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 [a,*] David Wright [a]

[a] *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', which will be used as a measure of efficacy. 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 ("Bayesian Belief Network") 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 and counter-intuitive subtleties here; for example an entirely supportive second leg can sometimes undermine an original argument, resulting overall 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.

Keywords: safety claims, safety arguments, software safety, software reliability, Bayesian belief networks

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. There may be doubt about the truth of some of the assumptions that underpin the reasoning used to support a claim, e.g. that a test oracle is correct.

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 multiple diverse arguments to support dependability claims: the idea is analogous to the use of fault tolerance to make systems reliable. Thus each of the diverse arguments could, in principle, support the claim but might be undermined by doubt about underlying assumptions, weakness of evidence, etc.

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 was suggested that one leg be based upon logical proof of correctness, the other upon statistical testing. 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 verification and validation. This last case 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', just as it is for achieving system dependability, 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

* Corresponding author.

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 argument 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: e.g., for a given outlay, would it be better to divide this between a proof and a testing leg, or to spend it all on a larger test?

At an informal level, one can take an argument leg to comprise: some *assumptions*, some *evidence*, and some *reasoning* that allow a dependability *claim* to be made at a certain level of *confidence*. Typically, such an argument leg will support an infinite number of different (claim, confidence) pairs – the more stringent the claim, the lower the confidence that will come from a particular argument leg. We shall interpret 'confidence' to be a probability (that the claim is true). This probability, in turn, will be interpreted as the usual Bayesian subjective strength of belief in the claim, held by an individual whom we shall refer to as 'the expert'. This person might be a regulator, or some other person who has to take a decision on the acceptability of the system.

It is thus confidence that allows us to discuss the 'strength' of arguments: for example, an argument that allows someone to place 99% confidence in a claim that a system's probability of failure on demand is less than 10^{-3} is clearly 'stronger' than an argument that only allows them to place 90% in this paper we shall always consider the claim to be fixed – perhaps arising from some wider safety case – and thus compare arguments solely via the confidence they engender in this claim. Clearly this is not the only way one could proceed, but it will suffice for our purposes here.

It is easy to see that confidence – and its complement, 'doubt' – will depend upon: confidence/doubt in the truth of the assumptions underpinning the argument; strength and/or extensiveness of the evidence; correctness of the reasoning. Continuing in this informal vein, it seems plausible that for *multi-legged* arguments the overall effectiveness will depend upon the same factors, and in addition the *dependence* between the legs. Thus we might expect a two-legged argument whose legs are 'very diverse' to be more effective – i.e. give greater confidence – than one where the legs are very similar (all things being equal).

Of course, as we have been at pains to state, all this is very informal. Our intention in the work reported here is to put these ideas onto a formal basis.

We shall use as the basis of the paper a more formal treatment of an example

first examined in [3, Example 1, p27]. In this example a two-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% 10^{-3} then the additional *verification leg* might allow this level of confidence in 10^{-3} to be increased.

Note that for these individual legs important sources of doubt in the claim are doubt about the correctness of the oracle (for the testing leg), and doubt about the correctness of the specification (for the verification leg). Furthermore, such doubts are likely to be dependent: in particular, doubt in the correctness of the specification is likely to affect doubt in the correctness of the oracle. We might expect that the greater the doubt in the specification, the greater the doubt in the oracle. Clearly, these doubts will propagate and affect the confidence associated with the use of both arguments in a two-legged configuration: we might expect this confidence to be less than would be the case if the specification and oracle doubts were independent.

Note also that dependence between legs in this example can arise in other ways. It could arise from the *evidence*, for example: the observation of a failure in the testing leg would completely refute the perfection claim of the second leg.

Issues of (lack of) dependence here are very similar to those that arise in *system* diversity, where it is has been established from theoretical [4,5] and empirical [6] studies that independence is extremely elusive. If this is also true of arguments we would need to be sceptical of simplistic claims, e.g. that 99% based on two legs each of which alone only allow 90% understanding of argument *dependence* is therefore an important goal of this research. It turns out, in fact, 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.

In this paper we mainly concentrate on the problems associated with the assessment of confidence. We only address issues of decision-making – e.g. whether to accept and deploy a system – briefly, and then only to treat the case of 'dangerous' argument failure, i.e. the acceptance of an untrue claim. The other kind of failure - rejecting a claim when it is true - will also be important (e.g. it may have important economic consequences), but will not be considered here.

The paper is organised as follows. We begin by describing a 6-variable Bayesian belief network (BBN) representing the structure of the two-legged argument example. A BBN topology for this system assessment is first presented, with an enumeration of the *independency model* [1] which it represents.

There follow 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 two-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.

Our allocation of node probability tables used to analyse this ideal observations case is parameterized, i.e. it specifies the conditional probabilities of each node, given its parents' values, numerically except for unspecified values of a set of independent model parameters. There are 12 independent model parameters significant for the analysis of the ideal observation case. A particular expert's beliefs will thus be represented by an assignment of numerical values to these parameters.

A substantial part of the remainder of the paper examines the consequences of different expert beliefs, concentrating on questions of when the two-legged approach is effective, and what are the factors that determine its effectiveness. We show that there are some unexpected subtleties here, and give examples of some surprising and non-intuitive results.

[1] Other synonymous terms and some references are given on p7.

2 Model Variable Definitions and BBN Topology

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, the discoveries and difficulties encountered during later stages may well feed back into adjustments and refinements to earlier stages.

The first stage in building a BBN is the identification of model variables. Our model variables are defined in the following list, which gives in square brackets the state-space we have used, in this paper, for each variable. For the sake of simplicity in this initial model formulation, the state-spaces are all Boolean, apart from the first. In some cases, this is a deliberate simplification of the real situation which we do not intend to retain in all our future work.

- S - The system's unknown, true probability of failure on demand (pfd) $[0 \leq S \leq 1]$.
- Z - system *specification* $[Z \in \{correct, incorrect\}]$. 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 $[V \in \{verified, not\ verified\}]$. 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) $[O \in \{correct, incorrect\}]$. 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* $[T \in \{no\ failures, failures\}]$. We shall be interested here only in the *ideal observation* case of *no failures*.
- C - acceptance (or otherwise) of final *claim* as to whether or not the system is fit for use $[C \in \{accepted, rejected\}]$. 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. For the purposes of illustration throughout this paper we will use the claim that the operational system pfd S is better than 10^{-3}.

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*, *(in)dependency relation, (in)dependency model*, or *Conditional Independence (CI) relation* – see p91 of [7], §2.4 of [8], and p5 of [9] – which enable the grand multivariate distribution of all model variables to be composed of several node conditional probability distributions of lower dimension. 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). Thus probabilistic CI assumptions and the BBN topologies that encode them are devices for constructing, communicating, and reasoning with multi-variate 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' set of nodes. The formal notions of 'd-separation', 'graphoid', 'semi-graphoid', and 'I-map' are central to this theory. See [7–10] for further details.

Our BBN topology is shown in Figure 1.

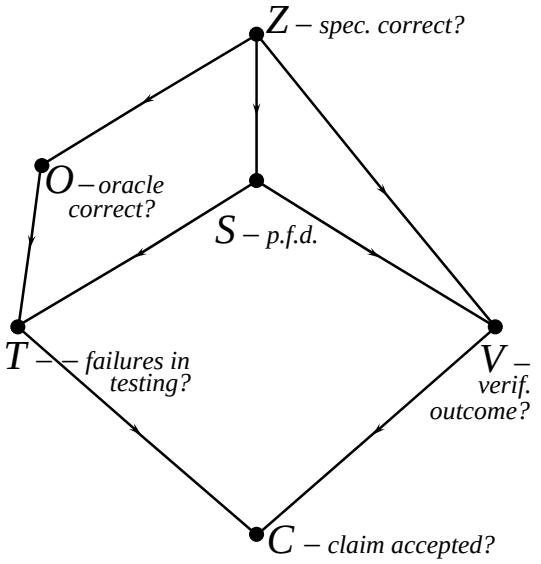

Fig. 1. BBN model topology

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:

$$\mathcal{A} \perp\!\!\!\perp \mathcal{B} | \mathcal{S} \qquad \text{means} \qquad P(\mathcal{A}, \mathcal{B}, \mathcal{S}) = P(\mathcal{A}|\mathcal{S}) P(\mathcal{B}|\mathcal{S}) P(\mathcal{S}), \tag{1}$$

where $\mathcal{A}, \mathcal{B}, \mathcal{S}$, here represent *sets* of model *random variables*[2]. 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*[3]. Our 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. Other CI statements are logical consequences of these. (See e.g. [7,11] for precise definitions and theory of how to use the net topology to determine all its logical CI consequences.) 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[4] set of CI 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 together completely specify the model is:

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

Expressed in terms of its 26 *elementary CI statements*[5] as in [8], the same dependency model can be written

[2] Other forms, such as $P(\mathcal{A}, \mathcal{B}|\mathcal{S}) = P(\mathcal{A}|\mathcal{S}) P(\mathcal{B}|\mathcal{S})$, 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, $P(\mathcal{A}, \mathcal{B}, \mathcal{S}) P(\mathcal{S}) = P(\mathcal{A}, \mathcal{S}) P(\mathcal{B}, \mathcal{S})$.

[3] *Conditional* independence means distinct factors may contain common variables.

[4] 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 *de*pendence beliefs expressed by the same expert.

[5] [8] shows that *every* probabilistic dependency model is completely characterised by listing its elementary CI statements.

$O \perp\!\!\!\perp S \mid Z$	$O \perp\!\!\!\perp S \mid ZV$			
$O \perp\!\!\!\perp V \mid Z$	$O \perp\!\!\!\perp V \mid ZS$	$O \perp\!\!\!\perp V \mid SZT$	$O \perp\!\!\!\perp V \mid SZTC$	
$Z \perp\!\!\!\perp T \mid OS$	$Z \perp\!\!\!\perp T \mid OSV$	$Z \perp\!\!\!\perp T \mid OSVC$		
$V \perp\!\!\!\perp T \mid OS$	$V \perp\!\!\!\perp T \mid OSZ$	$V \perp\!\!\!\perp T \mid SZ$		
$Z \perp\!\!\!\perp C \mid VT$	$Z \perp\!\!\!\perp C \mid VTO$	$Z \perp\!\!\!\perp C \mid VTS$	$Z \perp\!\!\!\perp C \mid VTOS$	$Z \perp\!\!\!\perp C \mid OSV$
$O \perp\!\!\!\perp C \mid VT$	$O \perp\!\!\!\perp C \mid VTS$	$O \perp\!\!\!\perp C \mid VTZ$	$O \perp\!\!\!\perp C \mid VTSZ$	$O \perp\!\!\!\perp C \mid SZT$
$S \perp\!\!\!\perp C \mid VT$	$S \perp\!\!\!\perp C \mid VTO$	$S \perp\!\!\!\perp C \mid VTZ$	$S \perp\!\!\!\perp C \mid VTOZ$	

When an expert declares himself satisfied with a BBN topology such as that in Figure 1, he is really saying that he believes the CI assertions that are entailed by the topology. Thus, part of the topology elicitation exercise would be an exposure to, and acceptance of, these.

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* $\mathcal{B}$*'*, [7]. 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 $\mathcal{A}$. See [7,11–16], and §1.2.1 of [9]. For example, our model assumes $V \perp\!\!\!\perp T \mid SZ$. A factorization such as

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

does *not* follow; whereas

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

is logically entailed within our model.

This particular CI assumption also illustrates another important point about the above informal interpretation 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. 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 $\mathcal{A} \perp\!\!\!\perp \mathcal{B} | \mathcal{S}$, 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 $\mathcal{A} \perp\!\!\!\perp \mathcal{B} | \mathcal{S}$ 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 $\mathcal{S}$.

Figures 2 and 3 show the BBNs representing an obvious pair of single-legged arguments whose 'combination', in some sense, we have discussed up to now, in the form of the BBN model shown in Figure 1:

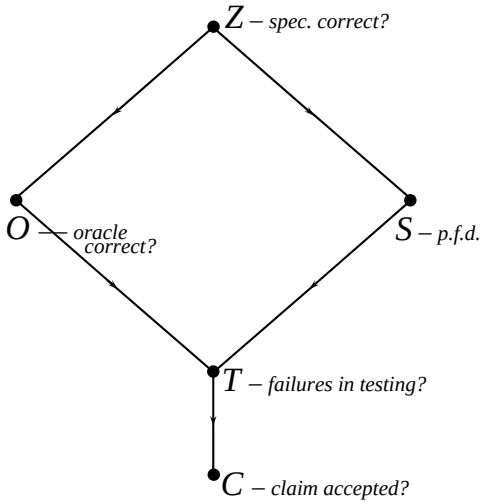

Fig. 2. Testing leg BBN topology

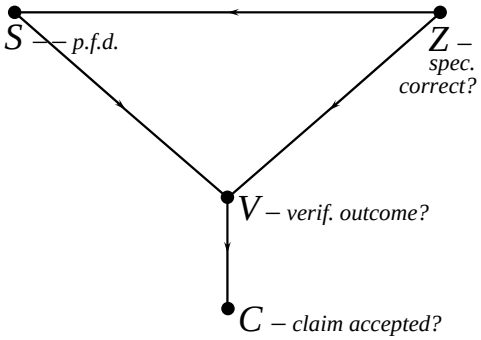

Fig. 3. Verification leg BBN topology

These topologies may be obtained from that of Figure 1 by removing one or other of the two observable nodes V and T. It may be that an expert who accepts the dependency model of Figure 1 as correct for the uncertainties inherent in the two-legged argument would feel the same about Figures 2 or 3 for an argument in which only one of these two source of evidence is available. More rigorously, the relationship between these one-legged and two-legged argument topologies is not trivial. We do not envisage that the appropriate one-legged dependency model should necessarily be merely a 'marginal' of the two legged model, obtained by summation over the observable V or T that is to be removed. For example, in all the example uses of these topologies which follow, we choose to assign the values of C, "claim acceptance", deterministically in terms of C's (one or two) parents' values. This means that our single-legged models will attribute to variable C a different stochastic relationship with the remaining model variables. (Below, in our model based on the topology of Figure 2, we make C a deterministic function of T alone: in our two-legged model based on Figure 1 we do not.) In what follows when we have assigned node conditional probabilities to our topologies, we will say a little more about the relationship between the three multivariate probability models that result.

Represented algebraically, the two single-leg dependency models are

Testing Leg	Verification Leg
$O \perp\!\!\!\perp S \,\vert\, Z$	$C \perp\!\!\!\perp SZ \,\vert\, V$
$T \perp\!\!\!\perp Z \,\vert\, OS$	
$C \perp\!\!\!\perp OSZ \,\vert\, T$	

the first OS-symmetric; and the second SZ-symmetric [6]. In the form of exhaustive lists of elementary CIs, we have:

Testing Leg				Verification Leg	
$O \perp\!\!\!\perp S \,\vert\, Z$					
$Z \perp\!\!\!\perp T \,\vert\, OS$	$Z \perp\!\!\!\perp T \,\vert\, OSC$				
$Z \perp\!\!\!\perp C \,\vert\, T$	$Z \perp\!\!\!\perp C \,\vert\, TO$	$Z \perp\!\!\!\perp C \,\vert\, TS$		$Z \perp\!\!\!\perp C \,\vert\, V$	$Z \perp\!\!\!\perp C \,\vert\, VS$
$Z \perp\!\!\!\perp C \,\vert\, OS$	$Z \perp\!\!\!\perp C \,\vert\, TOS$				
$O \perp\!\!\!\perp C \,\vert\, T$	$O \perp\!\!\!\perp C \,\vert\, TS$	$O \perp\!\!\!\perp C \,\vert\, TZ$	$O \perp\!\!\!\perp C \,\vert\, TSZ$		
$S \perp\!\!\!\perp C \,\vert\, T$	$S \perp\!\!\!\perp C \,\vert\, TO$	$S \perp\!\!\!\perp C \,\vert\, TZ$	$S \perp\!\!\!\perp C \,\vert\, TOZ$	$S \perp\!\!\!\perp C \,\vert\, V$	$S \perp\!\!\!\perp C \,\vert\, VZ$

Readers may have noticed that, although we have informally treated a two-legged argument as a combination of single legs, our more formal treatment above has moved in the reverse direction. Comparison of these last CI lists with the one on p9 clearly does not indicate any formal operation allowing composition of these two single-legged dependency models to create the two-legged model. The two-legged model of Figure 1 is a construction embodying various subjective beliefs about the dependencies arising in practice among the variables of our particular application. In particular, our two-legged dependency model can be characterised neither as weaker nor as stronger than the simple conjunction [7] of the two separate single-legged dependency models.

Precisely, using the elementary CI statement dependency model characterization, the two-legged model deletes all CIs from one single-legged model, most CIs from the other, and of course introduces several other CIs whose contexts [8]

[6] The latter symmetry is not inherited by Figure 3 itself, though there are derivable *canonical* graph representations which *always* show the same symmetries as the dependency model they represent, such as the "largest chain graph" [8,17] model representation.

[7] i.e. the pooled set of conditional independencies involving the 6 variables

[8] the *context* of a CI statement is the set of all the model variables it involves: $\text{context}(A \perp\!\!\!\perp B \,\vert\, C) = A \cup B \cup C$.

are not contained in the variable set for either single leg.

Deleted from Testing Leg		Deleted from Verification Leg		Appended	
				$O \perp\!\!\!\perp S \mid ZV$	
				$O \perp\!\!\!\perp V \mid Z$	$O \perp\!\!\!\perp V \mid ZS$
				$O \perp\!\!\!\perp V \mid SZT$	$O \perp\!\!\!\perp V \mid SZTC$
$Z \perp\!\!\!\perp T \mid OSC$				$Z \perp\!\!\!\perp T \mid OSV$	$Z \perp\!\!\!\perp T \mid OSVC$
				$V \perp\!\!\!\perp T \mid OS$	$V \perp\!\!\!\perp T \mid OSZ$
				$V \perp\!\!\!\perp T \mid SZ$	
$Z \perp\!\!\!\perp C \mid T$	$Z \perp\!\!\!\perp C \mid TO$	$Z \perp\!\!\!\perp C \mid V$		$Z \perp\!\!\!\perp C \mid OSV$	$Z \perp\!\!\!\perp C \mid VT$
$Z \perp\!\!\!\perp C \mid TS$	$Z \perp\!\!\!\perp C \mid TOS$	$Z \perp\!\!\!\perp C \mid VS$		$Z \perp\!\!\!\perp C \mid VTO$	$Z \perp\!\!\!\perp C \mid VTS$
$Z \perp\!\!\!\perp C \mid OS$				$Z \perp\!\!\!\perp C \mid VTOS$	
$O \perp\!\!\!\perp C \mid T$	$O \perp\!\!\!\perp C \mid TS$			$O \perp\!\!\!\perp C \mid VT$	$O \perp\!\!\!\perp C \mid VTS$
$O \perp\!\!\!\perp C \mid TZ$				$O \perp\!\!\!\perp C \mid VTZ$	$O \perp\!\!\!\perp C \mid VTSZ$
$S \perp\!\!\!\perp C \mid T$	$S \perp\!\!\!\perp C \mid TO$	$S \perp\!\!\!\perp C \mid V$		$S \perp\!\!\!\perp C \mid VT$	$S \perp\!\!\!\perp C \mid VTO$
$S \perp\!\!\!\perp C \mid TZ$	$S \perp\!\!\!\perp C \mid TOZ$	$S \perp\!\!\!\perp C \mid VZ$		$S \perp\!\!\!\perp C \mid VTZ$	$S \perp\!\!\!\perp C \mid VTOZ$

Clearly many of these changes involve variable C and relate to its changed role, mentioned above, in the dependence structure as we move between these three dependency models.

3 Computations from these BBN Topologies

We are primarily interested in the updated joint probability distribution $P(CS \mid \text{observations})$, particularly the value $P\Big(C{=}\textit{accepted}, S{>}10^{-3} \;\Big|\; \text{observations}\Big)$ concerning 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

$$P(CSVT) = P(C|VT)\left\{\sum_{Z} P(V|SZ)P(S|Z)\sum_{O} P(T|OS)P(OZ)\right\} \qquad (2)$$

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

$$P(CS|VT) = \frac{P(C|VT)\left\{\sum_Z P(V|SZ)P(S|Z)\sum_O P(T|OS)P(OZ)\right\}}{\sum_C \int_S P(C|VT)\left\{\sum_Z P(V|SZ)P(S|Z)\sum_O P(T|OS)P(OZ)\right\} dS}$$
$$= \frac{P(C|VT)\left\{\sum_Z P(V|SZ)P(S|Z)\sum_O P(T|OS)P(OZ)\right\}}{\int_S \left\{\sum_Z P(V|SZ)P(S|Z)\sum_O P(T|OS)P(OZ)\right\} dS} \quad (3)$$

We have used integration over our single continuous model variable S here, interpreting $P(S|Z)$ as a density function. In fact, notice that, since we are to accept that perfection (S=0) is possible, then we must allow mixed distributions for S (these often involving other model variables too, being joint or conditional distributions). In this notation, this means thinking of integrands over S as potentially exhibiting 'delta-function-like' behaviour, at S=0 (One could use Lebesgue-Stieltjes integrals [18] to notate this more rigorously.)

4 Node Probability Assumptions

4.1 Simplifying and Conservative Assumptions

We start with some assumptions that will simplify the mathematics. Some of these assumptions are quite strong. Below in §4.2 we consider some specific parametric refinements of some of these starting assumptions.

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

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

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

into (3) leaves us with the expression

$$P\Big((C,S){=}(\mathit{accepted},\mathrm{s}) \;\Big|\; \mathit{ideal\ obs.}\Big) = P\Big(S{=}s \;\Big|\; \mathit{ideal\ obs.}\Big) = \frac{\sum_Z P(V{=}\mathit{verified}\,|\,S{=}s,Z)P(S{=}s|Z)\sum_O P(T{=}\mathit{no\ failures}|O,S{=}s)P(OZ)}{\int_S \Big\{\sum_Z P(V{=}\mathit{verified}\,|\,SZ)P(S|Z)\sum_O P(T{=}\mathit{no\ failures}|OS)P(OZ)\Big\}\,dS} \tag{5}$$

for the updated probability density of S, given the *ideal observations* (for both argument legs). Here, the condition "$|$ *ideal obs.*" is a shorthand for "$|\,(V,T){=}(\mathit{verified,\ no\ failures})$". 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. (13) on p17 below.

It is the behaviour of this formula (5) for the distribution of the pfd S conditionally given the ideal observations that forms the focus of the remainder of the paper.

• **Verification Fallibility Against Correct Specification** Against a correct specification, an *infallible* verification procedure would pass the system precisely if its true pfd is zero: Provided the specification is correct, any positive pfd, however small, will always have been caused by a fault, which will certainly show up as a failure to verify the system. Conversely, one might assume that all systems which fail the verification have non-zero true pfds. Instead we introduce a pair of *verification fallibility* parameters α, ξ allowing the breakdown of both of these ideal behaviours of the verification process:

$$P(V{=}\mathit{verified}\,|\,S{=}s, Z{=}\mathit{correct}) = \begin{cases} \xi & \text{if } 0 < s \leq 1, \text{ or} \\ 1-\alpha & \text{if } s = 0, \end{cases} \tag{6}$$

Thus, our model allows that, against a correct specification:

– a perfectly reliable system *can* fail [9] the verification, with probability α;
– a system having a positive pfd *can* pass the verification, with probability ξ.

For simplicity, we assume that the probability (that is, conditionally given $\langle S{=}s, Z{=}\mathit{correct}\rangle$) of the latter kind of verification failure is independent of the actual (positive) value s of variable S. (Of course this constraint could be relaxed in future models, perhaps by using a parametric function $\xi(s)$.) Note

[9] 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.)

that the special case $(\alpha, \xi) = (0, 0)$ restores the *infallibility* assumption for the verification process (provided the specification is correct), as outlined above.

• **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.

$$P(V{=}\textit{verified} \,|\, S{=}s, Z{=}\textit{incorrect}) = 1\,. \tag{7}$$

• **Geometric Time-to-Failure Distribution** during testing. We assume that n test inputs cause failures independently with the operational pfd S, which are detected with certainty (and with no false alarms) if the oracle is *correct*. So the probability table of variable T has

$$P(T{=}\textit{no failures} \,|\, S{=}s, O{=}\textit{correct}) = (1-s)^n\,. \tag{8}$$

• **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:

$$P(T{=}\textit{no failures} \,|\, S{=}s, O{=}\textit{incorrect}) = 1. \tag{9}$$

• **Stochastic Ordering Constraint** We propose for the conditional distribution of S given Z a requirement for the following kind of stochastic ordering as a function of Z

$$P\left(S{>}s \;\middle|\; Z{=}\textit{correct}\right) < P\left(S{>}s \;\middle|\; Z{=}\textit{incorrect}\right), \qquad \text{for all } 0 \le s < 1\,. \tag{10}$$

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

4.2 Distributional Assumptions

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

We will use the symbol π 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, $\pi_{c*} = \pi_{cc}+\pi_{ci} = P(Z{=}\textit{correct})$. Later, we will sometimes display these prior probabilities of variables ZO using a 2×2 matrix layout

		O: correct	O: incorrect	
Z	correct	π_{cc}	π_{ci}	π_{c*}
	incorrect	π_{ic}	π_{ii}	π_{i*}
		π_{*c}	π_{*i}	1

(11)

This matrix represents an important set of prior beliefs, since it is here that is captured the dependence between our doubts about specification and oracle correctness. There is likely to be positive "assumption dependence" here, which will presumably cause dependence between the argument legs and undermine, to some extent, the efficacy of the two-legged approach.

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 $p_{0|c}$ and $p_{0|i}$, where the c or i indicates the conditioning value of Z. (We will require $p_{0|c}{>}p_{0|i}$ in compliance with assumption (10).) For the sake of statistical conjugacy [19, Ch. 9] with the discrete time model (8), we will use a beta distribution for the continuous component. So for $0{<}s{\leq}1$, use

$$\text{pdf}(s|Z) = \begin{cases} \frac{(1-p_{0|i})}{\beta(a,b)} s^{a-1}(1-s)^{b-1}, & \text{if } Z{=}\textit{incorrect}, \text{ or} \\ \frac{(1-p_{0|c})}{\beta(a',b')} s^{a'-1}(1-s)^{b'-1}, & \text{if } Z{=}\textit{correct}, \end{cases} \tag{12}$$

for some $a, b, a', b' > 0$. Note that if $\xi{=}0$, the latter of these two distributions is 'masked out', by the zero value in assumption (6), from the distributions (3,5,13), so that parameters a', b' disappear with ξ from the 'ideal observations' model in that case. See columns 5 and 6 (of 8) in Table 1 on p43. In other cases, in which $\xi{>}0$, note that our assumption (10) translates into a messy but computable constraint on $a, b, a', b', p_{0|i}, p_{0|c}$. (One may simply use analytic differentiation w.r.t. s, and then numerical zero-finding of monotonic functions, to determine the local minima of RHS−LHS in inequality (10), which, with the limiting values at the two end-points, $s \to 0, 1$ can be used to formulate the constraint that the inequality shall hold over the whole interval $0{\leq}s{<}1$.)

4.3 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 $P(S\&\textit{ideal obs.})$, meaning, more precisely

$$P(s\&\textit{ideal obs.}) \;=\; P\Big((C,S,V,T)=(\textit{accepted},s,\textit{verified},\textit{no failures})\Big)\,. \quad (13)$$

Table 1 on p43 shows how the value of (13) 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 (in each single column, the three entries in the three rows immediately under the double line that separates the headers) weighted also by the prior probability of the value of the pair ZO which selects the column. Without this $P(ZO)$ weighting factor, the column-product

$$P(VTS \,|\, ZO) = P(T \,|\, OS)P(S \,|\, Z)P(V \,|\, SZ)$$

is the probability (density, for S>0) of seeing the ideal observations VT, *and* of the pfd having a true (unknown) value S, conditioned (as if these were known) on specified ZO values. Note that the case s>0 becomes much simplified under the verification infallibility assumption $(\alpha,\xi) = (0,0)$, with the two zero ξ values in the third row of the body of the table causing some terms from subsequent rows to disappear: if we assume it impossible to verify an imperfect system against a correct specification, then a part of the total probability (13) disappears. The second to last row of the table gives the value of (13), 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 we have assumed C to be determined by VT in this ideal case (4).

For these *ideal observations* VT = (*verified, no failures*), the S=0–cases of the three equations (7-9) produce the six [10] 1's that occur in the left half of Table 1. These assumptions therefore mean that $P(\textit{ideal obs.}|S{=}0, Z{=}\textit{incorrect}, O) = 1$, and $P(\textit{ideal obs.}|S{=}0, Z{=}\textit{correct}, O) = 1{-}\alpha$. I.e., under our conservative assumptions, and irrespective of assumed correctness or otherwise of the oracle, the conditional probability of seeing the ideal observations from a system *known to be perfect* becomes certainty, if the specification is assumed *in*correct, and $1{-}\alpha$ if the specification is assumed correct.

[10] Note how just three equations produce six 1 s here essentially because the topology says that $T \perp\!\!\!\perp Z \,|\, OS$ and $V \perp\!\!\!\perp O \,|\, ZS$. 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.

The last row of Table 1 gives the *denominator of (5)*. This normaliser is just $P(\textit{ideal obs.}) = P\Big((V,T)=(\textit{verified,no failures})\Big)$. Substituting the entries of Table 1 into (5), gives the conditional probability $P\Big(S{=}0 \;\Big|\; \textit{ideal obs.}\Big)$ and also, for $S{>}0$, the conditional probability density $\text{pdf}\Big(S \;\Big|\; \textit{ideal obs.}\Big)$, 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 comment under the table.

5 Expressions for Confidence and Doubt

Substituting these further assumptions and abbreviated notations into the penultimate row of Table 1 on p43 (which originated from the numerator of (5)) gives the concentrated mass

$$P(S{=}0 \;\&\; \textit{ideal obs.}) = (1{-}\alpha)p_{0|c}\pi_{c*} + p_{0|i}\pi_{i*} \tag{14}$$

and, for strictly positive values of S, the pdf

$$\text{pdf}(s \;\&\; \textit{ideal obs.}) = \\ \xi\frac{(1-p_{0|c})}{\beta(a',b')}s^{a'-1}(1-s)^{b'-1}\left[\pi_{cc}(1-s)^n + \pi_{ci}\right] + \frac{(1-p_{0|i})}{\beta(a,b)}s^{a-1}(1-s)^{b-1}\left[\pi_{ic}(1-s)^n + \pi_{ii}\right], \\ 0 < s \leq 1. \tag{15}$$

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

$$P(\textit{ideal obs.}) = P(S{=}0 \;\&\; \textit{ideal obs.}) + \int_{0<s\leq 1} \text{pdf}(s \;\&\; \textit{ideal obs.})\, ds \tag{16}$$

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 (14) and (15) yields

$$P(\textit{ideal obs.}) = (1{-}\alpha)p_{0|c}\pi_{c*}{+}p_{0|i}\pi_{i*}{+}\xi(1{-}p_{0|c})\left[\pi_{cc}\mu' + \pi_{ci}\right]{+}(1{-}p_{0|i})\left[\pi_{ic}\mu + \pi_{ii}\right] \tag{17}$$

using notation μ for the n^{th} non-central moment of the beta distribution [11],

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

[11] —to be precise, of the beta distribution with its usual parameters a and b interchanged, because $\{1{-}X \text{ is distributed Beta}(b,a)\} \Leftrightarrow \{X \text{ is distributed Beta}(a,b)\}$.

in order to shorten some expressions in what follows. Note the dependence of μ and μ' on n, as well as on the beta distribution parameters. For any fixed $a, b, a', b' > 0$, we always have μ, μ' strictly decreasing in n, both being 1 at n=0, and having $\mu, \mu' \to 0$ as $n \to \infty$ (though the convergence can be slow for small a, a').

From (15) and (17), we can express the *doubt*, or probability of 'unsafe failure' of the entire two-legged assessment procedure represented by these modelling assumptions, with the formula

$$P(S{>}s \,|\, \textit{ideal obs.}) = \frac{\xi(1{-}p_{0|c})\left[\pi_{cc}\mu' I_{1-s}(b'{+}n, a') + \pi_{ci} I_{1-s}(b', a')\right] + (1{-}p_{0|i})\left[\pi_{ic}\mu I_{1-s}(b{+}n, a) + \pi_{ii} I_{1-s}(b, a)\right]}{(1{-}\alpha)p_{0|c}\pi_{c*} + p_{0|i}\pi_{i*} + \xi(1{-}p_{0|c})\left[\pi_{cc}\mu' + \pi_{ci}\right] + (1{-}p_{0|i})\left[\pi_{ic}\mu + \pi_{ii}\right]} \tag{18}$$

where I_{1-s} denotes the (regularised) *incomplete beta function* using the notation

$$I_x(a, b) = \frac{1}{\beta(a, b)} \int_0^x u^{a-1}(1-u)^{b-1}\, du, \qquad a, b > 0,\ 0{\leq}x{\leq}1, \tag{19}$$

which is a strictly increasing function of $x \in [0, 1]$ for all $a, b > 0$. [See [20, p944] for its other properties, which include $I_0(a, b) = 0$, $I_1(a, b) = 1$, and $I_{1-x}(a, b) + I_x(b, a) = 1$.] Equation (18), with the strict inequality in the probability on the left hand side, actually holds for all (non-negative) s, including s=0.

We shall refer to the conditional probability (18) 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 of the most important kind of 'argument failure' we first identified on p5. 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 ($S{>}s$) *and* will be (incorrectly) accepted by the two-legged argument as sufficiently safe (the *numerator* of (18)). 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 π's, and $p_{0|i}, a, b, p_{0|c}, a', b'$;

and in the latter case (the marginal probability $P(C{=}\textit{rejected})$) by the entire set of 12 independent model parameters.

We note again the significant degree of simplification occurring in the case $\xi{=}0$. In that case, not only do parameters a', b' become irrelevant to the posterior probability distribution (18) of S given the ideal observations, but also (18) depends on only 2 of the 3 independent degrees of freedom of the prior ZO distribution π. The conditional probability (18), in this special $\xi{=}0$–case of the *ideal observations* scenario, depends on the marginal probability $P(Z)$ and on the conditional probability $P(O|Z{=}\text{incorrect})$. Its lack of dependence on $P(O|Z{=}\text{correct})$ is explained by the fact that the infallibility assumption $\xi{=}0$ for the verification process in the case $(S{>}0, Z{=}\text{correct})$, given by the top line of (6) creates two zeros in the 5^{th} and 6^{th} 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 = correct (the pair of unequal entries in Table 1 located two rows above this pair of zeros). Conditionally given the ideal observation $Z{=}\textit{correct}$, this masked-out $(S{>}0, Z{=}\textit{correct})$ scenario is the only means by which our posterior distribution of S could have otherwise (without the masking effect of putting $\xi{=}0$ in assumption (6)) depended on the state of O. This is because the conservative assumption (9) removes any dependence on O from the LHS of Table 1, where $S{=}0$. As soon as we relax either the verification infallibility, $\xi{=}0$, or the conservative assumption (9) which interact in this simplifying way here, we obtain the greater complexity of a conditional probability $P(S{>}s \,|\, \textit{ideal obs.})$ which involves all 12 independent parameters of our parametric node probabilities.

If s is small, it may be numerically more accurate (depending e.g. on the precision properties of the incomplete beta algorithm at its extreme arguments) to compute instead the *confidence* using

$$
\begin{aligned}
&P(S{\leq}s \,|\, \textit{ideal obs.}) = \\
&P(S{=}0 \,|\, \textit{ideal obs.}) + P(0{<}S{\leq}s \,|\, \textit{ideal obs.}) = \\
&\frac{(1{-}\alpha)p_{0|c}\pi_{c*} + p_{0|i}\pi_{i*}}{(1{-}\alpha)p_{0|c}\pi_{c*} + p_{0|i}\pi_{i*} + \xi(1{-}p_{0|c})\left[\pi_{cc}\mu' + \pi_{ci}\right] + (1{-}p_{0|i})\left[\pi_{ic}\mu + \pi_{ii}\right]} + \\
&\frac{\xi(1{-}p_{0|c})\left[\pi_{cc}\mu' I_s(a', b'{+}n) + \pi_{ci} I_s(a', b')\right] + (1{-}p_{0|i})\left[\pi_{ic}\mu I_s(a, b{+}n) + \pi_{ii} I_s(a, b)\right]}{(1{-}\alpha)p_{0|c}\pi_{c*} + p_{0|i}\pi_{i*} + \xi(1{-}p_{0|c})\left[\pi_{cc}\mu' + \pi_{ci}\right] + (1{-}p_{0|i})\left[\pi_{ic}\mu + \pi_{ii}\right]}
\end{aligned}
\tag{20}
$$

To look briefly at some actual numbers, we will consider first an infallible verification assumption of the form $(\alpha, \xi) = (0, 0)$, which reduces the number of active model parameters significantly, from 12 to 7 for the reasons just explained. Under this simplifying assumption, all entries in the third row of

Table 1 on p43 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 3rd, 4th, 7th, and 8th, 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 affected by two kinds of 'infallibility' assumption, α=0 and ξ=0, for the verification activity. We cannot justify these from an argument for conservatism. The assumption ξ=0 affecting the 5th and 6th 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.

If, under this $(\alpha, \xi) = (0, 0)$ assumption, we set [12]

$$(p_{0|i}, a, b, p_{0|c}, n, \pi_{c*}) = (0.2, 1, 999, 0.5, 4602, 0.8). \tag{21}$$

in (18), we obtain doubt

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

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 our assumption ξ=0, it makes no difference to (18) how the probability of specification correctness, π_{c*}=0.8 is divided between the probabilities of (Z, O)=(*correct,correct*) and (Z, O)=

[12] An interesting figure to use for the number of test-cases in (8) is $n = 4,602$. This is the number of tests required to give a Bayesian 99% than 10^{-3} when no failures are observed, *under the simpler model assumptions used in [21]*. 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.

(*correct,incorrect*). However, the distribution of the other 20% $\pi_{i*} = 0.2$, does make a difference, and in fact can be used to change the doubt 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 values for doubt

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

respectively, while values of (π_{ic}, π_{ii}) between these two extremes lead to values intermediate between these two, producing a curve which is hyperbolic in form and monotonic increasing [13] as the 0.2 probability mass is steadily shifted from π_{ic} to π_{ii}. For instance, if we put $s{=}10^{-3}$, these two end points of this increasing hyperbolic segment are $P(S{>}s \mid \textit{ideal obs.}) = 0.00022$ at $\pi_{ic}{=}0.2$, and $P(S{>}s \mid \textit{ideal obs.}) = 0.098$ at $\pi_{ii}{=}0.2$

These numbers illustrate how the joint prior beliefs concerning the two unobservable variables ZO – incorporating both the prior doubt about Z and O individually, as well as beliefs about the *association* between their likely values – can influence the doubt about the claim produced by the two-legged argument. Although the model we are using is very simplified at this stage, we expect that such 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.

6 How Effective is this Multi-Legged Argument Approach in Gaining Confidence in Dependability Claims?

We have mentioned earlier that 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 (18) expresses the monotonic 'doubt function' $P(S{>}s \mid \textit{ideal obs.})$ of s as a function of the twelve independent model variables: $\pi_{ic}, \pi_{ii}, \pi_{cc}, p_{0|c}, p_{0|i}, a, b, a', b', \alpha, \xi, n$: 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? How does the two-legged argument doubt (18) compare with its 'naive independence' version in which,

[13] assuming $s \neq 0, 1$

for a fixed claim, the doubts emanating from the two single argument legs are simply multiplied to produce the supposed two-leg-argument doubt?

6.1 Benefit of the Two-legged Argument and its Dependence on Stochastic Association Between Doubts in Assumptions for each Leg

We show next that it is possible to assign the parameters of our two-legged model such that we leave only one value possible, either of V, or of T, respectively, or of each of V and T. This single value is, in each case, the "ideal observation" value V= *verified* or T= *no failures*. Thus, we can choose parameters which make, in (2) and the equations following, $P(V$=*verified*$|SZ)$, or $P(T$=*no failures*$|OS)$, respectively, become a constant, 1, (so no longer depending on the values of SZO). Note that this procedure of altering the model's local conditional probability tables until only one value of a variable (of V, or of T, here) has positive probability is mathematically distinct from summation over that variable (with tables such that two or more of its values have positive probability). It is easy to verify in the case of our model that the first procedure does in fact produce a factorization of the joint distribution which – at least in the cases of "ideal observations" – is identical to that which would arise from one or other of our proposed single argument topologies Figures 2 and 3 on p10. More precisely stated: although it is possible to derive, for each of the two single-legged arguments of Figs. 2 & 3, a parametric representation of the 'doubt' (18), just as we have done above for the two-legged argument, it is actually simpler, and in fact equivalent (in the case where we condition on ideal observations), to deduce directly the analogous consequences of the single-leg arguments of Figures 2 & 3 as *special cases of the results for the combined argument* obtained by the following special parametric assignments.

Our removal of testing evidence from the argument is equivalent to substituting n=0 in (18) to produce a doubt

$$P(S{>}s \,|\, V{=}\textit{verified}) = \frac{\xi(1{-}p_{0|c})\pi_{c*}I_{1-s}(b',a') + (1{-}p_{0|i})\pi_{i*}I_{1-s}(b,a)}{\Big[(1{-}\alpha)p_{0|c} + \xi(1-p_{0|c})\Big]\pi_{c*} + \pi_{i*}} \qquad (22)$$

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 $(\alpha, \xi){=}(0,1)$ in (18)

to produce

$$P(S{>}s \,|\, T{=}\textit{no failures}) = \frac{(1{-}p_{0|c})\Big[\pi_{cc}\mu' I_{1-s}(b'{+}n, a') + \pi_{ci} I_{1-s}(b', a')\Big] + (1{-}p_{0|i})\Big[\pi_{ic}\mu I_{1-s}(b{+}n, a) + \pi_{ii} I_{1-s}(b, a)\Big]}{p_{0|c}\pi_{c*} + p_{0|i}\pi_{i*} + (1{-}p_{0|c})[\pi_{cc}\mu' + \pi_{ci}] + (1{-}p_{0|i})[\pi_{ic}\mu + \pi_{ii}]} \tag{23}$$

for our *ideal observations* case of the single, testing-argument leg. To see this [14], consider the third row (of 5 rows) in the body of Table 1 on p43. The above substitutions make all entries in this row equal. That is to say, the substitutions make $P(V{=}\textit{verified}|ZS)$ independent of the values of both Z and S, in just the manner we described at the beginning of this subsection. (Consider the effect on equations (3,5).) It is not difficult to confirm from our algebraic conclusions 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 2, as required.

We now compare some plots of the three doubt functions (18, 22, & 23), for fixed $s{=}10^{-3}$, as we vary certain other model parameters. The plots of

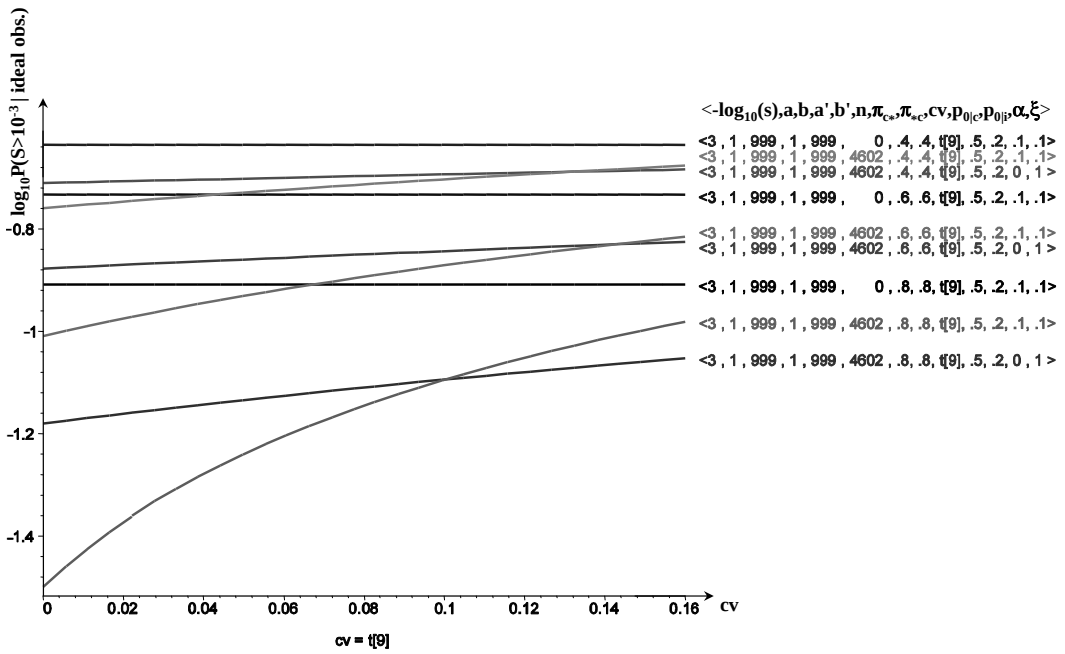

Fig. 4. Doubt functions (18) vs. *cv* – with params. as in (24) and following text.

Figure 4 show the value of $\log_{10}$ of the doubt function (18) and were produced by setting the threshold s value to 10^{-3} and the model parameters

$$(p_{0|i}, a, b, p_{0|c}, a', b') = (0.2, 1, 999, 0.5, 1, 999). \tag{24}$$

[14] 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 nine plots shown may be thought of as three groups of three. These groups were produced by setting (n, α, ξ) 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 *two-legged argument*, which may be compared against *the second two* which are the special cases (23) and (22) of (18), corresponding respectively to *testing-only*, and *verification-only, single-legged arguments*, with the other parameters (24) 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 π-matrix as follows. π was reparameterised in terms of the triple (π_{c*}, π_{*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).

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

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 π_{c*} and π_{*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 4 to state the following descriptive observations about the shapes of the 9 curves. We will refer to individual curves here by numbering them $\mathbf{1}\ldots\mathbf{9}$, counting vertically upwards at the right-hand end of the graph, with the lowest plot numbered $\mathbf{1}$. (Because of some curves crossing, this means that the plots at the left-hand end of the graph, are numbered in the order $\mathbf{2, 1, 5, 3, 4, 8, 6, 7, 9}$, moving upwards):

- The three flat, horizontal plots $(\mathbf{3, 6, 9})$ 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 π_{c*} of Z (22).
- The prior marginals (π_{c*}, π_{*c}) have the values $(0.8, 0.8)$ in plots $(\mathbf{1, 2, 3})$, $(0.6, 0.6)$ in plots $(\mathbf{4, 5, 6})$, and $(0.4, 0.4)$ in plots $(\mathbf{7, 8, 9})$. For each fixed value of the prior marginals, towards the left-hand end (small cv) of the

graph [15], 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 π_{c*}, and π_{*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 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**)).

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

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

for the correctness of the oracle and specification, accordingly as the marginal correctness probabilities π_{c*}, π_{*c} are assigned values 0.8, or 0.6, or 0.4. These three π-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 $(0.23, 0.20, 0.19)$. Notice that the 2-legged argument gives the greatest confidence in each case.

In comparing the two-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 is the first of several, at first sight, counter-intuitive model behaviours that we shall meet. We see below that our current model topology, with its uncertainties concerning a potentially defective oracle or specification and use of conservative assumptions,

[15] but continuing, as before, to refer to the plots by number, according to their order at the *right-hand* end

allows some complex kinds of model behaviour which may either 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 with which, it will perhaps eventually be possible to argue, all competent expert beliefs will necessarily conform.

The examples above all involve a symmetric π-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 two-legged and single-legged arguments, based on this model. To examine one more numerical example, which has a slightly higher prior "covariance" $cv = 0.075$ between Z and O correctness, but coupled with some rather more optimistic values of other parameters, and a greater 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 α=0.01, ξ=0.04, and slightly more optimistic prior beliefs than above about the likely values of S when Z is incorrect,

$$(p_{0|i}, a, b, p_{0|c}, a', b') = (0.4, 1, 999, 0.5, 1, 999)\,, \tag{25}$$

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 $P(S > 10^{-3} | \textit{ideal obs.})$, 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.

6.2 Doubt as a function of the number n of test cases

Under *'ideal observations'* (no detected failure), one might expect, as with earlier models [21], that S should stochastically decrease – in terms of its posterior distribution (18), or (23) 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.

We will not attempt to solve for the general parametric assumptions required such that our two-legged argument, with testing evidence incorporated, should provide higher confidence than the single, verification-only leg (the n=0 case (22) of (18)). We only note that for 'very large' n this is so whenever (22) exceeds (18) with 0 substituted for μ and μ'. This is because of the limiting property of the beta moments noted on p18 (recalling that the incomplete beta terms in (18) are bounded by 1 as $n\to\infty$). This is the case in all of the

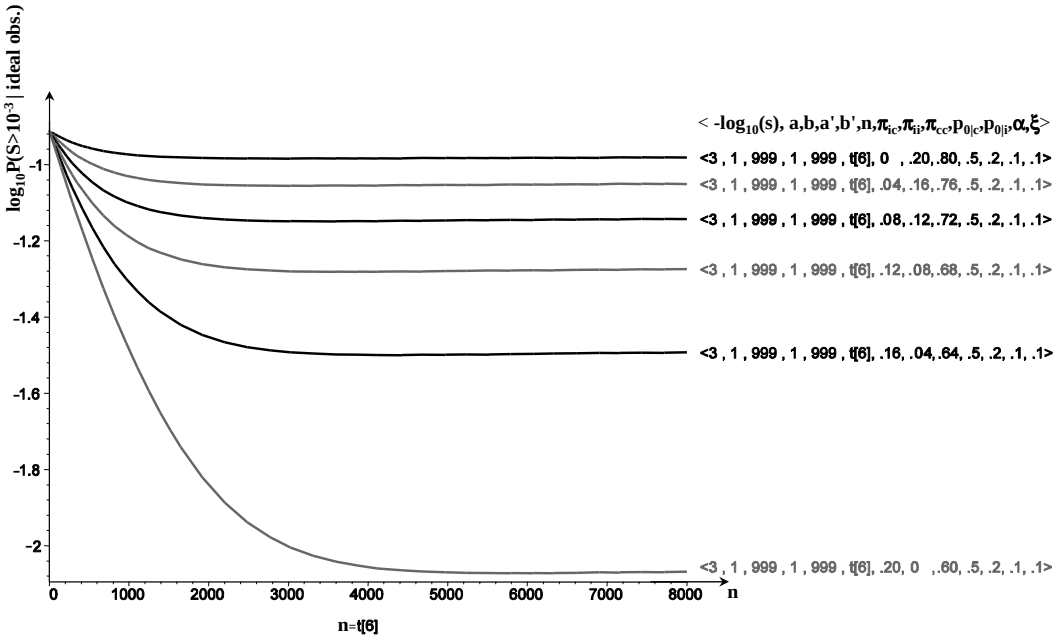

Fig. 5. Doubt functions (18) vs. n – with params. as in (26)

example plots of Figure 5. These show the value of $\log_{10}$ of the doubt function (18) plotted against the number n of test cases, using a claim s=10^{-3}, and model parameters

$$(p_{0|i}, a, b, p_{0|c}, a', b', \alpha, \xi) = (0.2, 1, 999, 0.5, 1, 999, 0.1, 0.1)\,. \tag{26}$$

The plots differ from each other only in the values of the π matrix, which were produced by fixing the marginal probabilities of specification/oracle-correctness at π_{c*}=π_{*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 4 above. Notice that there appears to be a very 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) occurring before the limiting $n \to \infty$ value is approached. We give some explanation of this quirk in the next section, and of other model characteristics which might appear counter-intuitive on first inspection.

7 Some Counterintuitive Results

7.1 Supportive and Non-Supportive Single Argument Legs

As we have seen in the previous section, the acquisition of evidence that is 'obviously good news' – what we have called ideal evidence – can sometimes result in a *reduction* in confidence. In this section we shall examine in some detail examples of such apparently non-intuitive results. Our approach is to use some numerical search and optimization tools to investigate the question of whether or not, under the present model, the arrival of ideal 'supportive' evidence results in increased confidence in the pfd.

We begin with a *single* argument. At first sight it might appear self-evident what 'supportive evidence' should look like: For a verification leg, the evidence is supportive precisely if the system is *verified correct*. For a testing leg (though there may also be a grey area), it seems at first sight clear that *no failures*, or very few failures, amongst a large number of test cases is supportive evidence. In contrast, a system's failure to be verified correct, or to succeed sufficiently often during test, should reduce our confidence.

In fact we can show that this is not always the case. Consider, for example, a testing argument leg in which no failures have been observed among the test cases, and the parameter values of the model are: s=0.001, n=17,921, a= 2.58276, b=4.77020, a'=16.68483, b'=41,133.7, p_{0i}=2.00200×10^{-3}, p_{0c}= 4.21724×10^{-3}, and

$$\pi = \begin{bmatrix} 0.994192 & 1.63910\times10^{-3} \\ 7.81537\times10^{-5} & 4.09042\times10^{-3} \end{bmatrix}.$$

The beliefs about the unobservable variables ZO are changed by this evidence from the above prior π to

$$P(ZO \mid T) = \begin{bmatrix} 0.53406 & 0.13329 \\ 1.2724\times10^{-5} & 0.33263 \end{bmatrix}.$$

The prior confidence in the claim here is 0.99583, but the posterior confidence, in spite of the extensive 'good news' from testing, is *decreased* to 0.66803. At

first glance such a result is quite surprising and counter-intuitive. On closer inspection, however, we believe there is an intuitive explanation in terms of two 'rival explanations' for a long period of failure-free testing. Reasoning informally, guided by the property of adjacency in the topology of Figure 2 on p10, we see that on the one hand extensive failure-free testing, applied as evidence to node T, may be taken to indicate a low system pfd S. Realistic node conditional probabilities for T will surely capture this effect. But, symmetrically on the left-hand side of Figure 2, a parallel *negative* inference of a *potentially low oracle quality* O also suggests itself, as a rival explanation of the apparently positive testing evidence: perhaps a defective oracle is missing failures. These two inferences from T may *both* follow when n becomes very large without failure. One can imagine a tension arising at node S between these two competing tendencies of the testing evidence. This raises the question of how the two inferences interact, and whether one 'dominates' in its effect upon posterior beliefs about S. The answer will depend partly on the conditional probability distribution $P(T|OS)$. (Note that – while in no sense *essential* to this explanation – the 'conservative' assumption (9) would appear to increase the viability of the second potential explanation for apparently successful testing.) But our model in Figure 2 also contains a prior belief structure 'higher up', usually with stochastic association of some kind between the variables ZOS. The combined effect of the two available inferences from observing large n without failure will depend also on this 'upper part' of the topology. In particular, 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 it seems plausible that, for certain model parameter values, *evidence of a defective oracle could have the effect of reducing confidence* in a low system pfd. E.g., in the last numerical example the increased doubt in the correctness of the oracle (sum of terms in last column of matrices above) is associated with an increased doubt about the correctness of the specification (sum of terms of last row of the matrices). Thus when we see very many failure-free test cases we may increase our mistrust in the oracle, increase our mistrust in the specification *and thus increase our mistrust in the pfd* – see constraint (10). We stress that this informal conception of competing effects at S of successful testing is not intrinsic to the topology of Figure 2, but relies heavily on node conditional probability assumptions. The model topology is sufficiently flexible to allow such possibilities, but whether or not they are found will depend also on the 2×2 prior matrix π for ZO, the parameter values used for our two point-mass-augmented beta distributions (12) for $P(S|Z)$, and on the conditional probabilities (8) & (9) comprising $P(T|OS)$.

Examples of non-supportive verification arguments may also be obtained. These might be explained by a similar inference concerning the likely correctness of Z, whenever the model parameter values are such as to create

the needed associations. For example, a successful system verification with s=0.001, a=1.2742, b=0.2106, a'=3.2095, b'=27, 095, α=0.3950, ξ=1.2006×10^{-4}, p_{0i}=1.5547×10^{-4}, p_{0c}=1.3812×10^{-3}, π_{c*}=0.9997156, π_{i*}=2.844×10^{-4} gives a decrease from a high prior confidence 0.99972 to the much lower 0.77064 after the positive verification. We suggest that while this, perhaps counter-intuitive, behaviour may well have been accentuated by our conservative assumption (7) (for the verification outcome in the case that the specification Z=*incorrect*), it would still be possible after a degree of relaxation of this conservative assumption, given the 'right' conditional probabilities applied to our topology. In more detail, our tentative explanation for this effect is as 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. Specifically, beliefs about the unobservable variable Z become changed (by the news of the successful system verification) from the prior probability 0.9997156 of Z's correctness to $P(Z{=}correct \,|\, V) = 0.77060$. This opens a chain of subsequent inference of the following kind. The prior *stochastic association between correct* Z *and small* S required by (10) gives this *increased* Z*-doubt* a tendency to increase the probability of large S.

In the case of the numbers used above, we can perhaps best illustrate this by dividing the range of S up into two bins $[0, 10^{-3}]$ and $(10^{-3}, \infty)$. Using the same 2×2 matrix layout that we have used above for joint beliefs about ZO, we can compute the prior beliefs

$$P(ZS) = \begin{bmatrix} 0.99971563 & 1.09403\times10^{-9} \\ 5.18016\times10^{-8} & 2.84319\times10^{-4} \end{bmatrix}.$$

which quantify (an aspect of) the association between Z and S mentioned above. Compare the posterior counterpart of these beliefs, having observed the ideal verification evidence

$$P(ZS \,|\, V) = \begin{bmatrix} 0.77060 & 1.05960\times10^{-10} \\ 4.17888\times10^{-5} & 0.22936 \end{bmatrix}.$$

which illustrates how the doubt cast on Z by the successful verification has caused belief to 'shift along the main diagonal' from ($Z{=}$ *correct*, $S{\leq}10^{-3}$) to ($Z{=}$ *incorrect*, $S{>}10^{-3}$).

We do not assert that the chains of inference shown in these examples will be realistic in every model application. Much will depend on the parameter values assigned to the node conditional probabilities, and how, for example, these stochastically associate variables Z, O, and S. We might well later choose to replace the assumptions we have called 'conservative' by alternatives that we

might elicit as real experts' beliefs, about the likely effects of defective oracles and specifications. However, our examples of quite subtle interactions between evidence, assumptions, and assumption doubt do illustrate the richness of the kinds of competing inferences modelled within even a highly simplified BBN structure, and alert us to the fact that symbolic analysis may identify model consequences that initially will seem surprising and counter-intuitive.

7.2 Adding a Supportive Leg May Not Improve Confidence

Having made these observations that the effects of 'ideal' evidence are not always of the kind one might naively expect, we now focus on *improvement in confidence* as our stricter definition of a 'supportive argument' (as opposed to mere 'ideal evidence': V= *verified* and/or T= *no failures*). Thus, we will call an argument *supportive if it improves upon the prior confidence*. So the testing leg with outcome T is called 'supportive' whenever $P(S{\leq}s \,|\, T) > P(S{\leq}s)$. Similarly a verification leg with outcome V is called 'supportive' whenever $P(S{\leq}s \,|\, V) > P(S{\leq}s)$. The dual-legged argument is 'supportive' when $P(S{\leq}s \,|\, VT) > P(S{\leq}s)$.

For each set of model parameters, there are actually four different cases for which one can ask whether a confidence improvement occurs on receipt of ideal evidence from an argument leg. A system assessor may obtain either *testing-leg* evidence T= *no failures*, or formal *verification-leg* evidence V= *verified*, in each case with or without evidence of the other kind being already present. We explained in §6.1 that, as our model stands, all these four questions may be framed mathematically as the questions as to whether a single expression, which can be either of (18) or (20), increases or decreases when ideal evidence is received. In terms of this expression, ideal testing evidence is added by changing from n=0 to a general non-zero n value (and making the associated change to μ and μ', from 1 to the corresponding values). Similarly ideal verification evidence is added by changing from (α, ξ)=$(0, 1)$ to the general pair.

Thinking of the values of our doubt expression (18) as laid out as a square with each side corresponding to the receipt of "ideal evidence" V or T, we can depict the conditions described in the questions above diagrammatically. In our diagrams, a downward arrow corresponds to the receipt of ideal testing leg evidence and a rightward arrow corresponds to the receipt of ideal verification leg evidence. So the top, left vertex corresponds to prior confidence and the bottom right to the final confidence after evidence from both legs is in. Following an arrow represents 'progress', in the sense that further ideal evidence is received. The '$>$' (beside a right-pointing arrow) and '$\vee$' (beside a downward-pointing arrow) symbols are used to indicate a decrease in confi-

dence as we move along the arrow. Thus the counter-intuitive situations which we have called non-*supportive* single legs look like

$$\begin{array}{ccc} 00 & \xrightarrow{>} & 0V \\ \downarrow & & \downarrow \\ T0 & \longrightarrow & TV \end{array} \tag{27}$$

for verification leg evidence, and like

$$\begin{array}{ccc} 00 & \longrightarrow & 0V \\ \downarrow_{\vee} & & \downarrow \\ T0 & \longrightarrow & TV \end{array} \tag{28}$$

for testing evidence. In these diagrams a *zero* represents the *absence of evidence* for one or other leg.

We are now in a position to ask whether counter-intuitive results are possible within our model for two-legged arguments. It is worth recalling here the *systems* metaphor that underpins the intuition behind the use of such arguments. It is well-known that the reliability of a 1-out-of-2 system will always be greater than or equal to the reliability of the best of the two components. Is this true of our two-legged arguments: in particular, is it always the case that confidence will increase if we add a second *supportive* argument leg to an initial supportive leg? That is, can either of the following two cases occur:

$$\begin{array}{ccc} 00 & \xrightarrow{<} & 0V \\ \downarrow_{\wedge} & & \downarrow \\ T0 & \xrightarrow{>} & TV \end{array} \tag{29}$$

or

$$\begin{array}{ccc} 00 & \xrightarrow{<} & 0V \\ \downarrow_{\wedge} & & \downarrow_{\vee} \\ T0 & \longrightarrow & TV \end{array} \tag{30}$$

We have examined this question numerically and tentatively conclude that, for our current model, the answers are: no, the scenario depicted in (30) is ruled

out; but yes it *is possible* for the case of (29) to occur, in which a supportive verification leg can *depress* confidence when it is added to pre-existing ideal testing evidence, which itself is supportive when considered alone.

For a numerical example of this case (29), consider the two-legged argument with evidence s=0.001, n=10,006, a=0.0807, b=0.0192, a'=8.2408×10^{-3}, b'=0.044813, α=0.12419, ξ=4.9315×10^{-6}, $p_{0\not C}$=6.91181×10^{-3}, p_{0C}=9.69767×10^{-3}, π^i=$\begin{bmatrix} 0.99965 & 5.50587\times10^{-6} \\ 1.19185\times10^{-5} & 3.28401\times10^{-4} \end{bmatrix}$.
With these parameters, the prior confidence 0.8001 is considerably improved, to 0.9659, by a positive system verification outcome. *Without* any such verification leg, but with instead ideal testing evidence from the 10,006 trials, confidence improves to as much as 0.999627. However, when the positive verification evidence is added, as one of two legs, to this ideal testing outcome, then much of this large confidence gain from the testing leg alone is *lost*, with a confidence now of only 0.9671, which – while still an improvement over the prior confidence – is not *as great* an improvement as was obtained from the testing leg alone. I.e., the act of *adding* a supportive verification leg *to* a testing leg has actually *lowered* our confidence in the claim. When comparing the confidence resulting from a two-legged argument against the prior confidence and against that from each single leg individually, we can lay out these four confidence values in a 2×2 matrix format

$$P(S{\leq}10^{-3}\,|\,obs.) = \begin{pmatrix} P(S{\leq}10^{-3}) & P(S{\leq}10^{-3}\,|\,V) \\ P(S{\leq}10^{-3}\,|\,T) & P(S{\leq}10^{-3}\,|\,VT) \end{pmatrix} = \begin{pmatrix} 0.8001 & 0.9659 \\ 0.999627 & 0.9671 \end{pmatrix} \tag{31}$$

matching the layout of our square diagrams above. We have used curved brackets around the matrix to distinguish it visually from our other frequently used 2×2-matrix notation for joint distributions of variables ZO (for which we will reserve square-bracketed matrix notation). Using the same layout as that in (27) to (30) – with the prior beliefs as the top left hand square-bracketed matrix, and the beliefs emanating from a two-legged argument at the bottom right, etc – the four possible sets of beliefs about the ZO pair for this example,

under the four possible conditions of evidence discussed, appear as

$$P(ZO \mid obs.) = \begin{pmatrix} P(ZO) & P(ZO \mid V) \\ P(ZO \mid T) & P(ZO \mid VT) \end{pmatrix}$$

$$= \left(\begin{bmatrix} 0.999654 & 5.5059\times10^{-6} \\ 1.1919\times10^{-5} & 3.2840\times10^{-4} \end{bmatrix} \begin{bmatrix} 0.96148 & 5.2956\times10^{-6} \\ 1.3489\times10^{-3} & 0.037168 \end{bmatrix} \right.$$

$$\left. \begin{bmatrix} 0.999572 & 7.0415\times10^{-6} \\ 1.4354\times10^{-6} & 4.2000\times10^{-4} \end{bmatrix} \begin{bmatrix} 0.96264 & 5.3027\times10^{-6} \\ 1.2720\times10^{-4} & 0.037218 \end{bmatrix} \right) \tag{32}$$

Some further insight into the apparently counter-intuitive result – that adding a supportive verification leg to the testing leg can reduce confidence in the dependability claim – comes from examining these intermediate numerical results. Consider first the prior belief represented by the top-left matrix in (32): in particular note that the leading diagonal of this matrix suggests that the *a priori* beliefs about Z and O are positively associated. That is, belief that the specification is incorrect results in a stronger belief the oracle is incorrect, and vice-versa.

Seeing 10,006 test cases executed without failure, in the first argument leg, results in increased doubt about the correctness of the oracle: from 5.5059×10^{-6} $+3.2840\times10^{-4} = 3.3391\times10^{-4}$ *a priori* to $7.0415\times10^{-6}+4.2000\times10^{-4} = 4.2704\times10^{-4}$ after the test data is seen. Because of the association between beliefs about Z and O, this supportive evidence from testing undermines confidence in the specification correctness: doubt increases from $1.1919\times10^{-5}+3.2840\times10^{-4} = 3.4032\times10^{-4}$ *a priori*, to $1.4354\times10^{-6}+4.2000\times10^{-4} = 4.2144\times10^{-4}$ after seeing the testing evidence.

The verification leg is supportive when we have no testing evidence (i.e. it increases confidence from its *a priori* value). But the above reasoning shows that the presence of (successful) testing can undermine the contribution that the verification leg makes to overall confidence in the dependability claim when both argument legs are present. In fact this undermining can be so severe that adding the verification leg makes things worse, compared with having only the testing leg.

Examination of the parameter values used to construct this example might cause one to conclude that some of them seem unlikely to be realistic beliefs of experts about real systems. E.g. consider the virtual prior certainty, according to these parameters, that the specification is correct; or the asymptotes $(b, b' < 1)$ of the two beta distributions at the 1 end of the unit interval $[0, 1]$. It remains

to determine whether actual realistic beliefs could also exhibit property (29).

Numerically we have been unable to obtain a similar example with T and V interchanged, i.e. satisfying (30). We conjecture that such an example may prove to be analytically impossible for our current model, while we impose our stochastic ordering constraint (10). Although we have not obtained experimentally the reversal of ordering occurring on the right hand side of (30), one finding that is almost as surprising is that we can find examples in which – in our terminology explained earlier – a significantly supportive testing leg, when added to a significantly supportive verification leg creates only a negligibly small improvement in confidence of the two-legged argument over the confidence obtained from the verification leg alone. A numerical example of this is provided by the values s=0.001, n=19,921, a=0.092728, b=2.4768, a'=0.13423, b'=3.8705, α=9.8691×10^{-3}, ξ=2.8029×10^{-7}, $p_{0\not C}$=1.39760×10^{-3}, p_{0C}=0.18737

$$\pi = \begin{bmatrix} 0.47491 & 0.09055 \\ 1.80783\times10^{-4} & 0.43436 \end{bmatrix}.$$

With these values, the ideal evidence produces three confidence levels, laid out next to the prior confidence as in the last example

$$P(S\leq10^{-3} \mid obs.) = \begin{pmatrix} 0.59125 & 0.67018 \\ 0.70759 & 0.67025 \end{pmatrix}. \tag{33}$$

Here, the supportive testing evidence produces an improvement of the two-legged over the verification-only argument which is less than one in the fourth-significant decimal digit.

8 Special Case of Claims for Perfection, S=0

If, instead of a claimed upper bound $S\leq s$ on pfd, we make a claim for perfection, i.e. S=0, then we obtain a special case of the above expressions for confidence and doubt for which certain of the counter-intuitive results demonstrated above become no longer possible. This substitution corresponds to the special case where our confidence refers to a claim that the system is *perfectly reliable*, rather than merely that its pfd *does not exceed some positive threshold* value $S\leq s>0$.

Firstly, we can show that for such a perfection claim, operation that is completely failure-free throughout testing *always* constitutes a *supportive* argument leg in the sense we identified in §7.1, for the simple reason that our model assumptions clearly make S=0 and S=0 & T identical events. Thus, we

must have a confidence, from the testing leg only given by

$$P(S{=}0\,|\,T) = \frac{P(S{=}0\,\&\,T)}{P(T)} = \frac{P(S{=}0)}{P(T)} \geq P(S{=}0)\,, \tag{34}$$

that is no less than the prior confidence. Further, for this perfection claim we can prove easily the special case of the result that, for the more general claim, we have been so far able to verify only numerically without an analytic proof: ideal testing evidence added to a positive verification outcome to produce a two-legged argument *always* [16] improves upon the confidence emanating from the verification leg alone. Essentially the same short proof as that given above works again here. Our model assumptions make $S{=}0\,\&\,V$ and $S{=}0\,\&\,VT$ identical events. So we have a final confidence $P(S{=}0\,|\,VT)$ in this perfection claim from the *two*-legged argument which exceeds the confidence $P(S{=}0\,|\,V)$ emanating from the *single*-leg verification argument, because

$$P(S{=}0\,|\,VT) = \frac{P(S{=}0\,\&\,VT)}{P(VT)} = \frac{P(S{=}0\,\&\,V)}{P(VT)} \geq \frac{P(S{=}0\,\&\,V)}{P(V)} = P(S{=}0\,|\,V)\,. \tag{35}$$

9 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 also artificially reduced some of the state spaces to Boolean. Finally, 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 [22]. Simply populating the node probability tables in this way results only in a single numerical posterior distribution for the goal variable S.

[16] even without a restriction that the verification leg must be *supportive*

Our intention here was to obtain greater insight into the factors that determine the efficacy of arguments, and in particular of multi-legged arguments. We started from the position that efficacy would be judged by the confidence that the arguments engendered in dependability claims. We wanted a better understanding of how confidence is determined by factors such as the doubt in the truth of assumptions underpinning the arguments. In particular, we sought to understand the importance of association – e.g. between the doubts associated with assumptions for different argument legs – in determining the effectiveness of the multi-legged argument approach.

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 [23,24]. 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.

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). It turns out that this analogy breaks down in surprising ways.

One way in which the systems metaphor breaks down concerns composability of arguments. Our example illustrates that it can be straightforward to *decompose* a model for a two-legged argument, producing two derived, single-leg models as special cases corresponding to a degenerate observation for one or other argument leg. However there is no standard reverse operation of *composition* starting from a given pair of single-leg models. The representation of *dependence* via variables which link the two argument legs is an additional and difficult modelling task, whose solution is integral to any meaningful model of a two-legged argument.

Another surprising way in which the systems/arguments analogy breaks down concerns efficacy: 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 drawing simplistic – albeit intuitively plausible – conclusions. It cannot be too strongly emphasised that it is only through the completely analytical treatment – difficult though it is – that we get these insights.

On the other hand, it is clear that in many cases – as might be expected – multi-legged arguments *do* bring benefits in terms of increased confidence in dependability claims compared with single arguments, as we have shown in Section 6. From a practical point of view, we would like to know exactly when such benefits can be expected, and how extensive they might be. Ideally, we would like to be able to design multi-legged arguments – before the expensive process of evidence-collection begins – so that confidence in a dependability claim will be gained *most cost-effectively*.

We do not claim that our work here enables this to be done. But we do believe that it is a useful beginning in understanding some of the key issues. Thus, for example, in Section 6 we show how it is possible with our analytical treatment of the BBN to perform 'what if' calculations on the effect of dependence, in assumption doubts for the two legs, upon the confidence that the two-legged argument provides in the claim. It is easy to see how this kind of study could be used to compare different possible multi-legged arguments before committing to the expense of deploying them in a particular dependability case.

Concerning the general efficacy of diverse arguments in real-life applications, 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). 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? The results presented here are largely neutral on such issues: they concern what *might* happen, rather than what *will* happen when real

experts assess real systems. It is worth stating, however, that when we constructed the simplified example used in the paper, we did not anticipate those consequences that we have called 'counter-intuitive': it seems possible, even after considerable reflection, to be surprised by what is implied by a complex model. This is likely to be true *a fortiori* for more realistic, and thus more complex, models.

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. We are extremely grateful to our three reviewers for their very thorough readings of, and thoughtful comments on, an earlier version of this paper. Their comments prompted us to conduct a thorough rewrite of the paper which has, we believe, improved it greatly.

References

[1] Ministry of Defence, Requirements for safety related 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] D. E. Eckhardt, L. D. Lee, A theoretical basis for the analysis of multi-version software subject to coincident errors, IEEE Trans. on Software Engineering 11 (1985) 1511–17.

[5] B. Littlewood, D. R. Miller, A conceptual model of multi-version software, Digest of 17^{th} Fault Tolerant Computing Symposium (1987) 150–55.

[6] J. C. Knight, N. G. Leveson, An experimental evaluation of the assumption of independence in multi-version programming, IEEE Trans. on Soft. Eng. 12 (1986) 96–109.

[7] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference, Mathematics and Its Applications, Morgan Kaufmann, San Mateo, California, 1988, revised 2nd printing 1991.

[8] 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/studeny/volstu.ps.

[9] 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/ (2003).

[10] A. P. Dawid, Conditional independence in statistical theory, Journal Royal Statistical Society, Series B 41 (1) (1979) 1–31, with discussion.

[11] 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.

[12] S. L. Lauritzen, Graphical Models, Oxford Statistical Science Series, Clarendon Press, Oxford, 1996.

[13] G. Shafer, Probabilistic Expert Systems, CBMS-NSF Regional Conf. Ser. in Applied Math., Society for Industrial & Applied Mathematics, Philadelphia, 1996.

[14] A. P. Dawid, Conditional independence for statistical operations, Annals of Statistics 8 (3) (1980) 598–617.

[15] 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).

[16] 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.

[17] D. Wright, Elicitation and validation of graphical dependability models, in: S. Anderson, M. Felici, B. Littlewood (Eds.), SAFECOMP 2003, Edinburgh, UK, 23-6 Sept., Springer-Verlag Heidelberg, 2003, pp. 8–21.

[18] G. de Barra, Measure Theory and Integration, Mathematics and Its Applications, Ellis Horwood, Chichester, 1981.

[19] M. H. DeGroot, Optimal Statistical Decisions, Wiley, 2004, wiley Classics Library Edition. ISBN 0-471-68029-X.

[20] M. Abramowitz, I. A. Stegun (Eds.), Handbook of Mathematical Functions, Dover, New York, 1970, http://www.math.sfu.ca/∼cbm/aands/.

[21] B. Littlewood, D. Wright, Some conservative stopping rules for the operational testing of safety-critical software, IEEE Trans. on Software Engineering 23 (11) (1997) 673–83.

[22] 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.

[23] S. Yih, C.-F. Fan, Search for the unnecessary, Nuclear Engineering International (2001) 24–6.

[24] 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 $P(S\&ideal\ obs.)$

S	0				>0			
Z	*correct*		*incorrect*		*correct*		*incorrect*	
O	*correct*	*incorrect*	*correct*	*incorrect*	*correct*	*incorrect*	*correct*	*incorrect*
$P(T{=}no\ failures\,\|OS)$	1	1	1	1	$(1{-}S)^n$	1	$(1{-}S)^n$	1
$P(S\|Z)$	$P(S{=}0\|Z{=}corr.)$		$P(S{=}0\|Z{=}incorr.)$		$P(S\|Z{=}corr.)$ (a pdf)		$P(S\|Z{=}incorr.)$ (a pdf)	
$P(V{=}verified\|SZ)$	$1-\alpha$	$1-\alpha$	1	1	ξ	ξ	1	1
$P(S\&ideal\ obs.)$	$(1{-}\alpha)P(S{=}0\|Z{=}corr.)P(Z{=}corr.)+P(S{=}0\|Z{=}incorr.)P(Z{=}incorr.)$				$\xi P(S\|Z{=}corr.)\Big[(1{-}S)^n P\left(ZO{=}(corr.,corr.)\right)+P\left(ZO{=}(corr.,incorr.)\right)\Big]$ $+P(S\|Z{=}incorr.)\Big[(1{-}S)^n P\left(ZO{=}(incorr.,corr.)\right)+P\left(ZO{=}(incorr.,incorr.)\right)\Big]$			
$P(ideal\ obs.)$	$(1{-}\alpha)P(S{=}0\|Z{=}corr.)P(Z{=}corr.)+P(S{=}0\|Z{=}incorr.)P(Z{=}incorr.)+$ $\int_{0^*<S\le 1}\xi P(S\|Z{=}corr.)\Big[(1{-}S)^n P\left(ZO{=}(corr.,corr.)\right)+P\left(ZO{=}(corr.,incorr.)\right)\Big]+P(S\|Z{=}incorr.)\Big[(1{-}S)^n P\left(ZO{=}(incorr.,corr.)\right)+P\left(ZO{=}(incorr.,incorr.)\right)\Big]dS$							

($*$ understanding that this integral specifically *excludes* any probability masses at $S{=}0$ of the conditional distributions of S given Z.)