\textit{correct})$  is a simplifying effect of the zero value assumed in (6). We will see later that if this verification infallibility assumption (of zero probability for the verification of an imperfect system against a correct specification) is dropped, then an additional analytic component is introduced into this model. It then becomes necessary to produce a *pair* of distributions here: one conditioned on  $Z = \textit{incorrect}$ , as (13), and the other to cover the case  $Z = \textit{correct}$ . In particular, this additional complication is even present when variable  $V$  is altogether *removed* from the model – i.e. where we consider a single, testing-leg-only safety case argument.

are observed, under the simpler model assumptions used in [16]. In terms of our model here, those assumptions say essentially that there is no verification leg, that the oracle is known to behave perfectly, and that the prior distribution of  $Z$  and the conditional distribution  $P(S|Z)$  are such that  $S$  is initially uniformly distributed on the unit interval. The adjacent value of  $n=4,603$  had previously been shown to be the number of tests required to obtain a similar-valued, *classical* 99% upper confidence bound on  $S$  (using no prior distribution for  $S$ ).

Substituting these further assumptions and abbreviated notations into the penultimate row of Table 1 on p44 (which originated from the numerator of (5)) gives the concentrated mass

$$P(S=0 \ \& \ \textit{ideal obs.}) = P(S=0) = p_{0c} \pi_{c*} + p_{0i} \pi_{i*} \quad (15)$$

and, for strictly positive values of  $S$ , the pdf

$$\text{pdf}(S \ \& \ \textit{ideal obs.}) = \frac{(1 - p_{0i})}{\beta(a, b)} S^{a-1} (1-S)^{b-1} [\pi_{ii} + \pi_{ic} (1-S)^n], \quad 0 < s \leq 1. \quad (16)$$

To obtain the conditional distribution of  $S$  given the ideal observations, we require a normaliser from the above joint probability density (and point mass) corresponding to the last row of Table 1

$$P(\textit{ideal obs.}) = P(S=0 \ \& \ \textit{ideal obs.}) + \int_{0 < S \leq 1} \text{pdf}(S \ \& \ \textit{ideal obs.}) dS \quad (17)$$

where the left-hand term is the point mass contribution, which is understood to be excluded from the domain of the right-hand, integral term. Substituting from (15) and (16) yields

$$\begin{aligned} P(\textit{ideal obs.}) &= p_{0c} \pi_{c*} + p_{0i} \pi_{i*} + (1 - p_{0i}) \pi_{ii} + (1 - p_{0i}) \pi_{ic} \frac{\beta(a, b+n)}{\beta(a, b)} \\ &= p_{0c} \pi_{c*} + p_{0i} \pi_{ic} + \pi_{ii} + (1 - p_{0i}) \pi_{ic} \frac{\beta(a, b+n)}{\beta(a, b)} \end{aligned} \quad (18)$$

The last fraction  $\frac{\beta(a, b+n)}{\beta(a, b)}$  is the  $n^{\text{th}}$  non-central moment of the beta distribution<sup>15</sup>. We will shorten some expressions in what follows by using notations

$$\mu = \frac{\beta(a, b+n)}{\beta(a, b)}, \quad \mu' = \frac{\beta(a', b'+n)}{\beta(a', b')} \quad (19)$$

for these. Note the dependence of  $\mu$  and  $\mu'$  on  $n$ , as well as the beta distribution parameters.

<sup>15</sup>—to be precise, of the beta distribution with its usual parameters  $a$  and  $b$  interchanged

From (16) and (18), we can express the probability of ‘unsafe failure’, of the entire two-legged assessment procedure represented by these modelling assumptions in a variety of ways, with the formulas

$$\begin{aligned}
P(S>s \mid \text{ideal obs.}) &= \frac{\int_s^1 \text{pdf}(S \ \& \ \text{ideal obs.}) \, dS}{P(\text{ideal obs.})} \\
&= \frac{\int_s^1 \frac{(1-p_{oi})}{\beta(a,b)} S^{a-1} (1-S)^{b-1} [\pi_{ii} + \pi_{ic}(1-S)^n] \, dS}{p_{oc}\pi_{c*} + p_{oi}\pi_{ic} + \pi_{ii} + (1-p_{oi})\pi_{ic}\mu} \\
&= \frac{(1-p_{oi}) [\pi_{ii}I_{1-s}(b,a) + \pi_{ic}\mu I_{1-s}(b+n,a)]}{p_{oc}\pi_{c*} + p_{oi}\pi_{ic} + \pi_{ii} + (1-p_{oi})\pi_{ic}\mu} \quad (20) \\
&= \frac{\pi_{ii}I_{1-s}(b,a) + \pi_{ic}\mu I_{1-s}(b+n,a)}{\frac{p_{oc}\pi_{c*} + p_{oi}\pi_{i*}}{1-p_{oi}} + \pi_{ii} + \pi_{ic}\mu}
\end{aligned}$$

where  $I_{1-s}$  denotes the *incomplete beta function* using the notation

$$I_x(a,b) = \frac{1}{\beta(a,b)} \int_0^x u^{a-1} (1-u)^{b-1} \, du, \quad a, b > 0, \quad 0 \leq x \leq 1, \quad (21)$$

which is a strictly increasing function of  $x \in [0, 1]$  for all  $a, b > 0$ . [See [17, p944] for its other properties, which include  $I_0(a,b) = 0$ ,  $I_1(a,b) = 1$ , and  $I_{1-x}(a,b) + I_x(b,a) = 1$ .] Equations (20), with the strict inequality in the probability on the left hand side, actually hold for all (non-negative)  $s$ , including  $s=0$ .

Notice one combined effect of several of our strong simplifying assumptions: that the posterior probability distribution (20) of  $S$  given the ideal observations depends on only 2 of the 3 independent degrees of freedom of the prior  $ZO$  distribution  $\pi$ . The conditional probability (20) depends on the marginal probability  $P(Z)$  and on the conditional probability  $P(O|Z=\text{incorrect})$ . Its lack of dependence, in this *ideal observations* case, on  $P(O|Z=\text{correct})$  is explained by the fact that the infallibility assumption for the verification process (specifically the  $S>0$  part of this, given by the top line of (6)) creates the two zeros in the 5<sup>th</sup> and 6<sup>th</sup> columns of Table 1 which, by multiplication, effectively ‘mask out’ from the final rows of Table 1 (i.e., from (2) and the numerator and denominator of equations (3) & (5)) the *only* case of dependence of any term in these equations on the state of variable  $O$  when  $Z = \text{correct}$ . (The pair of entries in Table 1 located two rows above this pair of zeros are not equal.) The fact that this masked-out ( $S>0, Z = \text{correct}$ ) scenario *is* the only means, under ideal observations, by which our posterior distribution of  $S$  could otherwise (without this masking effect of assumption (6)) depend on the state of  $O$  relies on the conservative assumption (9) which removes any dependence on  $O$  in the LHS of Table 1 where  $S=0$ . As soon as we begin to relax any one of the infallibility and conservative assumptions which interact in this simpli-

fying way here, as we will do later, the conditional probability (20) will have to be replaced by a more complex expression, involving all the parameters of the parametric node probabilities.

To look briefly at some actual numbers, if we set

$$(p_{0i}, a, b, p_{0c}, n, \pi_{c*}) = (0.2, 1, 999, 0.5, 4602, 0.8). \quad (22)$$

in (20), we obtain

$$P(S>s | ideal obs.) = \frac{\pi_{ii}I_{1-s}(999, 1) + 0.178\pi_{ic}I_{1-s}(999+4602, 1)}{0.55 + \pi_{ii} + 0.178\pi_{ic}}$$

where the incomplete beta term  $I_{1-s}(999, 1)$  may be thought of as the corresponding probability  $P(X>s)$  for a random variable  $X$  which is beta distributed with parameters  $(a, b)=(1, 999)$ ,

$$P(X>s) = \frac{\int_s^1 u^{a-1}(1-u)^{b-1} du}{\beta(a, b)} = 1 - I_s(a, b) = I_{1-s}(b, a).$$

The same applies to the other incomplete beta term, but instead with parameters  $(a, b)=(1, 5601)$ . The latter is a smaller beta probability – considerably smaller for many  $s$ -values: The respective means of these two beta distributions being 0.001 and 0.00018. Under the assumptions of this section, it makes no difference to (20) how the probability of specification correctness,  $\pi_{c*} = 0.8$  is divided between the probabilities of  $(Z, O) = (correct, correct)$  and  $(Z, O) = (correct, incorrect)$ . However, the distribution of the other 20% of prior belief,  $\pi_{i*} = 0.2$ , does make a difference, and in fact can be used to shift the result considerably as it is differently allocated between  $(Z, O) = (incorrect, correct)$  and  $(Z, O) = (incorrect, incorrect)$ . It is easy to see that the two extreme cases  $(\pi_{ic}, \pi_{ii}) = (0.2, 0)$  and  $(\pi_{ic}, \pi_{ii}) = (0, 0.2)$  result in probabilities of

$$P(S>s | ideal obs.) = \frac{0.178 \times 0.2 I_{1-s}(5601, 1)}{0.55 + 0.178 \times 0.2}, \quad \text{and} \quad \frac{0.2 I_{1-s}(999, 1)}{0.55 + 0.2},$$

respectively, while values of  $(\pi_{ic}, \pi_{ii})$  between these two extremes lead to values intermediate between these two, producing a curve which is hyperbolic in form and monotonic increasing<sup>16</sup> as the 0.2 probability mass is steadily shifted from  $\pi_{ic}$  to  $\pi_{ii}$ . For instance, if we put  $s=10^{-3}$ , these two end points of this increasing hyperbolic segment are  $P(S>s | ideal obs.) = 0.00022$  at  $\pi_{ic}=0.2$ , and  $P(S>s | ideal obs.) = 0.098$  at  $\pi_{ii}=0.2$

These numbers illustrate how the joint prior beliefs concerning the two unobservable variables  $ZO$  – incorporating both the prior confidence in  $Z$  and  $O$  individually, as well as beliefs about the association between their likely

<sup>16</sup> assuming  $s \neq 0, 1$

values – can influence the confidence in the claim produced by the 2-legged argument. Although the model we are using is very simplified at this stage, we expect that the prior beliefs represented at ‘the top part’ of our BBN topology may well continue, under more sophisticated and realistic models, to be a driver for the amount of extra confidence gained when safety arguments are combined in this kind of formal way.

## 5 Introducing Fallibility of Verification Process

Until now we have made very simple assumptions (6,7) about the node probability table for  $V$  given  $SZ$ . Specifically, all entries in the third row of Table 1 on p44 are zeros and ones. These eight entries in fact – given the CI assumptions of the model topology – may be taken in pairs, reading across, and actually represent only four CI tables entries (because of the lack of dependence on variable  $O$ , the oracle correctness). The 3<sup>rd</sup>, 4<sup>th</sup>, 7<sup>th</sup>, and 8<sup>th</sup>, entries in this row of the table are justified as a single conservative assumption to cover the case of an incorrect system specification. Here we have assumed pessimistically, with equation (7), that the verification against such an incorrect yardstick will always produce a positive conclusion about any built system, irrespective of its actual failure probability  $S$ . We will retain this conservative assumption below. In contrast, for the other case—that the specification  $Z$  is correct—the remaining four entries of the same row are equivalent to two kinds of ‘infallibility’ assumption for the verification activity. We cannot justify these from an argument for conservatism. The assumption represented by the 5<sup>th</sup> and 6<sup>th</sup> columns can even be viewed as over optimistic, depending on how we define correctness of a specification, and how the verification is carried out in practice. These two assumptions – see equation (6) – state that two kinds of verification process failure are both impossible, against a correct specification. These two possibilities which we exclude above are that, against a correct specification:

- a perfectly reliable system could ever fail<sup>17</sup> the verification; or
- a system having a positive probability of failure – however small – could pass the verification.

We will now relax both of these assumptions, assuming instead only that the probability (that is, conditionally given  $\langle S = s, Z = \textit{correct} \rangle$ ) of the latter kind

---

<sup>17</sup>We have to be careful about terminology here: Surely there may be systems which contain ‘faults’ while being perfectly reliable. (E.g. defective functionality may not be exercised by a particular operational profile; or the system may contain internal fault-tolerance which eliminates the possibility that a certain ‘fault’ could ever result in a system failure.)

of verification failure is independent of the actual (positive) value  $s$  of variable  $S$ . (Of course this new constraint may itself be relaxed later.) If we assign notation  $\alpha, \xi$  for the probabilities of these two kinds of failure, replacing (6) by

$$P(V=verified|S=s, Z=correct) = \begin{cases} \xi & \text{if } 0 < s \leq 1, \text{ or} \\ 1 - \alpha & \text{if } s = 0, \end{cases} \quad (23)$$

then we are no longer able to ignore, in our ‘ideal observations’ scenario, the case  $S > 0$  given  $Z = correct$  of the node probability table for variable  $S$ . (So our stochastic ordering assumption (10) now becomes relevant as a constraint on any specific distributional assumptions we wish to make.)

These less simplistic assumptions complicate the model consequences. Table 2 on p45 now replaces Table 1 on p44. Equations (15-16, 18-20) are all replaced by more complicated versions, reflecting the newly introduced possibilities, under a correct system specification, for a mismatch between the perfection, or otherwise, of the system, and the outcome of the verification activity:

$$P(S=0 \ \& \ ideal \ obs.) = (1-\alpha)p_{0c}\pi_{c*} + p_{0i}\pi_{i*} \quad (24)$$

and, if we define a counterpart

$$\text{pdf}(S|Z=correct) = \frac{(1-p_{0c})}{\beta(a', b')} S^{a'-1} (1-S)^{b'-1}, \quad \text{for some } a', b' > 0 \quad (25)$$

of definition (13), for the continuous component of the distribution of failure rate  $S$  given a *correct* specification (having parameter assignments to  $a', b'$  subject to our assumption (10), which translates into a messy but computable constraint determined by  $a, b, p_{0i}, p_{0c}$ ), then we obtain for *strictly positive* values of  $S$

$$\begin{aligned} \text{pdf}(S \ \& \ ideal \ obs.) = \\ \xi \frac{(1-p_{0c})}{\beta(a', b')} S^{a'-1} (1-S)^{b'-1} [\pi_{cc}(1-S)^n + \pi_{ci}] + \frac{(1-p_{0i})}{\beta(a, b)} S^{a-1} (1-S)^{b-1} [\pi_{ic}(1-S)^n + \pi_{ii}], \\ 0 < s \leq 1. \end{aligned} \quad (26)$$

Combining (24) and (26) in the same manner as for the simpler ‘infallible-verification’ case gives now

$$P(ideal \ obs.) = (1-\alpha)p_{0c}\pi_{c*} + p_{0i}\pi_{i*} + \xi(1-p_{0c}) [\pi_{cc}\mu' + \pi_{ci}] + (1-p_{0i}) [\pi_{ic}\mu + \pi_{ii}] \quad (27)$$

Finally, integrating (26) from  $s$  to 1, and then normalising with (27), just as in the derivation of (20) above, yields a similar, if more complicated, function for the probability of unacceptably high pdf  $S$  in the case of the ideal observations,

again involving incomplete beta functions. For  $0 \leq s \leq 1$ ,

$$P(S > s \mid \text{ideal obs.}) = \frac{\xi(1-p_{0c}) [\pi_{cc}\mu' I_{1-s}(b'+n, a') + \pi_{ci} I_{1-s}(b', a')] + (1-p_{0i}) [\pi_{ic}\mu I_{1-s}(b+n, a) + \pi_{ii} I_{1-s}(b, a)]}{(1-\alpha)p_{0c}\pi_{c*} + p_{0i}\pi_{i*} + \xi(1-p_{0c}) [\pi_{cc}\mu' + \pi_{ci}] + (1-p_{0i}) [\pi_{ic}\mu + \pi_{ii}]} \quad (28)$$

Notice that equations (15,16,18 & 20) are the  $(\alpha, \xi)=(0, 0)$ -cases, respectively, of (24,26,27 & 28). We shall refer to the conditional probability (28) in what follows as the ‘doubt function’. This function of 13 independent arguments (the threshold  $s$  value; and the 12 independent parameters of our node conditional probabilities) is an important consequence of our model as it stands, capturing the probability<sup>18</sup> of the most important kind of ‘argument failure’ we first identified on p4.

If  $s$  is small, it may sometimes<sup>19</sup> be numerically more accurate to compute instead

$$P(S \leq s \mid \text{ideal obs.}) = P(S=0 \mid \text{ideal obs.}) + P(0 < S \leq s \mid \text{ideal obs.}) = \frac{(1-\alpha)p_{0c}\pi_{c*} + p_{0i}\pi_{i*}}{(1-\alpha)p_{0c}\pi_{c*} + p_{0i}\pi_{i*} + \xi(1-p_{0c}) [\pi_{cc}\mu' + \pi_{ci}] + (1-p_{0i}) [\pi_{ic}\mu + \pi_{ii}]} + \frac{\xi(1-p_{0c}) [\pi_{cc}\mu' I_s(a', b'+n) + \pi_{ci} I_s(a', b')] + (1-p_{0i}) [\pi_{ic}\mu I_s(a, b+n) + \pi_{ii} I_s(a, b)]}{(1-\alpha)p_{0c}\pi_{c*} + p_{0i}\pi_{i*} + \xi(1-p_{0c}) [\pi_{cc}\mu' + \pi_{ci}] + (1-p_{0i}) [\pi_{ic}\mu + \pi_{ii}]} \quad (29)$$

<sup>18</sup> There are several related measures that could be substituted here. To be exact, this one is the *conditional* probability that such an unsafe argument failure has occurred, given that a system is deemed *accepted* by the two-legged argument. Of course this is distinct from the unconditional probability that a randomly selected system will be truly unsafe ( $S > 10^{-3} = s$ ) and will be (incorrectly) accepted by the 2-legged argument as sufficiently safe (the *numerator* of (28)). It is also different from the probability that a randomly selected *unsafe* system will be deemed sufficiently safe by the two-legged argument. The conversions between these are straightforward, depending on quantities such as the ‘base rate’ of truly unsafe systems among systems which are submitted for evaluation by this procedure, and the rates at which rejections of randomly submitted systems will occur. Note that in our model as presently formulated, these marginal background rates are not outside the scope of the model. They *are* implied by our chosen values for model parameters: in the first case (the marginal probability  $P(S > 10^{-3})$ ) by  $\pi$ ’s, and  $p_{0i}, a, b, p_{0c}, a', b'$ ; and in the latter case (the marginal probability  $P(C = \text{rejected})$ ) by the entire set of 12 independent model parameters.

<sup>19</sup> depending partly on the precision properties of whatever algorithm one has available for computing the incomplete beta function in its extreme argument regions

## 6 A Mathematically Simpler Special Case

The above conclusions appear quite complicated, and lead to some computational difficulties, particularly for the very small values of  $s$  and of  $\frac{a}{b}$ , and  $\frac{a'}{b'}$  that often arise<sup>20</sup> in the safety context. Treating an arguably plausible special case of (13) and (25) can simplify the mathematics and avoid the computational problems associated with the incomplete beta function. This case is the assumption

$$\text{pdf}(S|S>0, Z) = 1,$$

i.e., that, given  $Z$ ,  $S$  is either zero, with a finite conditional probability, or otherwise is conditionally, uniformly distributed over the positive unit interval  $(0, 1]$ . This is equivalent to the substitutions  $a=b=a'=b'=1$  in equations (13) and (25) and all of their consequences. Thus (13) and (25) are replaced by

$$P(S=0|Z) = \begin{cases} p_{0c} & \text{if } Z = \text{correct, or} \\ p_{0i} & \text{if } Z = \text{incorrect,} \end{cases} \quad (30)$$

and, for strictly positive  $0 < S \leq 1$

$$\text{pdf}(S|Z) = \begin{cases} 1-p_{0c} & \text{if } Z = \text{correct, or} \\ 1-p_{0i} & \text{if } Z = \text{incorrect,} \end{cases} \quad (31)$$

where we must continue to assume  $p_{0c} > p_{0i}$  in compliance with (10). It follows that  $\mu = \mu' = \frac{1}{n+1}$ , which, when substituted with  $I_x(n+1, 1) = x^{n+1}$ , ( $n \geq 0$ ), yields

$$\begin{aligned} P(S > s | \text{ideal obs.}) = & \\ \frac{\xi(1-p_{0c}) [\pi_{cc}(1-s)^{n+1} + \pi_{ci}(n+1)(1-s)] + (1-p_{0i}) [\pi_{ic}(1-s)^{n+1} + \pi_{ii}(n+1)(1-s)]}{(n+1) [(1-\alpha)p_{0c}\pi_{c*} + p_{0i}\pi_{i*}] + \xi(1-p_{0c}) [\pi_{cc} + (n+1)\pi_{ci}] + (1-p_{0i}) [\pi_{ic} + (n+1)\pi_{ii}]} & \quad (32) \end{aligned}$$

as this special case of (28). This 9-parameter (including  $s$ ) special case provides a computationally easy sanity check on the more general conclusion (28) of our modelling assumptions. Under the infallible verification assumptions  $(\alpha, \xi) =$

<sup>20</sup>  $\frac{a'}{a'+b'}$  is the beta distribution mean: the expected value  $E(S|S>0, Z=\text{correct})$  of  $S$ , conditionally given that the specification is correct and that  $S$  is not zero (in the absence of any evidence  $V, T$  from either argument leg). For a high integrity system, this expectation may arguably be quite small.

(0, 0) we would obtain the still simpler

$$P(S > s | \textit{ideal obs.}) = \frac{(1-p_{0\bar{i}}) [\pi_{ic}(1-s)^{n+1} + \pi_{ii}(n+1)(1-s)]}{(n+1) [p_{0c}\pi_{c*} + p_{0\bar{i}}\pi_{i*}] + (1-p_{0\bar{i}}) [\pi_{ic} + (n+1)\pi_{ii}]} \quad (33)$$

which is the corresponding special case of (20) on p17 of §4.3.

## 7 How effective is this multi-legged argument approach in gaining confidence in dependability claims?

The use of multiple argument legs arises informally from reasoning similar to that used to justify the use of diverse channel redundancy to obtain system reliability. Questions that are of interest for systems have similar counterparts here. How much of a confidence gain do we obtain, via the above two-legged argument model, over the verification-only argument model, or the testing-only argument model? Equation (28) expresses the monotonic ‘doubt function’  $P(S > s | \textit{ideal obs.})$  of  $s$  as a function of the twelve independent model variables:  $\pi_{ic}, \pi_{ii}, \pi_{cc}, p_{0c}, p_{0\bar{i}}, a, b, a', b', \alpha, \xi, n$ : how can we best gain an understanding of the ‘shape’ of this functional dependence? What are the important drivers of the benefit coming from the use of multiple legs? E.g. does correlation between doubts in the assumptions play an important role, as intuition would suggest? Is there some formal version of a ‘naive independence’ argument claiming to combine claim confidences from single argument legs to produce a value which we can then compare with our ‘not quite so naive’ two-legged argument (28)?

### 7.1 Comparison of Two-legged Arguments Against Each Single Leg Alone

Here are the BBNs representing the obvious versions of the two, single-legged arguments whose combination we have discussed up to now, in the form of the BBN model shown in Figure 1:

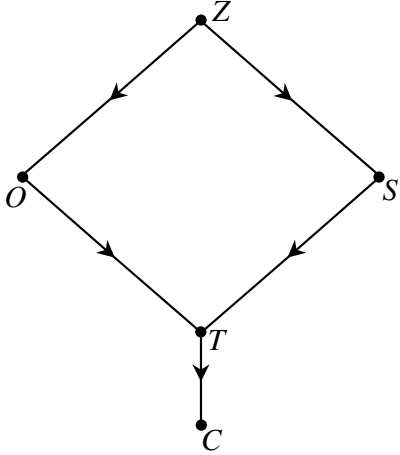


Fig. 3. Testing leg BBN topology

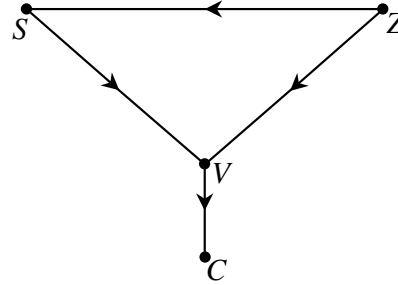


Fig. 4. Verification leg BBN topology

These topologies are respectively obtained from Figure 1 by assuming no observational information is available at  $V$ , or at  $T$ . This is done by the equivalent of assuming a state space of only one value, either at  $V$ , or at  $T$ . These two dependency models also can be represented canonically by their two LCG instantiations:

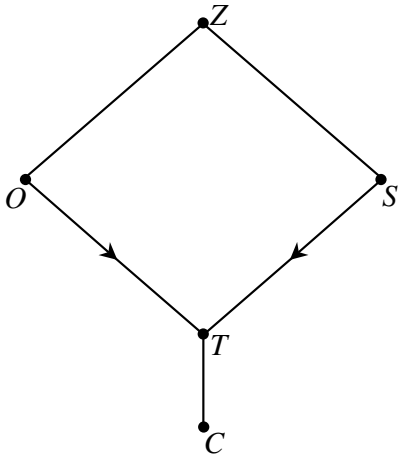


Fig. 5. Testing leg in LCG-form

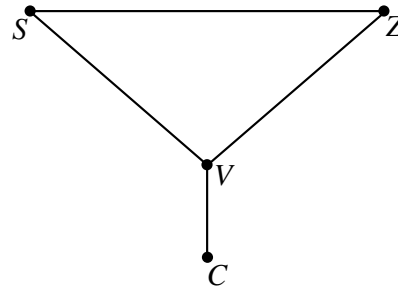


Fig. 6. Verification leg in LCG-form

Notice that our *dependency model* for the *testing leg* is symmetric under the transposition of  $O$  and  $S$ . The *verification leg* dependency model is symmetric under the transposition of  $S$  and  $Z$ <sup>21</sup> Represented algebraically, the two

<sup>21</sup>To avoid misinterpretation here, note that the *probabilistic* models that can be based upon these BBN topologies are not constrained to retain these symmetries in their individual *numerical (conditional) probability values*: it is only the *Markov model* – i.e., the *factorization properties* of the functions assigning these proba-

dependency models are

<u>Testing Leg</u>	<u>Verification Leg</u>
$O \perp\!\!\!\perp S \mid Z$	$C \perp\!\!\!\perp SZ \mid V$
$T \perp\!\!\!\perp Z \mid OS$	
$C \perp\!\!\!\perp OSZ \mid T$	

or, in the form of exhaustive lists of elementary CI statements<sup>22</sup>:

<u>Testing Leg</u>	<u>Verification Leg</u>
$O \perp\!\!\!\perp S \mid Z$	
$Z \perp\!\!\!\perp T \mid OS(C)$	
$Z \perp\!\!\!\perp C \mid T(OS)$	
$Z \perp\!\!\!\perp C \mid OS$	$Z \perp\!\!\!\perp C \mid V(S)$
$O \perp\!\!\!\perp C \mid T(SZ)$	
$S \perp\!\!\!\perp C \mid T(OZ)$	$S \perp\!\!\!\perp C \mid V(Z)$

Notice that comparison of these lists with the one on p9 demonstrates that, despite our relatively straightforward derivation of the single-legged argument dependency models from the 2-legged model, there is no simply defined formal “reverse operation” allowing composition of the two single-legged dependency models to create the 2-legged model. The 2-legged model of Figure 1 is a construction embodying various (first-pass) subjective beliefs about the dependencies arising in practice among the variables of our particular application. In particular, our 2-legged dependency model can be characterised neither as weaker nor as stronger than the simple conjunction<sup>23</sup> of the 2 separate single-legged dependency models. Precisely, using the elementary CI statement dependency model characterization, the 2-legged model deletes all CIs from one single-legged model, most CIs from the other, and of course introduces several other CIs whose contexts<sup>24</sup> are not contained in the variable set

---

bilities – to which the stated symmetries must apply. For example, variables  $S$  and  $Z$  in Figure 6 are interchangeable in terms of the underlying factorization  $P(S, Z, V, C) = f(S, Z, V)g(V, C)$  required by the BBN, but are not even required to have state-spaces of equal cardinality, let alone be interchangeable as arguments of the numerical conditional probability distributions defined on those state spaces.

<sup>22</sup> again using the bracket shorthand on the RHS of the “|” as explained on p9

<sup>23</sup> with *nothing* further assumed about conditional independencies (or their absence) involving the pooled set of 6 variables

<sup>24</sup> the *context* of a CI statement is the set of all the model variables it involves:  $\text{context}(A \perp\!\!\!\perp B \mid C) = A \cup B \cup C$ .

for either single leg.

<u>Deleted from Testing Leg</u>	<u>Deleted from Verification Leg</u>	<u>Appended</u>
		$O \perp S   ZV$
		$O \perp V   Z(S)$
		$O \perp V   SZT(C)$
$Z \perp T   OSC$		$Z \perp T   OSV(C)$
		$V \perp T   OS(Z)$
		$V \perp T   SZ$
$Z \perp C   T(OS)$	$Z \perp C   V(S)$	$Z \perp C   OSV$
$Z \perp C   OS$		$Z \perp C   VT(OS)$
$O \perp C   T(S)$		$O \perp C   VT(SZ)$
$O \perp C   TZ$		
$S \perp C   T(OZ)$	$S \perp C   V(Z)$	$S \perp C   VT(OZ)$

It is possible to follow through, in each of these two single-legged arguments, similar derivations of the two parametric representations of the conditional probability of “dangerous argument failure”, just as we have done above for the two-leg argument, to produce the doubt function (28). However, it is easier to obtain the analogous consequences of the single-leg arguments of Figures 3 & 4 as special cases of the results for the combined argument. The removal of testing evidence from the argument is equivalent to substituting  $n = 0$  in (28) to produce

$$P(S > s | V = \text{verified}) = \frac{\xi(1-p_{0c})\pi_{c*}I_{1-s}(b', a') + (1-p_{0i})\pi_{i*}I_{1-s}(b, a)}{[(1-\alpha)p_{0c} + \xi(1-p_{0c})]\pi_{c*} + \pi_{i*}} \quad (34)$$

as the doubt function for the verification-only argument. (Unsurprisingly, the initial beliefs about the likely oracle-correctness are not present in this expression.)

Similarly, for the testing-only argument, we can substitute  $\alpha = 0$ ,  $\xi = 1$  in (28) to produce

$$P(S > s | T = \text{no failures}) = \frac{(1-p_{0c})[\pi_{cc}\mu' I_{1-s}(b'+n, a') + \pi_{ci}I_{1-s}(b', a')] + (1-p_{0i})[\pi_{ic}\mu I_{1-s}(b+n, a) + \pi_{ii}I_{1-s}(b, a)]}{p_{0c}\pi_{c*} + p_{0i}\pi_{i*} + (1-p_{0c})[\pi_{cc}\mu' + \pi_{ci}] + (1-p_{0i})[\pi_{ic}\mu + \pi_{ii}]} \quad (35)$$

for our *ideal observations* case of the single, testing-argument leg. To see this<sup>25</sup>, consider the third row (of 5 rows) in the body of Table 2 on p45. The above substitutions make all entries in this row equal. That is to say,

<sup>25</sup> This  $(\alpha, \xi) = (0, 1)$  substitution procedure would not produce the single testing leg model if we relaxed the conservative assumption of equation (7) .

the substitutions make  $P(V|ZS)$  independent of the values of both  $Z$  and  $S$ , which – in terms of the topology – altogether ‘detaches’ variable  $V$  from the remainder of the model. It is not difficult to confirm from our algebraic conclusions (consider equations (3) and (5)) that the effect of this is equivalent to the removal of node  $V$  from the model, i.e. the conversion of Figure 1 to Figure 3, as required.

We now compare some plots of the three doubt function (28, 34, & 35), for fixed  $s=10^{-3}$ , as we vary certain other model parameters. The plots of Figure 7

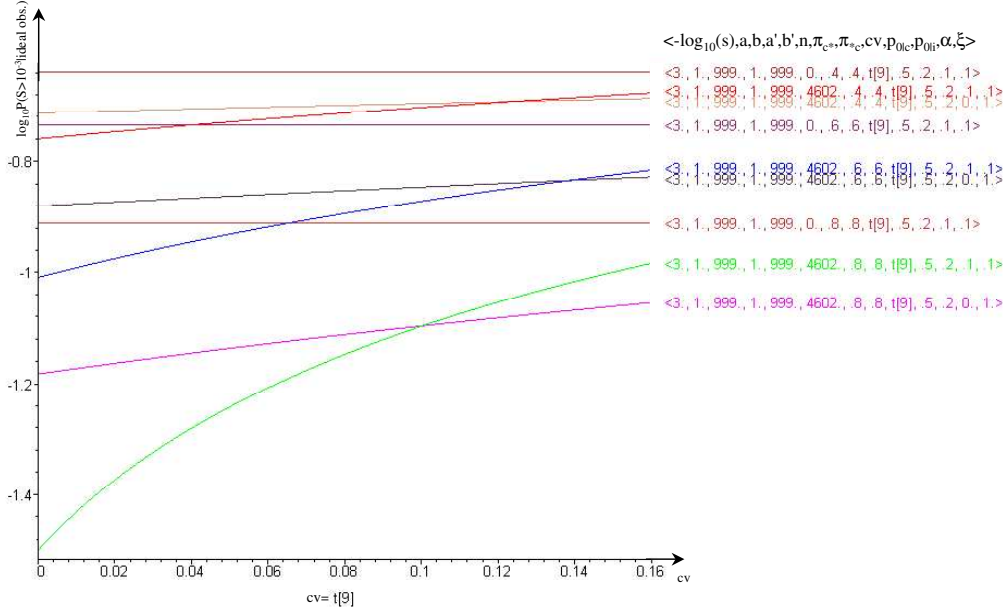


Fig. 7. Doubt functions (28) vs.  $cv$  – with params. as in (36) and following text.

show the value of  $\log_{10}$  of the doubt function (28) and were produced by setting the threshold  $s$  value to  $10^{-3}$  and the model parameters

$$(p_{0i}, a, b, p_{0c}, a', b') = (0.2, 1, 999, 0.5, 1, 999). \quad (36)$$

The nine plots shown may be thought of as three groups of three. These groups were produced by setting  $(n, \alpha, \xi)$  in turn to the values  $(4602, 0.1, 0.1)$ , then  $(4602, 0, 1)$ , then  $(0, 0.1, 0.1)$ . The *first* of these three vectors represents a *2-legged argument*, which may be compared against *the second two* which are the special cases (35) and (34) of (28), corresponding respectively to *testing-only*, and *verification-only, single-legged arguments*, with the other parameters (36) common to all three groups. *Within* each group of three, the individual plots are distinguished, and the variation of each along the horizontal axis is determined, by manipulating the  $\pi$ -matrix as follows.  $\pi$  was reparameterised in terms of the triple  $(\pi_{c*}, \pi_{*c}, cv)$ . The first two of these three parameters are the prior marginal probabilities of correctness for the specification  $Z$  and the oracle  $O$ . Variation of these is what separates the three different plots within each group. The third parameter  $cv$ , used as the horizontal axis, is an analog

of the *covariance* between  $Z$  and  $O$  (not strictly a covariance in the usual sense since the state-spaces of these two variables are not numeric, but equal to the covariance if they were converted to a pair of Bernoulli random variables)

$$\pi = \begin{bmatrix} \pi_{ci} & \pi_{ci} \\ \pi_{ic} & \pi_{ii} \end{bmatrix} = \begin{bmatrix} \pi_{c*}\pi_{*c} + cv & \pi_{c*}\pi_{*i} - cv \\ \pi_{i*}\pi_{*c} - cv & \pi_{i*}\pi_{*i} + cv \end{bmatrix}.$$

Thus *each individual plot* considered alone represents the effect on the value of the doubt function of an increasing positive, prior *covariance*  $cv$  between the correctness of specification and of oracle, while keeping the prior, *marginal* correctness probabilities of both  $Z$  and  $O$  fixed. *Comparison between* the 9 plots illustrates both the effect of changing marginal correctness probabilities  $\pi_{c*}$  and  $\pi_{*c}$ , as well as the comparison of the two single-legged arguments, against each other, and against the two-legged argument. It perhaps helps to disentangle the plots in Figure 7 to state the following descriptive observations about the shapes of the 9 curves. We will refer to individual curves here by numbering them **1** . . . **9**, counting vertically upwards at the right-hand end of the graph, with the lowest (pink) plot numbered **1**. (Because of some curves crossing, this means that the plots at the left-hand end of the graph, are numbered in the order **2, 1, 5, 3, 4, 8, 6, 7, 9**, moving upwards):

- The three flat, horizontal plots (**3, 6, 9**) are the doubt functions for the verification-only argument. They are flat because the verification-only argument does not depend on the correlation  $cv$  of prior correctness of specification  $Z$  and oracle  $O$ . It depends only on the prior marginal distribution  $\pi_{c*}$  of  $Z$  (34).
- The prior marginals  $(\pi_{c*}, \pi_{*c})$  have the values  $(0.8, 0.8)$  in plots (**1, 2, 3**),  $(0.6, 0.6)$  in plots (**4, 5, 6**), and  $(0.4, 0.4)$  in plots (**7, 8, 9**). For each fixed value of the prior marginals, towards the left-hand end (small  $cv$ ) of the graph<sup>26</sup>, there is a clear ordering in ‘performance’ (for these particular chosen model parameter values) between the three arguments: two-legged argument (**2, 5, 8**) is better than testing-only argument (**1, 4, 7**) which is better than verification-only argument (**3, 6, 9**). However the ordering of the two-legged, as compared to the testing-only is reversed towards the right-hand end of the graph. (See further discussion below.)
- There are six plotted doubt functions which are not flat. Of these, three (**2, 5, 8**) are from the two-legged argument, and three (**1, 4, 7**) are from the testing-only argument. For each  $\pi_{c*}$ , and  $\pi_{*c}$ , the two-legged argument doubt function has a steeper gradient – as a function of the correlation  $cv$  – than the corresponding testing-only argument. (**2** is steeper than **1**, etc.) The latter does increase as a function of  $cv$ , but, in each case, rather more

<sup>26</sup> but continuing, as before, to refer to the plots by number, according to their order at the *right-hand* end

gently than it would with the addition of verification evidence.

- For arguments of each of the three kinds, doubt always increases (confidence diminishes) as the prior marginal probabilities of correctness of  $Z$  and  $O$  decrease (from  $(\pi_{c^*}, \pi_{*c}) = (0.8, 0.8)$  in curves **(1, 2, 3)**, down to  $(0.6, 0.6)$  in **(4, 5, 6)**, down to  $(0.4, 0.4)$  in **(7, 8, 9)**).

These points should be sufficient for the reader visually to classify the nine plots according to type of argument (verification-only, or testing-only, or 2-legged), or according to prior marginal probabilities  $\pi_{c^*}, \pi_{*c}$  (which we have assumed equal to each other in each plot) of correctness of specification and of oracle. For a numerical example of the ordering observed in the second bullet point above, if we fix  $cv = 0.06$ , we obtain prior joint  $ZO$  distributions

$$\pi = \begin{bmatrix} 0.7 & 0.1 \\ 0.1 & 0.1 \end{bmatrix}, \quad \text{or} \quad \begin{bmatrix} 0.42 & 0.18 \\ 0.18 & 0.22 \end{bmatrix}, \quad \text{or} \quad \begin{bmatrix} 0.22 & 0.18 \\ 0.18 & 0.42 \end{bmatrix}$$

for the correctness of the oracle and specification, accordingly as the marginal correctness probabilities  $\pi_{c^*}, \pi_{*c}$  are assigned values 0.8, or 0.6, or 0.4. These three  $\pi$ -matrices result, respectively, in values of our doubt function, ordered as (*verification-leg, testing-leg, two-legged*), of  $(0.12, 0.074, 0.062)$ , or  $(0.18, 0.14, 0.12)$ , or  $(0.23, 0.20, 0.19)$ .

In comparing the 2-legged argument with the testing-only, single-legged argument, one notable observation is that, as the correlation  $cv$  between specification and oracle correctness becomes very (perhaps implausibly?) high towards the right-hand sides of the plots, we seem to arrive at a situation in which the fact of being informed that the system has been successfully verified against its specification slightly *undermines* the high confidence that had been obtained from the failure-free testing alone. This counter-intuitive behaviour may, to an extent, be explained by our use of conservative assumptions. We meet this phenomenon again later and discuss possible explanations for it on p37 of §7.3.

The examples above all involve a symmetric  $\pi$ -matrix, so that  $Z$  and  $O$  have equal marginal probabilities of being correct. Of course we have no reason to suppose this is representative of real beliefs for this kind of system. We experimented with various other parameter values and obtained a varied set of conclusions as to the comparative efficacy of the 2-legged and single-legged arguments, based on this model. To examine one more numerical example, which has, overall, some slightly more optimistic parameter values, a slightly higher “covariance”  $cv = 0.075$  between  $Z$  and  $O$  correctness, and more prior

confidence in specification correctness than in oracle correctness, put

$$\pi = \begin{bmatrix} 0.25 & 0.40 \\ 0.25 & 0.10 \end{bmatrix}.$$

We shall express a rather high prior confidence in the reliability of the verification process against a correct specification  $\alpha=0.01, \xi=0.04$ , and slightly more optimistic prior beliefs than above about the likely values of  $S$  when  $Z$  is incorrect,

$$(p_{o\ddot{i}}, a, b, p_{o\ddot{c}}, a', b') = (0.4, 1, 999, 0.5, 1, 999), \quad (37)$$

and retain the same values as above for the amount of testing  $n=4, 602$  and the claim  $S \leq s=10^{-3}$ . Using these values produces approximately equal confidence in the claim from each single leg, and almost a two thirds reduction in doubt when we use our model to combine the evidence from these two separate legs. The doubt function values  $P(S > 10^{-3} | \text{ideal obs.})$ , in the order (*verification-leg, testing-leg, two-legged*), are (0.12, 0.12, 0.045). So, as might be expected, a two-legged argument *can* bring considerable increase in confidence about a claim, in comparison with each of its constituent legs. Whether this occurs in practice will, of course, depend upon the details of the expert beliefs as represented by the model parameters.

## 7.2 $P(S > s | \text{ideal obs.})$ as a function of the number $n$ of test cases

Under ‘*ideal observations*’ (no detected failure), one might expect, as with earlier models [16], that  $S$  should stochastically decrease – in terms of its posterior distribution (28), or (35) in the testing-only case – monotonically as the quantity  $n$  of positive testing evidence accumulates. Increasing confidence from continued failure-free testing ought surely to make this posterior probability decrease monotonically in  $n$ , for every fixed  $s$ . But our current model topology incorporates uncertainties concerning a potentially defective oracle or specification, which enable more complex kinds of model behaviour. These behaviours may be realistic features of rational uncertainty, or only spurious model artifacts. In the latter case their exclusion may require parameter constraints on the node probability distributions.

Convert (28) and (35) to an odds representation

$$\frac{P(S > s \mid \text{ideal obs.})}{P(S \leq s \mid \text{ideal obs.})} = \frac{P(S > s \ \& \ \text{ideal obs.})}{P(S \leq s \ \& \ \text{ideal obs.})} = \frac{\int_s^1 \xi(1-p_{0c})\pi_{cc} \frac{u^{a'-1}(1-u)^{b'+n-1}}{\beta(a', b')} + (1-p_{0i})\pi_{ic} \frac{u^{a-1}(1-u)^{b+n-1}}{\beta(a, b)} du + C_1}{\int_0^s \xi(1-p_{0c})\pi_{cc} \frac{u^{a'-1}(1-u)^{b'+n-1}}{\beta(a', b')} + (1-p_{0i})\pi_{ic} \frac{u^{a-1}(1-u)^{b+n-1}}{\beta(a, b)} du + C_2} \quad (38)$$

where  $C_1$ , and  $C_2$  are terms not depending on  $n$

$$C_1 = \xi(1-p_{0c})\pi_{ci}I_{1-s}(b', a') + (1-p_{0i})\pi_{ii}I_{1-s}(b, a) \quad (39)$$

$$C_2 = (1-\alpha)p_{0c}\pi_{c*} + p_{0i}\pi_{i*} + \xi(1-p_{0c})\pi_{ci}I_s(a', b') + (1-p_{0i})\pi_{ii}I_s(a, b) \quad (40)$$

with the special testing-only case (35) provided simply by deleting from these equations the four occurrences of a factor  $\xi$  and the singly occurring  $(1-\alpha)$  factor. In this form, we can see that as  $n$  increases, the  $n$ -dependent term in the numerator of (38) decreases by a higher proportion than the  $n$ -dependent term in the denominator. But in general, working from this topology, it cannot be guaranteed that this ‘proportion dominance’ property extends also to the change in the *whole* numerator over the change in the *whole* denominator. Whether this is so depends on the sizes of the two non-negative ‘constants’  $C_1$  and  $C_2$ . Below we make two simple observations on the behaviour of the ratio (38) as  $n$  increases, before investigating some example cases graphically.

First, reasoning informally, guided by the property of adjacency in the topology of Figure 3, we can infer backwards, in the familiar way, from extensive failure-free testing, applied as evidence to node  $T$ , that  $S$  is likely to be small. Realistic node conditional probabilities for  $T$  will capture this effect. But, symmetrically on the left-hand side, a parallel *negative* inference of a potentially low oracle quality  $O$  also suggests itself, as a rival explanation of the apparently positive testing evidence. (Note that the ‘conservative’ assumption (9) would appear to increase the viability of this alternative explanation for ‘successful’ testing.) So, in a manner which in detail will depend on the conditional probability distribution  $P(T|OS)$ , these two inferences may both follow when  $n$  becomes very large without failure. Their combined effect – the manner in which the two available inferences from observing large  $n$  without failure may interact – will depend also on the ‘upper part’ of the topology in Figure 3 which contains the prior belief structure concerning  $ZOS$ . Does this structure of prior belief contain a positive prior probabilistic association between correctness of the oracle  $O$  and quality of the system (small value for  $S$ ) – perhaps via a shared association with the correctness of  $Z$ ? If the answer to this question is “yes” then one can imagine a tension between two competing

tendencies of the testing evidence at node  $S$ . A question arises as to which of these two effects ‘dominates’ at  $S$ . We stress that this informal conception of competing effects at  $S$  of successful testing is not intrinsic to the topology of Figure 3, but relies heavily on node conditional probability assumptions. The model topology is sufficiently flexible to allow such possibilities, but they will depend also on the  $2 \times 2$  prior matrix  $\pi$  for  $ZO$ , and on the parameter values used for our two point-mass-augmented beta distributions (13) and (25) for  $P(S|Z)$ .

To begin examining these ideas algebraically, notice that when  $n$  is replaced by  $n+1$ , (38) decreases<sup>27</sup> precisely if the ratio of the decrease in the numerator over the decrease in the denominator exceeds the value of (38), with the original  $n$ . I.e., (38) decreases if (and only if) (38) is exceeded by

$$\frac{\int_s^1 \xi(1-p_{0c})\pi_{cc} \frac{u^{a'}(1-u)^{b'+n-1}}{\beta(a',b')} + (1-p_{0i})\pi_{ic} \frac{u^a(1-u)^{b+n-1}}{\beta(a,b)} du}{\int_0^s \xi(1-p_{0c})\pi_{cc} \frac{u^{a'}(1-u)^{b'+n-1}}{\beta(a',b')} + (1-p_{0i})\pi_{ic} \frac{u^a(1-u)^{b+n-1}}{\beta(a,b)} du}. \quad (41)$$

(Here, we have obtained the increment of the  $n$ -dependent terms in the numerator as a difference of integrals, subtracted their integrands, and so derived the factor  $u = 1 - (1-u)$ . Similarly for the denominator.) This observation can be used to derive properties of any turning points of (38) as a function of  $n$ , in terms of approximations to the incomplete beta function. For any given parameter values the turning points can then be obtained numerically. Here, we will instead observe only that (38) tends to  $C_1/C_2$  as  $n \rightarrow \infty$  (and so (28) tends to  $C_1/(C_1+C_2)$ ). This follows<sup>28</sup> from noting that the integrands tend to zero everywhere in the unit interval, except perhaps at the single point  $u=0$ . So whenever  $C_1 > 0$ , the doubt function (28) does *not* tend to zero as the amount  $n$  of testing becomes arbitrarily large without failure.

We will not study for general  $n$  the parametric assumptions required in order for a 2-legged argument, with testing evidence incorporated, to provide higher confidence than the single, verification-only leg (the  $n=0$  case of (28) and (38)). But we note that for ‘very large’  $n$  this is so whenever the  $n=0$  case of (38) exceeds the limit  $C_1/C_2$ , i.e. whenever replacing, in the ratio  $C_1/C_2$ , the four  $i$  2<sup>nd</sup>-subscripts of  $\pi$ ’s by instead four  $*$  subscripts has the effect of increasing this ratio. This is the case in all of the example plots of Figure 8. These show the value of  $\log_{10}$  of the doubt function (28) plotted against the

<sup>27</sup> and, equivalently, (28) also decreases, since  $p \mapsto \frac{p}{1-p}$  is a strictly monotonic function of probabilities

<sup>28</sup> To justify rigorously, use the Lebesgue dominated convergence theorem [18, §10.10],[14, §3.2] with the  $n=0$  case of the integrand in the role of the dominating function. (We assume throughout this paper that all our beta parameters are strictly positive, as stated in (13,25).)

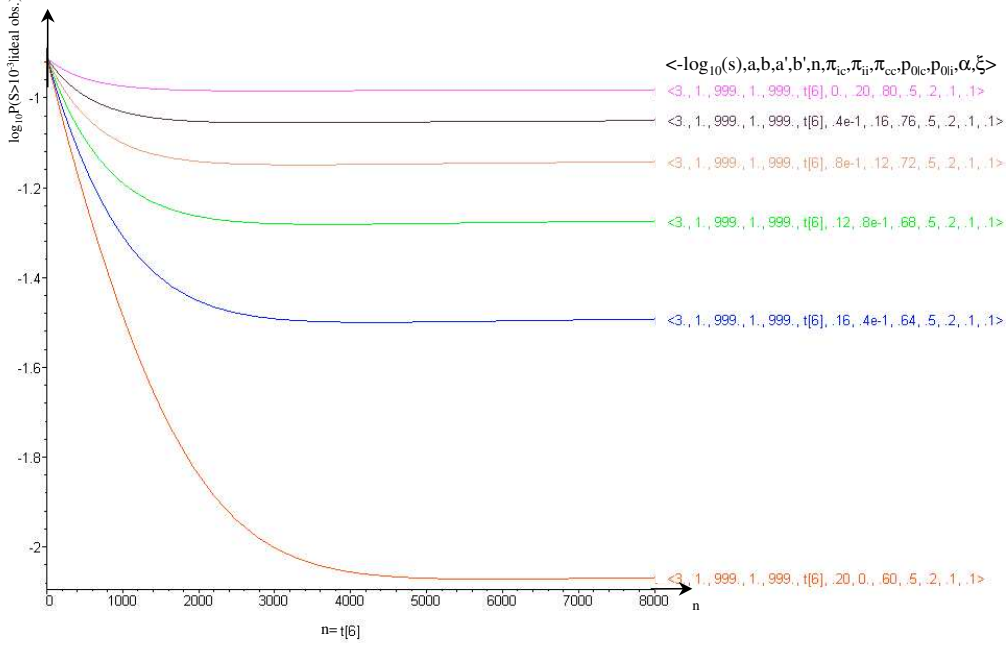


Fig. 8. Doubt functions (28) vs.  $n$  – with params. as in (42)

number  $n$  of test cases, using a claim  $s=10^{-3}$ , and model parameters

$$(p_{0i}, a, b, p_{0c}, a', b', \alpha, \xi) = (0.2, 1, 999, 0.5, 1, 999, 0.1, 0.1). \quad (42)$$

The plots differ from each other only in the values of the  $\pi$  matrix, which were produced by fixing the marginal probabilities of specification/oracle-correctness at  $\pi_{c*}=\pi_{*c}=0.8$ . The correlation coefficient between  $Z$ - and  $O$ -correctness varies as we read down the key on the RHS, with an unbelievable exact equality in the top plot (specification is correct precisely and only when oracle is correct; otherwise, both are incorrect), working down to an almost equally unbelievable negative correlation at the bottom. The second to bottom plot represents independence  $cv=0$  between the correctness of the oracle and of the specification. Clearly the general pattern here is again that low (or even negative) correlation of the prior beliefs in specification and oracle correctness yields an advantage in terms of confidence levels derivable from the two-legged argument, confirming our observation of the same tendency in Figure 7 above. Notice that there appears to be a slight deterioration in confidence at large  $n$  values, i.e. there is a turning point at some  $n$ -value (which varies from one plot to another). We informally propose to explain this deterioration by increasing suspicion of the oracle which, for large  $n$ , seems to compete successfully against the direct positive implications of continued successful testing.

### 7.3 $P(S>s \mid \text{ideal obs.})$ as a function of verification fallibility parameters $\alpha, \xi$

We shall view the pair of ‘verification fallibility levels’  $(\alpha, \xi)$  as lying within a Cartesian unit-square  $[0, 1]^2$ , with  $\alpha$  as the abscissa. (28) is then an ‘objective function’ defined on this square, for any fixed values of the threshold target reliability  $s$ , of  $n$ , and of the other 9 model parameters. Our conservative assumption (7) on p12, for the effect of a specification  $Z=\textit{incorrect}$ , makes the *upper left corner*  $(0, 1)$  of this square equivalent to a single-legged, *testing-only argument*. Any other point in the square represents the addition of an  $(\alpha, \xi)$ -fallible (when the specification  $Z=\textit{correct}$ ) verification-leg, the *lower left corner*  $(0, 0)$  corresponding to the *infallible verification* case, analysed in §4. Some of these pairs  $(\alpha, \xi)$  are more plausible than others: An adherent of the verification procedure should adopt values of  $(\alpha, \xi)$  lying somewhere ‘in the vicinity of  $(0, 0)$ ’. Certainly, at a large distance from this point, the verification methodology looks absurd. Above the  $\alpha + \xi = 1$  diagonal, we have a  $V$  more likely to reject a perfectly reliable system  $S=0$  than a system with positive pfd  $S>0$ . In the top, right-hand quarter  $\alpha>\frac{1}{2}, \xi>\frac{1}{2}$ ,  $V$  is in every respect ‘outperformed’ by its own negation,  $\neg V$ , whose  $\alpha$ - and  $\xi$ -probabilities are both smaller. So the behaviour of (28) in these upper right-hand parts is of mathematical interest only. Consequently, we might expect that under realistic values of other model parameters, prior beliefs  $(\alpha, \xi)$  sufficiently close to  $(0, 0)$  will have doubt (28) which is less than the testing-only value, at  $(0, 1)$ ?

We can outline the main characteristics of (28) over the unit square in terms of its monotonicity as a function of each of  $\alpha$  and  $\xi$  separately, for each fixed configuration of values of  $s$  and all other parameters. In all cases the monotonic dependence follows a rectangular hyperbola, by the linear fractional form of (28) in each parameter  $\alpha, \xi$ . The parameter  $\alpha$  only occurs once: in the denominator as a factor  $(1-\alpha)$ . It follows that, moving horizontally through any point in the unit square,  $P(S>s \mid \textit{ideal obs.})$  is a *monotonic non-decreasing function of  $\alpha$ , for fixed  $\xi$* . (The numerator and the denominator of (28) are both sums of products, whose factors (including the incomplete beta terms) are all probabilities – in the unit interval  $[0, 1]$ .) Fixing instead  $\alpha$ , and moving vertically, the rectangular-hyperbolic  $\xi$ -dependence can be either increasing or decreasing (or constant) throughout  $0\leq\xi\leq 1$ , depending on  $s, n$ , and the model parameters (including  $\alpha$ ). Denoting the linear fractional function in  $\xi$

$$P(S>s \mid \textit{ideal obs.}) = \frac{A\xi + B}{C\xi + D}, \quad (43)$$

there are essentially three different cases to consider, depending on all parameters other than  $(\alpha, \xi)$ , according to the value of  $\alpha$  at which this hyperbolic dependence on  $\xi$  switches from monotonic increasing to monotonic decreasing. This switch occurs whenever the determinant,  $d=AD-BC$ , of the fractional

form changes sign. (When  $d=0$ , the dependence of (43) on  $\xi$  disappears, since a change in  $\xi$  then represents an equal percentage change in both numerator and denominator.) It is for positive values of this determinant that (43) is an increasing function of  $\xi$ . For fixed values of all parameters except  $(\alpha, \xi)$ ,  $d$  itself is a monotonic *non-increasing* function  $d = d(\alpha)$ , which crosses zero at

$$\alpha_0 = 1 + \frac{\pi_{ii} + (p_{0i} + (1-p_{0i})\mu)\pi_{ic} - \frac{(1-p_{0i})(\mu'\pi_{cc} + \pi_{ci})(\mu I_{1-s}(b+n, a)\pi_{ic} + I_{1-s}(b, a)\pi_{ii})}{\mu' I_{1-s}(b'+n, a')\pi_{cc} + I_{1-s}(b', a')\pi_{ci}}}{p_{0c}\pi_{c*}} \quad (44)$$

Where does the value of this expression  $\alpha_0(s, n, \pi, p_{0c}, p_{0i}, a, b, a', b')$  lie in relation to the unit interval  $0 \leq \alpha \leq 1$ ? Before answering this question we first remark that expression (28) is a smooth function of  $(\alpha, \xi)$  so we could express what we have just stated in terms of the variation of its *gradient vector* throughout the unit square: We have shown that this gradient vector will point “eastwards” anywhere on the vertical line  $\alpha = \alpha_0$ . To the left of this line its direction will lie within the north-east quadrant, and to the right, in the south-east.

#### Case 1: $\alpha_0 \geq 1$

Here, no meaningful value of  $\alpha$  exceeds  $\alpha_0$ , so (28) is always increasing in  $\xi$  (or possibly constant along just the right-hand edge of the square, in the special case  $\alpha_0=1$ ), and the gradient vector is in the northeast quadrant throughout the whole of the unit square. It follows that the global minimum and maximum values of (28), as a function of  $(\alpha, \xi) \in [0, 1]^2$ , occur at  $(0, 0)$  and  $(1, 1)$ , respectively. This first case accords strongly with simple intuition, given that  $\alpha$  and  $\xi$  are our prior probabilities of two kinds of *errors* of the verification argument leg. We might well expect that an increased risk of either of the two kinds of verification-leg error might weaken the best (i.e. smallest  $P(S > s \mid \text{ideal obs.})$ ) confidence levels achievable from the ideal observations. By continuity, there is some neighbourhood of the maximum point  $(1, 1)$  throughout which (28) will exceed the testing-only value at  $(0, 1)$ , but, as we have argued above, for the two-legged argument, believable values of  $(\alpha, \xi)$  will be instead ‘in the vicinity of’  $(0, 0)$ .

#### Case 2: $\alpha_0 \leq 0$

We have found what appear to be plausible model parameter (and  $n, s$ ) values under which this most surprising case seems to apply. Here,  $d$  is negative (or zero) throughout the interval  $0 \leq \alpha \leq 1$ . Hence, (28) *decreases* in  $\xi$  for every fixed  $\alpha$  in  $[0, 1]$  (or perhaps is constant along just the left-hand edge of the square, in the special case  $\alpha_0=0$ ). The gradient vector everywhere points in

the south-east quadrant. Strangely, the *lowest* chance  $P(S>s | \text{ideal obs.})$  of accepting an unreliable system, is provided by the single-legged, testing only argument  $(\alpha, \xi)=(0, 1)$ , which ‘out performs’ *every* 2-legged argument obtainable while keeping  $s, n$ , and the other 9 parameters fixed. I.e., any fallibility levels (including complete infallibility  $(\alpha, \xi)=(0, 0)$ ) of an appended verification leg produce a two-legged argument whose confidence is *lower*, under ideal observations, than that of the testing-only argument. What, on the surface, appears to be unqualified ‘good news’ (that  $V=\text{verified}$ ) *lowers* our confidence in the system, when we try to extend beyond a single-legged argument. Now the worst possible (i.e. largest) value of (28) is obtained at  $(\alpha, \xi)=(1, 0)$ .

We suggest that this counter-intuitive behaviour may well be a consequence of our conservative assumption (7) for the verification outcome in the case that the specification  $Z=\text{incorrect}$ . I.e., the behaviour may be explained by the fact that (28) is – in both cases: testing-only and dual-legged – effectively only an upper bound, not an exact value, for the ‘true probability’ of large  $S$  (were we to realistically substitute for the conservative assumption). In Figure 10 we illustrate an instance of this behaviour, obtained using the values indicated in (45-46). Here the minimum value on the unit  $(\alpha, \xi)$ -square is  $P(S>s | \text{ideal obs.})=0.0727$ , occurring at  $(\alpha, \xi) = (0, 1)$ , the testing-only argument. In more detail, our tentative explanation for this discrepancy is as

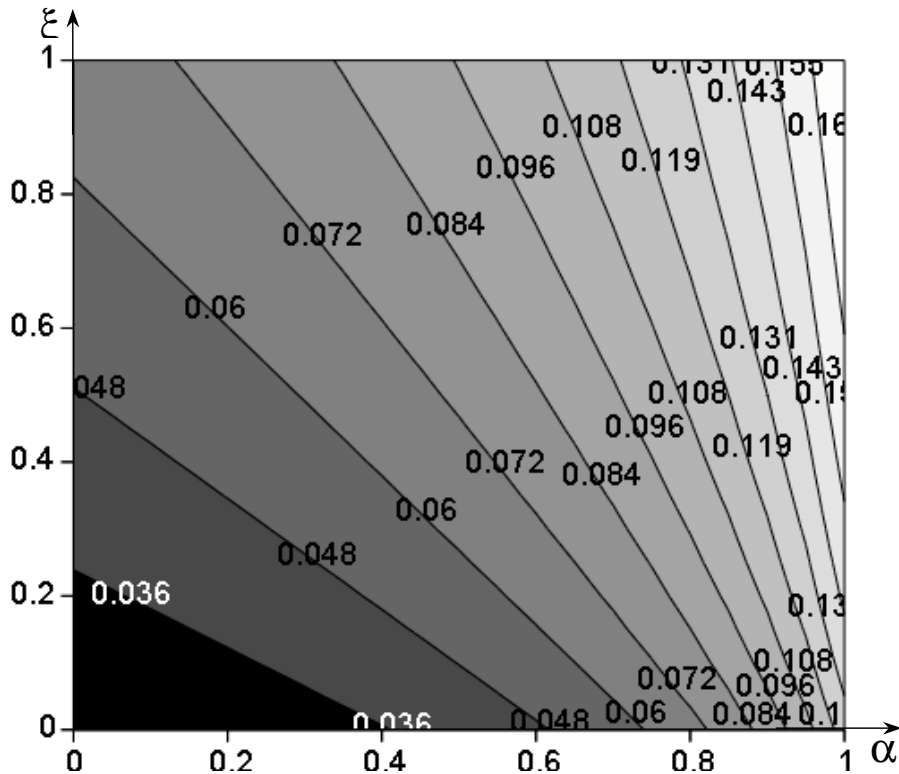


Fig. 9. Doubt function (28) vs.  $(\alpha, \xi)$  – with params. as in (45,46) on p38, gives  $\alpha_o = 1.10$

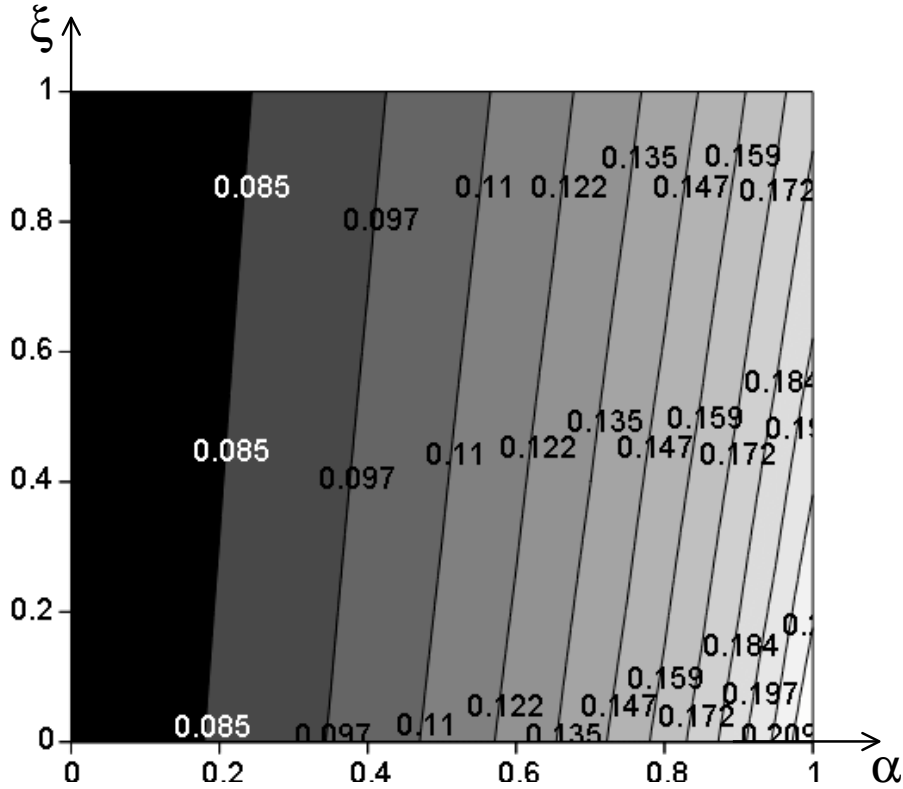


Fig. 10. Doubt function (28) vs.  $(\alpha, \xi)$  – with params. as in (45,46), on p38, gives  $\alpha_o = -0.408$

follows: While ignorant of the true value of  $S$ , the observation of a successful verification at  $V$  provides, under our conservative assumption, *stronger support for the incorrectness, than for the correctness, of the specification  $Z$* . This opens a chain of subsequent inference of the following kind. This accidental artifact of the conservative assumption undermines – *via a positive correlation between  $Z$  and  $O$  contained within the prior belief matrix  $\pi$*  – our confidence in the correctness of the oracle  $O$ . This *increased  $O$ -doubt* has a tendency to ‘explain away’<sup>29</sup> all the good news we observed from the testing leg. We do not assert that this chain of inference will be realistic in every model application—much will depend on the parameter values assigned to the node conditional probabilities. However, it does illustrate the richness of the kinds of competing inferences modelled within even a highly simplified BBN structure, and alerts us to the fact that symbolic analysis may identify model consequences that initially will seem surprising and counter-intuitive.

### Case 3: $0 < \alpha_0 < 1$

In this last case  $P(S > s \mid \text{ideal obs.})$  increases in  $\xi$  whenever  $\alpha$  is fixed at some value  $0 \leq \alpha < \alpha_0$ , but decreases in  $\xi$  for other  $\alpha_0 < \alpha \leq 1$ . So the latter set of points

<sup>29</sup> particularly in view of the other conservative assumption (9)

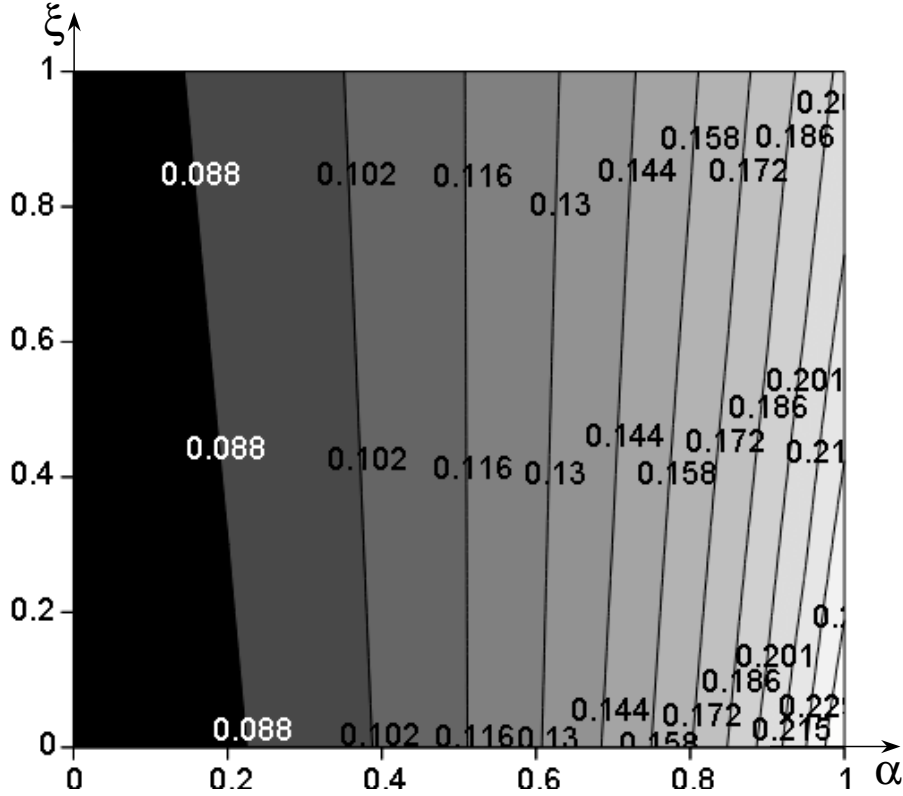


Fig. 11. Doubt function (28) vs.  $(\alpha, \xi)$  – with params. as in (45,46), on p38, gives  $\alpha_o = 0.524$

are all ‘out-performed’ by the testing-only argument (since (28) increases as we move along the sequence,  $(0, 1)$ ,  $(\alpha_0, 1)$ ,  $(\alpha, \xi)$ ). We cannot say how the testing-only case  $(0, 1)$  compares with other points, for which  $(0 \leq \alpha < \alpha_0)$ . (The exception is the points along the left hand and top boundaries of this rectangle, for which we have, in increasing (28)-order,  $(0, \xi)$ ,  $(0, 1)$ ,  $(\alpha, 1)$ .) Continuity tells us, though, that some neighbourhood of the infallible-verification point  $(0, 0)$  must outperform the testing-only case  $(0, 1)$ . On the whole unit square now we have an increasing sequence: minimum point  $= (0, 0)$ ,  $(0, \xi)$ ,  $(\alpha, \xi)$ ,  $(1, \xi)$ ,  $(1, 0)$  = maximum point.

*Numerical examples of each of the three cases:*

Assigning the parameter values

$$(n, p_{0i}, a, b, p_{0c}, a', b') = (4602, 0.2, 1, 999, 0.5, 1, 999) \quad (45)$$

it is possible to find examples at  $s=10^{-3}$ , of all three above cases (Figures 9–11, in the order discussed above) simply by manipulation of the  $\pi$ -matrix prior

distribution for  $ZO$ .

$$\begin{bmatrix} \pi_{cc} & \pi_{ci} \\ \pi_{ic} & \pi_{ii} \end{bmatrix} = \begin{bmatrix} 0.64 & 0.16 \\ 0.16 & 0.04 \end{bmatrix}, \begin{bmatrix} 0.74 & 0.06 \\ 0.06 & 0.14 \end{bmatrix}, \begin{bmatrix} 0.73 & 0.023 \\ 0.107 & 0.14 \end{bmatrix} \quad (46)$$

[Fig. 9]                      [Fig. 10]                      [Fig. 11]

The resulting  $\alpha_0$  (44) are quoted in the plot captions.

We remark that were we to relax the assumption, mentioned on p20, that  $\xi$  is independent of the value of positive  $S$ , then, rather than a point in the unit square, we would require a *function*  $\xi(s)$ , from the unit interval into itself, to characterize the ‘shape’ of the verification fallibility ( $1-\xi(0)$  replacing  $\alpha$ ).

## 8 Summary and Conclusions

As we have said earlier in this paper, the BBN we have studied here has been an over-simplified one. The first simplification is in only taking account of statistical testing and proof evidence: we have ignored other kinds of evidence that would clearly be relevant in practice – for example evidence concerning the quality and competence of the personnel involved at all stages. Furthermore, each of the argument legs considered here is itself unrealistically simplified: e.g. the testing leg ignores important issues concerning the accuracy of the operational profile. We have artificially reduced some of the state spaces to Boolean. To make the mathematics tractable, we have had to introduce some simplifying assumptions that (rather tentatively) we claim to be ‘conservative’.

Our main reason for these simplifications lay in our desire to carry out the analysis completely analytically. We wanted to obtain complete analytic expressions for posterior distributions in terms of parametric families of input node probability distributions. This contrasts with the more common approach to BBN analysis in which numerical expressions – e.g. involving elicited expert beliefs – are manipulated using tools like Hugin [19]. Simply populating the node probability tables in this way results only in a single numerical posterior distribution for the goal variable  $S$ .

In spite of the great simplification we have applied, the model turns out to be quite complex and difficult to understand. We were somewhat surprised by this, and we regard it as a strong warning against a naive trust in the results of a conventional numerical analysis of a BBN like this. In particular, we believe that the analytic approach has exposed some non-intuitive and surprising results that would not be noticed in a conventional numerical analysis. It would be possible to be lulled into a false sense of certainty and security, and

believe numerical consequences that one would not believe with the benefit of greater insight (e.g. offered by the kind of analysis we have conducted).

These remarks confirm our long-held view that BBNs need to be treated with great respect and humility [20,21]. The Bayesian approach is clearly the right one for the representation of uncertainty, and the BBN formalism has immeasurably aided understanding and construction of complex probability models. But the very seductiveness of the approach – particularly in its automated numerical form – can bring unwarranted confidence in the results.

Leaving aside these general comments, what can be said about the particular situation we have examined in this paper? Our intention was to try to understand better the idea of multi-legged arguments. In particular, we were interested in how much (if at all) the use of multi-legged arguments could increase confidence in a safety claim, over and above what could be claimed from a single argument alone. When we set out on this work, we had in mind an analogy with the use of diversity in systems – multiple diverse channels – to increase reliability and safety. It seemed plausible that multi-legged arguments could be used similarly to increase confidence in claims about dependability (e.g. safety).

One result from the paper is that this analogy breaks down in a surprising way. Whereas it is easy to show that for systems, a diverse 1-out-of-2 system is always more reliable than each of its component channels, the same is not true of arguments. We cannot be sure that the confidence in a dependability claim arising from a diverse 2-legged (‘1-out-of-2’) argument is greater than that arising from either of the single argument legs. Indeed, we have examples where a single leg is to be preferred to the same leg aided by a further supportive argument leg: i.e. additional ‘good news’ is not necessarily beneficial and can even be detrimental.

Such results are, at first glance, counter-intuitive: indeed, if they had been obtained from a purely numerical analysis they would be hard to explain. The more detailed analysis here has the advantage of showing how this kind of thing can happen, by revealing the subtle interplay between assumptions and evidence, both within and between legs. It thus provides warnings against making simplistic – albeit intuitively plausible – assumptions. It cannot be too strongly emphasised that it is only through the completely analytical treatment – difficult though it is – that we get these insights.

What are the practical implications of all this? It is very early days yet to be definitive, but we expect this kind of work to have an impact on the way safety claims are supported. The key here is the notion of ‘confidence’. We believe this is an essential complement to claim level. It is not sufficient to say “I believe that this system has a pfd smaller than  $10^{-3}$ ”, since you cannot

be certain of the truth of such a claim. You must therefore additionally state your confidence in the claim, e.g.: “I am 95% confident that this system has a pfd smaller than  $10^{-3}$ ”.

If the reader accepts that such a calculus of confidence (and claims) is needed, then our tentative results come into play. Requirements for systems (and sub-systems) need to be expressed in terms of claim and confidence. Decisions about whether systems should be allowed to be used will be based on arguments that demonstrate that a particular dependability claim can be believed at a particular level of confidence. Most importantly, the benefits of using diverse arguments to increase confidence in a claim need to be evaluated properly. Even if it is true that diverse arguments are generally a good thing (and our results appear to show that this may not always be so), as implied by their advocacy in certain standards [1,2], it is still necessary to know how much they deliver. Our kind of analysis can help here. Of course, in all this we are assuming the need for a higher level of formality than is customarily the case even for safety cases in some critical systems. Some may argue that qualitative approaches have served us well in the past: can we not hide all the complexity of analyses like the ones here, and rely on (say) the judgement of experienced experts? We think not. We think the complexity here is inherent, and you ignore it at your peril. If a purely numerical BBN can get it wrong (by ignoring some of the complexity), why should we believe that an even more informal qualitative approach can get it right, when it ignores even more of the complexity?

Concerning the general efficacy of diverse arguments, it is clear that more work is needed. It would be interesting to know, for example, whether the kinds of parameters that are realistic (e.g. for experts’ beliefs) in our simplified model result in 2-legged arguments that are almost always effective, in the sense of being better than the constituent legs. (We have presented an example in which the 2-legged argument gives only one third of the claim doubt of each of the two single legs.) Are the exceptions that we have identified in some sense ‘not believable’ when real experts assess real systems? To what extent are some of these results the consequence of our need to make ‘conservative’ assumptions for mathematical tractability? (Is the notion of ‘conservative’ more subtle than we appreciate?)

As things stand, our results demonstrate that there *can* be some unexpected subtleties involved here that can, *in extremis*, undermine the efficacy of the multi-legged approach, but our results are neutral on the issue of whether there *will* be such problems in particular cases (i.e. when particular experts assess the dependability of particular systems).

We plan to address issues like these in future work.

## Acknowledgement

This work was partially supported by the DIRC project ('Interdisciplinary Research Collaboration in Dependability of Computer-Based Systems') funded by UK Engineering and Physical Sciences Research Council (EPSRC), and by the DISPO (DIverse Software PrOject) projects, funded by British Energy Generation Ltd and BNFL Magnox Generation under the Nuclear Research Programme via contracts PP/40030532 and PP/96523/MB.

## References

- [1] Ministry of Defence, The procurement of safety critical software in defence equipment, UK Defence Standard Def-Stan 00-55, issue 2 (August 1997).
- [2] Civil Aviation Authority, Regulatory objective for software safety assurance in air traffic service equipment, CAA SW01 (2001).
- [3] R. E. Bloomfield, B. Littlewood, Multi-legged arguments: The impact of diversity upon confidence in dependability arguments, in: Proceedings DSN 2003, IEEE Computer Society, 2003, pp. 25–34.
- [4] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Mathematics and Its Applications, Morgan Kaufmann, San Mateo, California, 1988, revised 2<sup>nd</sup> printing 1991.
- [5] A. P. Dawid, Conditional independence in statistical theory, Journal Royal Statistical Society, Series B 41 (1) (1979) 1–31, with discussion.
- [6] D. R. Wright, Elicitation and validation of graphical dependability models, Tech. rep., City University, ROPA Project Report: [www.csr.city.ac.uk/people/david.wright/ropa/](http://www.csr.city.ac.uk/people/david.wright/ropa/) (2003).
- [7] M. Volf, M. Studený, A graphical characterisation of the largest chain graphs, International Journal of Approximate Reasoning 20 (3) (1999) 209–36, <ftp://ftp.utia.cas.cz/pub/staff/studený/volstu.ps>.
- [8] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, D. J. Spiegelhalter, Probabilistic Networks and Expert Systems, Statistics for Engineering and Information Science, Springer-Verlag, New York, 1999.
- [9] S. L. Lauritzen, Graphical Models, Oxford Statistical Science Series, Clarendon Press, Oxford, 1996.
- [10] G. Shafer, Probabilistic Expert Systems, CBMS-NSF Regional Conf. Ser. in Applied Math., Society for Industrial & Applied Mathematics, Philadelphia, 1996.

- [11] A. P. Dawid, Conditional independence for statistical operations, *Annals of Statistics* 8 (3) (1980) 598–617.
- [12] M. Studený, On mathematical description of probabilistic conditional independence structures, Dr. of Science Thesis, Institute of Information Theory and Automation, Academy of Sciences of the Czech Republic, Prague (May 2001).
- [13] N. Wermuth, S. L. Lauritzen, On substantive research hypotheses, conditional independence graphs and graphical chain models, *Journal Royal Statistical Society, Series B* 52 (1) (1990) 21–72, with discussion.
- [14] G. de Barra, *Measure Theory and Integration, Mathematics and Its Applications*, Ellis Horwood, Chichester, 1981.
- [15] M. H. DeGroot, *Optimal Statistical Decisions*, Wiley, 2004, wiley Classics Library Edition. ISBN 0-471-68029-X.
- [16] B. Littlewood, D. Wright, Some conservative stopping rules for the operational testing of safety-critical software, *IEEE Transactions on Software Engineering* 23 (11) (1997) 673–83.
- [17] M. Abramowitz, I. A. Stegun (Eds.), *Handbook of Mathematical Functions*, Dover, New York, 1970, <http://www.math.sfu.ca/~cbm/aands/>.
- [18] T. M. Apostol, *Mathematical Analysis*, World Student Series, Addison-Wesley, 1974.
- [19] S. K. Andersen, K. G. Olesen, F. V. Jensen, F. Jensen, Hugin—a shell for building Bayesian belief universes for expert systems, in: *Proceedings of the Eleventh International Joint Conference on Artificial Intelligence*, Detroit, 1989, pp. 1080–84.
- [20] S. Yih, C.-F. Fan, Search for the unnecessary, *Nuclear Engineering International* (2001) 24–6.
- [21] R. Bloomfield, P.-J. Courtois, L. Strigini, B. Littlewood, Letter to the editor, *Nuclear Engineering International* (2002) 11.

Table 1. How the Conditional Probability Tables Combine to Produce  $P(S \& ideal\ obs.)$  – Infallible Verification

$S$	0				$> 0$			
	correct		incorrect		correct		incorrect	
$Z$	correct	incorrect	correct	incorrect	correct	incorrect	correct	incorrect
$O$	correct	incorrect	correct	incorrect	correct	incorrect	correct	incorrect
$P(T=no\ failures OS)$	$1 = f(0)$	1	$1 = f(0)$	1	$f(S)$	1	$f(S)$	
$P(S Z)$	$P(S=0 Z=corr.)$		$P(S=0 Z=incorr.)$		$P(S Z=corr.)$		$P(S Z=incorr.)$	
$P(V=verified SZ)$	1	1	1	1	0	0	1	
$P(S \& ideal\ obs.)$	$P(S=0 Z=corr.)P(Z=corr.) + P(S=0 Z=incorr.)P(Z=incorr.)$				$P(S Z=incorr.) [f(S)P(ZO=(incorr., corr.)) + P(ZO=(incorr., incorr.))]$			
$P(ideal\ obs.)$	$P(S=0 Z=corr.)P(Z=corr.) + P(S=0 Z=incorr.)P(Z=incorr.)$				$\int_{0^* < S \leq 1} P(S Z=incorr.) [f(S)P(ZO=(incorr., corr.)) + P(ZO=(incorr., incorr.))]$			
	$= P(S=0 Z=corr.)P(Z=corr.) + P(ZO=(incorr., corr.)) E(f(S)   Z=incorr.) + P(ZO=(incorr., incorr.))$							

(\* understanding that this integral specifically *excludes* any probability mass at  $S=0$  of the conditional distribution of  $S$  given  $Z=incorr.$ . In fact the term immediately to the left of this integral is that excluded point-mass contribution, since  $f(0)=1$ . Combining these two terms makes an expectation w.r.t.  $S$  given  $Z=incorr.$ , which simplifies to produce an alternative representation of  $P(ideal\ obs.)$  as shown on final row of the table.)

