NASA/TM–2016–219195

An Investigation of Proposed Techniques for Quantifying Confidence in Assurance Arguments

Patrick J. Graydon and C. Michael Holloway
Langley Research Center, Hampton, VA

May 2016

NASA/TM–2016–219195

An Investigation of Proposed Techniques for Quantifying Confidence in Assurance Arguments

Patrick J. Graydon and C. Michael Holloway
Langley Research Center, Hampton, VA

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

May 2016

Acknowledgments

We thank Lian Duan, Janusz Górski, and Sunil Nair for their assistance in understanding and replicating their work. We thank John McDermid, Andrew Rae, and the branch and directorate reviewers for their feedback on this work.

The use of trademarks or names of manufacturers in this report is for accurate reporting and does not constitute an official endorsement, either expressed or implied, of such products or manufacturers by the National Aeronautics and Space Administration.

Available from:

NASA STI Program / Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199
Fax: 757-864-6500

Abstract

The use of safety cases in certification raises the question of assurance argument sufficiency and the issue of confidence (or uncertainty) in the argument's claims. Some researchers propose to model confidence quantitatively and to calculate confidence in argument conclusions. We know of little evidence to suggest that any proposed technique would deliver trustworthy results when implemented by system safety practitioners. Proponents do not usually assess the efficacy of their techniques through controlled experiment or historical study. Instead, they present an illustrative example where the calculation delivers a plausible result. In this paper, we review current proposals, claims made about them, and evidence advanced in favor of them. We then show that proposed techniques can deliver implausible results in some cases. We conclude that quantitative confidence techniques require further validation before they should be recommended as part of the basis for deciding whether an assurance argument justifies fielding a critical system.

1 Introduction

The *safety case approach* has been used in some industries and regulatory domains for many years [1]. An organization using the approach takes ownership of the risks to be controlled by adopting an appropriate safety management system, performing a hazard assessment, selecting appropriate controls, and implementing these. The main difference between the safety case approach and other systems safety approaches is the use of a *safety case* to document hazards, controls, and the controls' adequacy [2]. A safety case combines safety evidence such as fault tree analysis results and test reports with an *assurance argument*, typically defined as "a reasoned and compelling argument . . . that a system, service or organisation will operate as intended for a defined application in a defined environment" [3]. A safety case might serve many purposes. For example, a safety case might communicate the system safety rationale to engineers who will later modify the system. Alternatively, a safety case might explain the safety rationale and evidence to an assessor who must decide whether the hazard controls are adequate. Such use raises a question: how should the assessor determine whether the argument and its evidence are sufficient?

The question of assurance argument sufficiency leads to the concepts of *confidence* and *uncertainty* in the argument's claims. Researchers have defined methods for reviewing assurance arguments [4, 5] and means of associating reasoning about confidence with the parts of the assurance argument they relate to [6, 7]. Other researchers propose adopting quantitative models of argument from disciplines such as philosophy to the problem of assessing confidence in assurance arguments. Some vendors sell tools to perform the necessary calculations [8]. But despite the importance of knowing how far confidence estimates should be trusted, little is known about whether proposed techniques for quantifying confidence produce trustworthy results [9]. And as others have observed, the seductive appearance of computational rigor might cause decision-makers to mistakenly place trust in “superficially plausible nonsense” [10]. A frank appraisal of the evidence for and against the efficacy of proposed quantitative confidence techniques will be of value to safety engineers, assessors, and regulators who must decide how to assess safety arguments and interpret assessments. In this paper, we survey and assess proposed techniques for quantifying confidence in assurance arguments. We identify the proposers’ claims and the support given for these, provide specific counterarguments, identify common flaws in the proposals, and assess the evidential basis for quantifying confidence.

2 Background

There is a substantial literature on safety cases, a much larger philosophical literature on argument, and a growing body of work on applying ideas from the latter to the former.

2.1 Safety Cases

In the 1970s, the United Kingdom (UK) Committee on Safety and Health at Work observed that prescribing specific risk reduction measures had not ensured safety in diverse workplaces for two reasons [11]. First, prescription encouraged compliance without thought, resulting in missed opportunities for risk reduction. Second, making law or regulation takes so much time that prescriptions were often out of date before or shortly after they took effect. (This is still true four decades later [12].) Accordingly, the UK introduced the safety case process to compel operators to conduct risk assessments, implement appropriate mitigations, adopt an appropriate safety management plan, commission independent audits to verify effective safety management, and revisit safety as circumstances, operations, and technology change [1, 11, 13]. In later decades, the safety case process was expanded to applications such as offshore oil and gas installations [13] and railway operations [14]. The safety case process is now used in the oil and gas sector in the UK and Australia, and a similar process is used in Norway [1]. Safety cases are used in the UK defense sector [15], in automotive applications [16], and with some medical devices in the United States [17].

Safety cases and their assurance arguments are thought to serve multiple purposes. For example, safety cases communicate safety design intent to those who will modify existing systems so that safety can be maintained during and after the change [15]. Safety cases also explain the safety rationale and evidence to an assessor—a customer, regulatory agency, or third party—who uses that information to decide whether a system is adequately safe [18–25]. Common definitions of assurance cases call for them to be compelling [3, 15]. But if an authority is to decide whether to accept a system based on whether its assurance argument is compelling, that authority would need to assess the confidence that the argument justifies and to use that assessment in a test of sufficiency. Researchers have proposed non-quantitative means of doing this [4]. But it is commonly held that, *ceteris paribus*, quantitative methods produce more trustworthy results than qualitative methods. Perhaps for this reason, some researchers have proposed techniques for quantifying confidence in assurance arguments [18–32].

2.2 The Philosophy Literature

The literature on assurance arguments draws mainly on the philosophy of *informal argumentation*, particularly the work of Toulmin [33, 34]. This literature focuses mainly on how to structure, represent, and critically examine arguments as humans make them [33–35]. While philosophers recognize some arguments as stronger than others, the informal argumentation literature does not generally treat argument strength as a property to be measured or calculated.

Philosophers in a different tradition have been defining logics in terms of probability calculi for over a century [36]. Many of these approaches, including Keynes's, are based on *Bayes' Theorem* [36, 37]. More recent work by Dempster and Shafer has defined the *Dempster–Shafer theory* of evidence [38]. Researchers have proposed applying Dempster–Shafer theory to multiple attribute decision analysis problems; this application is known as *Evidential Reasoning* [39].

We will not recount the history of these approaches or identify their differences. It suffices to observe that no consensus exists among philosophers that any of these is a trustworthy means of computing confidence in an argument so as to reliably determine whether it is sufficient to justify an action such as putting a safety-critical system into service. Probabilistic logics have long been the target of criticism [40] and continue to be updated [41]. While it *might* be possible for safety professionals to use probabilistic logic to produce trustworthy estimates of confidence in assurance claims, this cannot be inferred from what is known about these logics. Careful, direct scientific assessment of efficacy is needed.

3 Method

Researchers have been examining techniques for quantifying confidence in assurance arguments since the 1990s [18–32]. In this paper, we survey and assess these proposals.

3.1 Selection of Proposals

We identified twelve candidate proposals (in fifteen papers) using a combination of web and repository searches (e.g., Google, IEEE Xplore, and ACM Digital Library), notification services (e.g., Google Scholar Alerts on terms such as ‘safety case’), following chains of references, and targeted review of the writings of selected researchers. Some literature surveys systematically obtain a sample of papers on a subject. This is necessary when the proportion of selected studies reaching a given conclusion is used to assess the truth of that conclusion. Our literature survey is not systematic in this sense, but we are not drawing conclusions about a balance of conflicting evidence. Our primary aim was to include as many proposals as we could find. Our secondary aim is to arm readers with the information they need to understand the maturity of current proposals and critically assess future proposals.

3.2 Assessment of Proposals

To assess the maturity of techniques for quantifying confidence in assurance cases, we first reviewed each proposal to capture (1) what its authors hypothesize about the efficacy of that method and (2) what evidence they provide or cite to support that hypothesis. We then attempted to refute the hypothesis that each technique produces output that is a suitable basis for making release-to-service decisions by identifying a counterexample in which the technique’s output is untrustworthy.

In many of the selected papers, the authors propose a technique and apply it to a specimen argument fragment to yield a plausible assessment of confidence. To predict whether a technique’s output will be trustworthy in a proposed application, one needs to know whether the output is known to be correct in all or most similar applications. A single example cannot show this, but a single counterexample can cast doubt on it. Since the burden of proving that a technique is effective lies with its proposers, the absence of sufficient evidence undercuts conclusions of efficacy. Nevertheless, we also attempted to find a counterexample for each technique.

To ensure that we understood the authors’ proposals and applied them correctly, we began by attempting to reproduce their examples (if given). We then attempted to find a counterexample for which the proposed technique produces an implausible result. Where possible, our counterexamples are variants of the original examples. The techniques we survey require analysts to supply input such as the choice of aggregation rules, weighting and scaling factors, and likelihood data. Strict guidelines for these parameters are not supplied and might be impossible to furnish. By using similar argument structures, weights, and input data, we reduce the risk of misapplying the technique. For example, suppose a technique provides multiple rules for aggregating confidence from premises to assess confidence in a conclusion. Using the same rule the authors used for a given reasoning step eliminates the possibility of undermining the purported counterexample by claiming a different rule should have been used.

The existence of a counterexample demonstrates that a technique cannot be trusted to produce acceptable output in all cases and raises the effective the burden of proof for proponents. For example, a proponent might hypothesize that a technique, while imperfect, nevertheless produces better output on average than plausible alternatives. But such a hypothesis cannot be confirmed by worked examples; controlled experiments or historical studies are needed.

Our counterexamples also provide value beyond assessing the identified quantitative confidence techniques: they arm readers with patterns for critically examining other such proposals. When assessing a new quantitative confidence technique, we suggest that readers begin their evaluation by determining whether it produces plausible output for the extreme cases that we make use of.

4 Results

Table 1 lists the twelve selected proposals for quantifying confidence in assurance cases. In this section, we briefly summarize each, identify the hypotheses and evidence for efficacy put forward by their authors, and give our counterarguments. We present detailed descriptions of each proposed technique, the claims made about them, the evidence for these, and our counterargument in Appendices A–L.

4.1 The Proposed Techniques

Of the twelve techniques we identified, five are based on Bayesian Belief Networks (BBNs), six are based on Dempster–Shafer Theory, Jøsang's Opinion Triangle, or Evidential Reasoning, and one makes use of weighted averages. In this section, we briefly introduce each in turn.

4.1.1 Techniques Based on Bayesian Belief Networks

In techniques based on Bayesian Belief Networks, the analyst constructs a network of nodes using a BBN tool. Each node represents a variable and the absence of an arrow from one node to another indicates independence. The analyst supplies probability data for the leaf nodes and conditional probability tables or formulae for non-leaf nodes. The BBN tool then computes the probability of the non-leaf nodes, including the node representing the safety claim of interest.

Denney, Pai, and Habli. Denney, Pai, and Habli propose to compute uncertainty in a safety claim by constructing a BBN that roughly mirrors

Technique	Described in	Underlying theory	Focus of technique	Empirical assessment	Example reproduced	Counterexample(s) found
Ayoub et al. [18]	Section 4.1.2, Appendix A	Dempster–Shafer theory	Safety	✗	✓	Section 4.4.1, Section 4.4.2
Cyra and Górski [26, 27]	Section 4.1.2, Appendix B	Dempster–Shafer theory	Trust (safety and security)	Section 4.3.2	✓	Section 4.4.1, Section 4.4.2
Denney et al. [28]	Section 4.1.1, Appendix C	Bayesian Belief Networks	Safety	✗	✓	Section 4.4.1
Duan et al. [19]	Section 4.1.2, Appendix D	Jøsang's opinion triangle	Safety	✗	✓	Section 4.4.1
Guiochet et al. [29]	Section 4.1.2, Appendix E	Dempster–Shafer theory	Safety	✗	✗	✗
Guo [30]	Section 4.1.1, Appendix F	Bayesian Belief Networks	Safety	✗	✗	✗
Hobbs and Lloyd [20]	Section 4.1.1, Appendix G	Bayesian Belief Networks	Safety	✗	✓	✗
Nair et al. [21, 22, 31]	Section 4.1.2, Appendix H	Evidential Reasoning	Safety	Section 4.3.3	✓	Section 4.4.3
SERENE Partners [23]	Section 4.1.1, Appendix I	Bayesian Belief Networks	Safety	✗	✓	Section 4.4.3
Yamamoto [32]	Section 4.1.3, Appendix J	Weighted average	Safety and security	✗	✓	Section 4.4.1
Zeng et al. [24]	Section 4.1.2, Appendix K	Dempster–Shafer theory	Safety	✗	✓	Section 4.4.1
Zhao et al. [25]	Section 4.1.1, Appendix L	Bayesian Belief Networks	Safety	✗	✓	Section 4.4.2, Section 4.4.3

Table 1: *Selected techniques for quantifying confidence in assurance arguments.*

the structure of a safety argument [28]. Each leaf node represents a source of uncertainty in the safety argument. The leaf node's value encodes the analyst's confidence (derived from "both quantitative data ... [and] qualitative means") on a five-point scale from *Very Low* to *Very High*. The paper gives an example in which a node labeled Proof Correct represents confidence in a proof used to verify software. Non-leaf nodes are computed as the weighted average of the nodes from which they derive their value. The paper gives an example in which Computation Correct is the weighted average of Specification Correct and Proof Correct. See Appendix C for details.

Guo. Guo proposes representing safety arguments as BBNs [30]. The BBN replaces representations of the safety argument such as diagrams in Goal Structuring Notation (GSN) [3]. The paper gives an example in which "quality of safety function requirement" depends on "omission," "misunderstanding," "conflict," and the "quality of overall safety requirement." See Appendix F for details.

Hobbs and Lloyd. Hobbs and Lloyd propose representing assurance arguments as BBNs "where each leaf ... represents elementary evidence and the network's structure represents the argument" [20]. They refer to the SafEty and Risk Evaluation using bayesian NEts (SERENE) manual [23] for patterns but suggest that their *noisy-or* and *noisy-and* functions might model assurance case reasoning better than simple *or* and *and* functions. See Appendix G for details.

The SERENE Partners. The partners of the Safety and Risk Evaluation Using Bayesian Nets (SERENE) project propose using BBNs to model system safety and compute risk [23]. The SERENE method manual defines a number of *idioms* (patterns) that the analyst uses to model the system in question. See Appendix I for details.

Zhao, Zhang, Lu, and Zeng. Zhao, Zhang, Lu, and Zeng propose to calculate confidence using BBNs [25]. The analyst (1) interprets an existing assurance argument as a sequence of Toulmin arguments [34], (2) creates a BBN by instantiating a given pattern for each instance, (3) supplies the requisite probability data, and (4) uses a BBN tool to compute the resulting confidence in safety. See Appendix L for details.

4.1.2 Techniques Based on Dempster–Shafer Theory, Jøsang's Opinion Triangle, or Evidential Reasoning

Approaches based on Dempster–Shafer theory differ from Bayesian approaches in that they reason about both the strength of *belief* in opinions and the *plausibility* of those opinions. This adds a second dimension to the assessment of each claim in a safety case. Jøsang's opinion triangle,

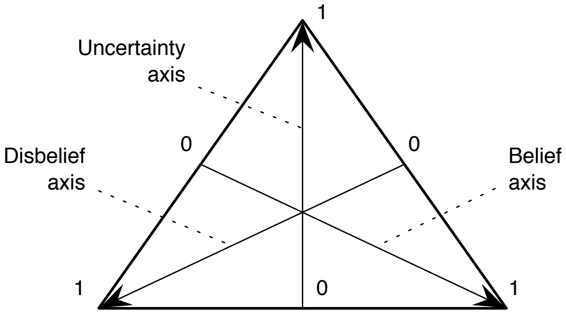


Figure 1: *Jøsang's opinion triangle redefines Dempster–Shafer theory's belief and plausibility measures in terms of belief, disbelief, and uncertainty [42].*

illustrated in Figure 1, reimagines the two dimensional space as three dimensions: *belief*, *disbelief*, and *uncertainty* [42]. As belief, disbelief, or uncertainty grows, the other two attributes must shrink. Evidential reasoning is an approach to using Dempster–Shafer theory in multi-attribute decision problems [39].

Ayoub, Chang, Sokolsky, and Lee. Ayoub, Chang, Sokolsky, and Lee propose to compute the sufficiency of an existing safety arguments using Dempster–Shafer theory [18]. The analyst first assesses the sufficiency and insufficiency of item of evidence cited by the safety case. The analyst then uses aggregation rules to compute the sufficiency and insufficiency of each claim in the safety argument, working from evidence toward the main safety claim. See Appendix A for details.

Cyra and Górski. Cyra and Górski propose to evaluate confidence in assurance cases using Dempster–Shafer theory. The analyst first creates an argument comprising a series of claims supported by *arguments*, *warrants*, and *premises*. Each premise is another claim, a *fact* supported by evidence, or an *assumption*. The analyst then assesses each assumption, fact, and warrant in terms of Jøsang's opinion triangle. The analyst finally identifies the type of each argument step, selects the appropriate formulae from those provided by the authors, and computes belief, disbelief, and uncertainty in each claim. See Appendix B for details.

Duan, Rayadurgam, Heimdahl, Sokolsky, and Lee. Duan, Rayadurgam, Heimdahl, Sokolsky, and Lee propose using the beta distribution and Jøsang's opinion triangle to model uncertainty in assurance case evidence [19]. The analyst first expresses an opinion about evidence cited by an existing assurance argument in terms of Jøsang's opinion triangle then uses the beta distribution and Jøsang's formulae to compute opinions in the argument's claims. See Appendix D for details.

Guiochet, Hoang, and Kaâniche. Guiochet, Hoang, and Kaâniche propose a technique for assessing confidence in an existing assurance case using Dempster–Shafer theory [29]. The technique models confidence in a conclusion as the strength of belief that the conclusion is true and defines uncertainty as the lack of belief that the conclusion is either true or false, but rules out disbelief (which presumably should not be included in the argument). The authors propose aggregation rules for various kinds of argument steps. See Appendix E for details.

Nair, Walkinshaw, Kelly, and de la Vara. Walkinshaw, Kelly, and de la Vara propose to use Evidential Reasoning (ER) to calculate belief in safety claims [21,22,31]. The proposed technique focuses on assessing the evidence cited by an existing safety argument. The analyst answers a series of questions specific to the type of evidence and a tool both (i) instantiates a GSN pattern depicting the reasons for having confidence in the claim supported by the evidence and (ii) calculates that confidence using Evidential Reasoning rules. See Appendix H for details.

Zeng, Lu, and Zhong. Zeng, Lu, and Zhong propose using Dempster–Shafer theory to assess the confidence in arguments recorded in GSN [3, 24]. The proposed technique defines confidence on a five-point Likert scale and uses an improved Dempster's rule to assess confidence in conclusions. See Appendix K for details.

4.1.3 Other Techniques

One of the papers we selected proposes an approach that does not use BBNs, Dempster–Shafer theory, Jøsang's opinion triangle, or Evidential Reasoning.

Yamamoto. Yamamoto proposes annotating GSN assurance arguments with attributes and using these to evaluate architectures "for quality claims, such as security and safety" [32]. In the proposed technique, analysts rate GSN solutions on a five-point Likert scale from *strongly unsatisfied* to *strongly satisfied*. The technique defines the assessment of a GSN goal as the weighted average of its supporting goals and solutions. See Appendix J for details.

4.2 Hypotheses

None of the selected papers formally defines a research hypothesis (e.g., using a sentence containing the word 'hypothesis' or notation such as H_0). But all of the papers begin with an introduction that describes and motivates the work. In this section, we characterize the hypotheses that we infer from the papers' narratives.

All selected papers assert the need for a reviewer to determine how much confidence in a safety claim an assurance argument should justify.

The papers for seven of the twelve techniques describe this assessment as the basis for determining whether a system is safe enough to field. While the papers for six techniques present conclusions that imply limited maturity, the papers for the remaining six make stronger claims. The papers for five techniques hypothesize that the techniques will help analysts make more accurate assessments than they could with unaided reason. And the papers presenting five techniques hypothesize that the technique better communicates the safety argument to stakeholders than alternatives.

4.2.1 The Purposes of Confidence Assessment

The selected papers related to eight techniques present them in a manner that suggests that those techniques are to be used to assess whether systems are safe or secure enough to deploy [18–24, 29–31] (see Appendices A, D–I, and K):

- The paper about one technique motivates it by asserting that "the objective of safety case development . . . is to facilitate mutual acceptance [among system stakeholders] of [a] subjective [safety claim]" [18]. Quantitative confidence assessment aids this mutual acceptance by "answering a question about the overall sufficiency of the argument."

- A paper about another technique notes that "certain safety critical systems must be demonstrated to be safe and certified or approved by some regulatory body before they are allowed to be taken into operation or sold to the general public" and that this raises the question of confidence in the systems' assurance cases [19].

- The paper about a third technique begins by noting that "when implementing a safety- or mission-critical application, it is necessary to convince the authors, the management of the development company, potential customers and auditors that the system meets its safety, availability and reliability requirements" and hypothesizes that BBNs are a means of doing this [20].

- The paper about a fourth technique begins with the observation that "safety cases are used in several critical industrial sectors to justify safety of installations and operations" (to unspecified stakeholders) and then notes that "an important and growing issue is to understand how much confidence one could have in the [safety] claim" [29].

- The paper about a fifth technique asserts that "BBN provides a reasonable frame to quantitatively evaluate the fulfilling quality of the requirements in safety standards" and "obtain the confidence of a system, e.g., to provide a quantitative figure about systematic failure rate" [30].

- A paper about a sixth technique begins by noting that "goal-based system safety standards such as [UK Defence Standard] 00-56 ... often require the construction and provision of a safety case" and notes that as a result "the assessor needs to establish confidence that the safety case adequately addresses the identification and mitigation of hazards" [15, 21]. Two more papers about the same technique note that "it is often left to the human assessor to decide whether or not the presented evidence is sufficient to support the safety claims made in the case" [22, 31].
- The manual about a seventh technique notes that "the justification of safety, sometimes called a safety case, may be submitted to safety regulators for approval" [23].
- A paper about an eighth technique notes in its introduction that "the acceptance of a safety case requires the assessors to be confident that the safety case meets [its] requirements" [24].

The paper presenting an additional technique does not state how confidence assessment is to be used save to note that an "overarching motivation for this work is, eventually, to integrate it into a quantitative framework for risk analysis" [28] (see Appendix C). The paper presenting another technique notes that "assurance cases support consensus building process" but defines the aim of the technique as "help[ing] engineers develop safety critical ... system architectures quickly" [32] (see Appendix J). The papers presenting the remaining two techniques assert the need to assess confidence in arguments without stating how that assessment will be used [25, 26] (see Appendices B and L):

- A paper presenting one technique notes the "necessity of expert assessment of the compelling power of [argument] structures" but does not explicitly connect this to release-to-service decisions [26].
- The paper presenting a third technique begins with the observation that "instead of assessing manufacturer compliance with process-based regulations and standards, recently the industry areas have paid much attention to the assurance cases which focused on demonstrating the dependability of product-specific system" but does not elaborate on who this is demonstrated to or why [25].

4.2.2 The Maturity of the Proposed Techniques

The selected papers about six techniques present conclusions in a manner consistent with a technique that is still being developed or assessed [18, 19, 22, 25, 28, 29] (see Appendices A, C–E, H, and L):

- The paper presenting one proposed technique notes that the underlying theory's assumption of evidence independence is often violated in safety arguments and proposes future work to account for this dependence [18].

- The paper presenting another technique describes assessment of the proposed technique as a “preliminary investigation” [28].
- The paper presenting a third technique notes that it discusses only two probabilistic operators and suggests examining others in future work [19].
- The paper presenting a fourth technique “focus[es on] the feasibility of a quantitative estimation of confidence” without addressing the accuracy of that estimation [29].
- A paper about a fifth technique indicates that in the future the authors will “further validate the completeness of the confidence argument pattern [they present] with more checklists used in practice, ... seek to improve the tool support, eliminate fallacies ... and mitigate fatigue when answering the questionnaire” used to assess confidence in an evidence citation [22].
- A paper about a sixth technique cautiously concludes that the proposal is “a potentially helpful way towards the measurement of confidence in assurance cases” [25].

The selected papers about three other techniques present conclusions that suggest that the proposed techniques are more mature [24, 27, 32] (see Appendices B, J, and K):

- A paper presenting one technique concludes that it “has already been fully implemented in the [authors’] tool” and “has been subjected to experimental validation” with “further experiments under preparation. Furthermore, the method has been applied for appraisal of arguments for patient safety and privacy, and for fulfilment of security requirements in two [European Union–funded] projects. It is also going to be used in a new project utilizing argument structures to demonstrate conformity with standards and regulations” [27].
- The paper presenting a second technique refers to the technique as a proposal yet concludes that “discussions based on the case study showed the effectiveness and appropriateness of the proposed method to resolve security and safety issues” and describes future work aimed at assessing “productivity” but not efficacy [32].
- The paper presenting a third technique also describes it as proposed yet asserts that the underlying theory “is fit for processing the subjective judgment and synthesizing the uncertain knowledge” and concludes that “it is proved that [Dempster–Shafer] theory has advantages in evaluating confidence in safety case which has some uncertainty. The usage of [Dempster–Shafer] theory reduced the effect of the uncertainty, improved the precision and the validity of the evaluation, and reduced the blindness and the subjectivity of evaluation of confidence in safety case” [24].

The selected papers about the remaining three techniques present them in a manner consistent with technology mature enough for practitioners to use [20,23,30] (see Appendices F, G, and I):

- The paper about one technique describes BBNs in a manner that gives the impression that they are a mature technology and asserts that an example application shows that it is appropriate to apply BBNs safety cases: "from this example, we show that BBN provides a mathematically sound approach for modelling and manipulating uncertainty that is inherent in safety assessment. Specifically, uncertainty can be expressed by conditional probability, and the Bayes theory underlying BBNs therefore offers them the capability to manage uncertainty" [30].
- The paper about a second technique presents BBN-based assurance arguments as a technology being used in practice, reports the authors' experience, and concludes that "applying a BBN to represent an assurance case, when backed by a suitable computation tool, appears to be a flexible and powerful technique" [20].
- The manual for a third technique presents detailed instructions for using the technique and describes expected benefits without mentioning what the evidence of efficacy is or what further research remains to be done [23].

4.2.3 Helping Analysts Make More Accurate Assessments

The selected papers about five techniques hypothesize that those techniques will help analysts to make more accurate assessments than they could with unaided reason [18,21,23,26,28] (see Appendices A–C, H, and I):

- A paper presenting one technique asserts that "the research in experimental psychology shows that human minds do not deal properly with complex inference based on uncertain sources of knowledge," citing a computer scientist's paper on the subject [26,43].
- The paper presenting a different technique cites the former paper when making the same assertion [18].
- The paper proposing a third technique observes that "subjectivity inherent in the structure of the argument and its supporting evidence ... pose[s] a key challenge to the measurement and quantification of confidence in the overall safety case" [28]. The same paper concludes that "linking qualitative safety arguments to quantitative arguments about uncertainty and confidence ... ensur[es] rigor in measuring confidence via probabilistic reasoning using [BBNs]."

- A paper presenting a fourth technique hypothesizes that the technique "will help safety case assessment to be more systematic and consistent" than using another technique for reviewing arguments expressed in GSN [21].
- The manual presenting a fifth technique asserts that it provides "a basis for empirical validation of the beliefs of safety experts," thus presumably leading to more accurate assessments than they could obtain with unaided reason [23].

The paper presenting one of the above techniques also hypothesizes that the technique limits the impact of the analysts' bias on their assessments [18] (see Appendix A). It suggests that asking analysts' opinions about both the sufficiency and insufficiency of evidence will counteract confirmation bias. The paper concludes, "preliminary experience of applying the proposed method has revealed that the assessing mechanism yields the expected benefits in guiding the safety argument reviewer and helping him/her to reduce the effect of the confirmation bias mindset."

4.2.4 Communicating the Safety Argument

The selected papers about five techniques hypothesize that they result in better communication of the safety argument to stakeholders than an alternative [19–21, 23, 30, 31] (see Appendices D and F–I):

- The paper presenting one technique hypothesizes that the beta distribution, mapped to Jøsang's opinion triangle, better represents human opinion than alternatives such as the truncated normal distribution [19].
- The paper presenting another technique hypothesizes that using a BBN to represent an assurance argument provides "improved transparency" in relation to an unspecified alternative [30].
- The paper presenting a third technique motivates a discussion of BBN-based arguments by noting that organizing and presenting a safety case "in an unstructured, natural language presents difficulties both to the authors of the case, who have to struggle to maintain coherence as the case evolves, and to the auditors reviewing it, who have to extract the threads of the argument" [20].
- A paper presenting a fourth technique asserts that it "explicitly captures any uncertainty in the [assessor's] judgement" [21]. A further paper on the same technique asserts that the technique "enables these judgements to be presented within the context of an overall argument of confidence" [31].
- A manual presenting a fifth technique asserts that the technique provides "improved communication of safety" and a "greater fo-

> cus on the properties which lead to safety" than an unspecified alternative [23].

4.3 Evidence of Fitness for Use in Release-to-Service Decisions

As described in Section 4.2, all of the techniques described in Section 4.1 and detailed in Appendices A–L assume the need to quantify confidence in assurance arguments. While some papers describe quantified confidence assessment as part of a certification process, others do not. But that purpose is the focus of this paper. Accordingly, this section identifies the lines of argument that best show that a quantified confidence technique is fit for use as the basis of release-to-service decisions. The selected papers present three such lines of argument: (1) appeal to properties of the underlying theory, (2) the Cyra and Górski experiment, and (3) the Nair et al. survey.

4.3.1 Properties of the Underlying Theory

Several of the selected papers appeal to the properties of the underlying theory. In some cases, this is simply an appeal to the theory's quantitative nature or the breath of its acceptance. For example, one paper concludes that using BBNs "ensures rigor in measuring confidence via probabilistic reasoning" [28]. Another reasons that because "BBNs are based on Bayes' theory and have a sound mathematical background," they "provide a quantitative figure for evaluating safety" [30]. A paper proposing one technique explicitly identified relevant properties of the underlying theory [22]. Citing Yang and Xu, that paper reports that the Evidential Reasoning approach underpinning the proposed technique

> obeys certain desirable axioms that ensure the following:
>
> 1. If [no premise of] y is assessed at [confidence] grade H_n, then $\beta_{n,y}$ [(belief that y has confidence H_n)] $= 0$.
> 2. If all [premises] are assessed to a grade H_n then $\beta_{n,y} = 1$.
> 3. If all [premises] are completely assessed to a subset of evaluation grades then y should be completely assessed to the same subset of grades.
> 4. If, for a [premise] z, $\sum_{i=0}^{n} \beta_{i,z} < 1$, then the same holds for y: $\sum_{i=0}^{n} \beta_{i,y} < 1$ [22].

These properties, while desirable, are not sufficient to show that the proposed technique produces trustworthy confidence assessments. Rules that would produce different confidence figures also satisfy these criteria. For example, one might define the combination of assessments of premises as either the categorical weaker or categorical mean of the two assessments. (See Section H.3.)

If an appeal to the underlying theory's properties is to show that a technique produces trustworthy confidence assessments, the maker of the appeal must show that the theory's properties entail the accuracy of the technique's results. None of the surveyed papers shows this.

4.3.2 The Cyra and Górski Experiment

The papers presenting one selected technique present the results of "an experimental calibration and validation" of that technique [27] (see Appendix B for details). The technique uses functions that map opinions expressed on a five-point linguistic scale (e.g., "with very high confidence") to and from the numeric forms used in computation. The authors used an experiment to both derive these parameters and assess the technique. The participants, thirty-one Master's students, were divided into three groups that each focused on a distinct kind of argument step. The roughly ten subjects in each group assessed the strength of warrants, chose weights for premises, and, given assessments of premises, assessed strength in the conclusion. The experimenters assessed the *consistency of assessments* by measuring the root-mean-squared "of the difference between the first and repeated assessment (by the same participant) of the same conclusion with the same assessments assigned to the premises." The experimenters also assessed the *accuracy of assessments* by measuring the root-mean-square "of the difference between a participant's assessment and the result of application of the [proposed technique]."

The experimental results show that individual students assess arguments somewhat inconsistently: the root-mean-squared value of consistency ranged between 0.62 and 1.03 of a linguistic scale category depending on the type of argument step. Individual assessments varied from the proposed technique's predictions by about as much: the root-mean-squared value of accuracy ranged between 0.66 and 1.10 of a category depending on the type of argument step.

These results are encouraging. But there are threats to validity that limit the experiment's strength as evidence of efficacy. Two stand out: the problem of using the same data to both define and assess the technique, and the representativeness of the sample. The danger in using the same data to both define and assess a technique is that the technique might be effective only for the specimen problem from which it derives. This problem can be mitigated using an approach such as *n-fold cross validation*. But the papers do not mention whether or how the experimenters addressed this problem. And it is not clear that the reasoning a small number of students used to assess five example argument steps is normative for all assurance cases. Students and seasoned security professionals might assess arguments differently. The five selected arguments of each type might not be representative of all arguments. And it is possible that people generally assess the specimen arguments incorrectly.

4.3.3 The Nair et al. Survey

The papers presenting one selected technique present the results of a survey of twenty-one safety experts [22] (see Appendix H for details). The subjects viewed a presentation and then responded to survey questions. When asked whether "use of the approach will lead to more accurate safety evidence assessments," four strongly agreed, nine agreed, and eight neither agreed nor disagreed.

It is not clear that the reported survey provides much support for the hypothesis that the technique is a trustworthy basis for making release-to-service decisions. This is in part due to weaknesses that affect survey research generally: surveys are affected by *response biases*, including those related to *demand characteristics*. For example, survey respondents sometimes feel compelled to be 'good subjects' by giving responses they think support the researchers' hypothesis [44]. Researchers can limit response bias through measures such as deliberately hiding the hypothesis from subjects, but the papers describe no such controls.

The other main threat to validity is that respondents are not known to be capable of accurately assessing efficacy. Efficacy might depend on subtle matters of degree such as analysts' skill in choosing weights and making judgments. If experimental and observational assessments of efficacy and its contributory factors were available to respondents, their opinions might synthesize this evidence. Without such data, it is not clear that participants' opinions on efficacy are a measure of efficacy.

4.4 Counterargument

As reported in Table 1, we identified counterexamples for nine of the twelve selected techniques. In two of the remaining cases, the papers did not present a worked example we could replicate. (See Appendices E and F.) In the third, we replicated calculations but found the original example unsuitable as a basis for plausible counterexamples. (See Appendix G.) As described in Section 3.2, we derived counterexamples from original examples so that our choices of weights and parameters are consistent with the authors' intent. Without an original example to follow, we could not create counterexamples without risking these being regarded as misapplications of the proposed technique.

Our counterexamples came in three main forms: *masking missing evidence or counterevidence*, *sensitivity to the arbitrary scope of hazards*, and *technique-specific counterexamples.* In this section, we discuss all three forms and then report other issues we identified when assessing the selected techniques.

4.4.1 Masking Missing Evidence or Counterevidence

When verification and validation techniques produce a result other than the desired (safe) result, it is accepted safety practice to investigate the

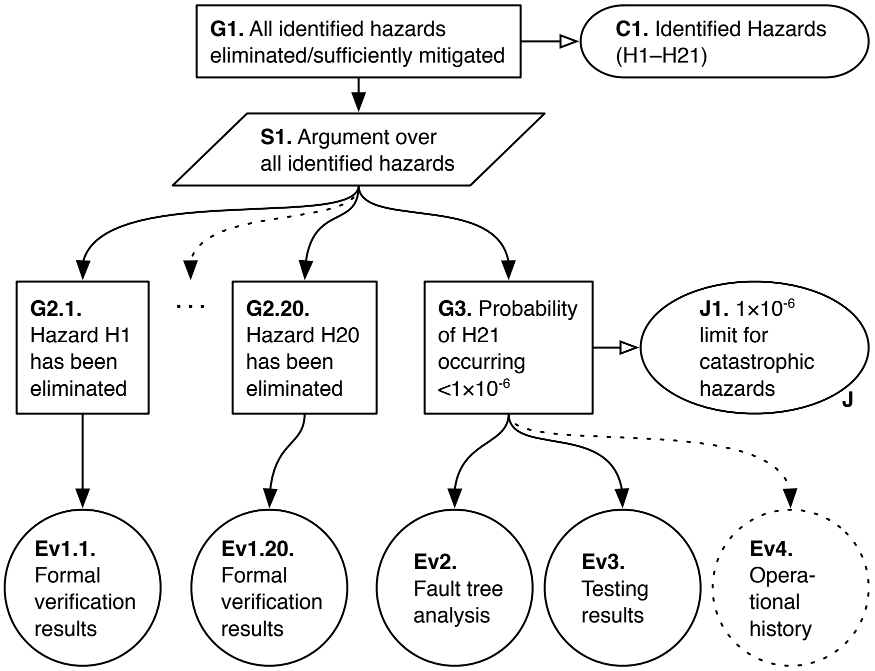

G1. All identified hazards eliminated/sufficiently mitigated
C1. Identified Hazards (H1–H21)
S1. Argument over all identified hazards
G2.1. Hazard H1 has been eliminated
. . .
G2.20. Hazard H20 has been eliminated
G3. Probability of H21 occurring <1×10-6
J1. 1×10-6 limit for catastrophic hazards
J
Ev1.1. Formal verification results
Ev1.20. Formal verification results
Ev2. Fault tree analysis
Ev3. Testing results
Ev4. Opera-tional history

Figure 2: *Our Many Hazards (with G2.2–G2.19 and Ev1.2–Ev.19 but not Ev4), Undermined Evidence 1 (with G2.2–G2.19 and Ev1.2–Ev.19 but not Ev4), Undermined Evidence 2 (without G2.2–G2.19, Ev1.2–Ev.19 or Ev4) and Counter Evidence (with G2.2–G2.19, Ev1.2–Ev.19, and Ev4) variants of the example safety argument shown in Figure A1 (itself from [18]).*

cause of that outcome. For example, suppose that a test fails. This might happen for reasons other than a dangerous flaw. For example, testers might misinterpret a requirement or perform a test procedure incorrectly. Even if the test has revealed a flaw, it might be prudent to live with the flaw because correcting it could introduce more dangerous defects. But the cause of any undesired test, review, or analysis result should be investigated before making a decision to overlook it.

We identified counterexamples for six of the twelve selected techniques that demonstrate that evidence for one claim is allowed to mask missing evidence or counterevidence for the same or different claims. (See Table 1 and Appendices A–D, J, and K.) For example, consider the counterexample we identified for one technique based on Dempster–Shafer theory [18] (see Appendix A). Figure 2 depicts the argument on which we base three variants of the original example: *Many Hazards*, *Undermined Evidence 1*, and *Counterevidence*. (We will define a fourth case, *Undermined Evidence 2*, in Section 4.4.2.) These arguments differ from the original in two ways: (a) they show twenty-one hazards rather than two, and (b) one has additional operational history evidence. Table 2 gives our assessment of the sufficiency, insufficiency, and uncertainty of the evidence and the resulting computed assessment of the main safety

Element(s)	Variant	Suff.	Insuff.	Uncert.
Ev1.1	All	0.800	0.100	0.100
Ev1.2–19	All but Undermined Evidence 2	0.800	0.100	0.100
Ev1.20	Many Hazards, Counterevidence	0.800	0.100	0.100
	Undermined Evidence 1 and 2	0.050	0.900	0.050
Ev2	All	0.700	0.100	0.200
Ev3	All	0.500	0.200	0.300
Ev4	Counterevidence	0.050	0.900	0.050
G1	Many Hazards	0.801	0.101	0.099
	Undermined Evidence 1	0.765	0.139	0.096
	Undermined Evidence 2	0.079	0.053	0.868
	Counterevidence	0.777	0.127	0.096

Table 2: *The sufficiency, insufficiency, and uncertainty of the evidence and main safety claim in the example arguments shown in Figure 2.*

claim, G1. In all three cases, the sufficiency of the safety claim is much higher than uncertainty in the safety claim. As a result, the proposed technique assesses all three counterexample arguments as acceptable.

The danger of a technique producing such results is that analysts might use those results to justify forgoing investigation of the disconfirming evidence. This is not necessarily a fatal flaw. But this possibility must be kept in mind when specifying how quantified confidence will be used in the safety engineering process. In the case of these techniques, it is insufficient to rely solely on calculated confidence in safety to judge whether there are problems in the safety case.

4.4.2 Sensitivity to the Arbitrary Scope of Hazards

The number of significant hazards identified for a system depends on the analysts' choice of scope for each: one might increase the number of hazards by dividing one state into two distinct substates or decrease their number by combining two states into one. This change in scope need not change how the hazards are managed or what evidence of their management engineers produce. Simply the changing scope of hazards without changing how they are managed or the evidence of their management should not substantially affect confidence in system safety.

We identified counterexamples for three of the twelve selected techniques that show that those techniques produce substantially different assessments of confidence depending on the arbitrary choice of hazard scope. (See Table 1 and Appendices A, B, and J.) For example, consider a counterexample for the same technique discussed in Section 4.4.1. In addition to the Undermined Evidence 1 case shown in Figure 2 and Table 2, consider a new variant of the example, Undermined Evidence 2. The Undermined Evidence 1 and 2 examples are identical except that the system states identified as hazards H1–H19 in the former are identified as the single hazard H1 in the latter. The systems behave identically and we have identical evidence about their behavior. Yet the calculated confidence in safety differs dramatically: the argument is acceptable in the former case and unacceptable in the latter. This arbitrary difference in hazard scope should not so dramatically change our confidence in system safety.

In all of the examples in the selected papers in which the authors supplied weights for an argument over hazards, the example used equal weights. This would make sense if the mitigation of each hazard had an equal effect on system safety, but that is rarely the case. Some hazards might lead to more dire consequences than others, and accidents are more likely to result when in some hazardous states than in others. We might attribute the effect these counterexamples illustrate to the authors' choice to use equal weights in their examples. Yet the papers say nothing about choosing weights except through the examples they provide. And following those examples leads to implausible results.

4.4.3 Technique-Specific Counterexamples

In producing counterarguments for three of the twelve selected papers, we produced counterexamples that are unique to the techniques in question. (See Table 1 and Appendices H, I, and L.)

Specific Counterexample 1. One of these techniques includes four questions to be used to assess the trustworthiness of a hazard log [22]:

Q1. "If independence is required, is the person doing the verification different than the one responsible for developing the hazard log?

Q2. "Is the manager to whom the team reports identified so that it can be confirmed that the requirements on independence are met?

Q3. "Are there records of attendance/participation in hazard identification workshops/exercises of the personnel that include the name, organisation and role?

Q4. "Is there information on the competency of the personnel?"

The analyst responds to the questions using a five-point Likert scale and uses the proposed technique to calculate confidence in the hazard log. Our counterexample supposes (a) that two safety engineers are known, by long history, to be hopelessly incompetent to practice, (b) that they work independently (and that a manager has documented this) and (c) that there is record of them attending training. The answers to questions Q1–Q4 above might be *absolutely, with very high confidence.* (The engineers'

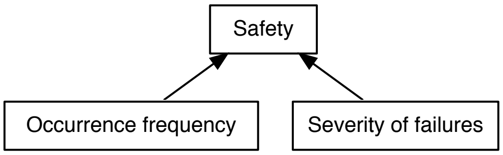


Figure 3: *A BBN from [23] instantiating the definition/synthesis idiom for safety. For consistency with other figures in this paper, we represent BBN nodes using rectangles.*

Variant	**Failure modes** ($<$ **Frequency, Severity** $>$)	**Risk**
Good	$\langle 10^{-9}, 10^{9}\rangle$, $\langle 10^{-9}, 10^{5}\rangle$, $\langle 10^{-7}, 10^{3}\rangle$, $\langle 10^{-3}, 10^{1}\rangle$	1×10^{1}
Bad	$\langle 10^{-9}, 10^{1}\rangle$, $\langle 10^{-9}, 10^{3}\rangle$, $\langle 10^{-7}, 10^{5}\rangle$, $\langle 10^{-3}, 10^{9}\rangle$	1×10^{6}

Table 3: *The failure modes and total risk of system variants to be modeled using the argument in Figure 3. Risk is computed using a typical expected value formulation, i.e.* $\text{Risk} = \sum_{i:\{\textit{Failure modes}\}} \text{Frequency}_i \times \text{Severity}_i$.

history of incompetence is "information on the competency of the personnel" even if it shows incompetence.) Using the proposed technique, one might conclude with very high confidence that the trustworthiness of the hazard log that one of the incompetent engineers compiles and the other verifies is very high. This is not plausible.

Specific Counterexample 2. The second of these techniques define the BBN pattern shown in Figure 3 [23]. Our counterexample comprises two variants of a system, *Good* and *Bad*, defined as shown in Table 3. Using continuous interval nodes, we instantiate the pattern for our systems as shown in Table 4. Even though we defined Good and Bad so that they produce maximally distinct levels of risk, the technique's pattern models them identically. This is not plausible.

It might be argued that we should have used a different equation for the Safety node, that we have inappropriately combined information about multiple failure modes into the Occurrence frequency and Severity of failures nodes, or that the manual used the word 'safety' where it meant 'risk.' But this counterexample survives such objections. The problem that the counterexample reveals is the result of first reasoning separately about the frequency and severity of system failure and then reasoning about the resulting effect on safety or risk.

Specific Counterexample 3. The third of these techniques includes a generic BBN pattern for assessing confidence in one argument step [25]. The paper presenting this technique gives an example in which this pattern is instantiated for an argument step from the premises "hazard A, B have been eliminated" to the conclusion "system S is safe." Our counterexample comprises two variants of the original example, namely

Node	**Distribution**				
Occurrence frequency	$[0, 10^{-8}]$	$, 10^{-6}]$	$, 10^{-4}]$	$, 10^{-2}]$	$, 1]$
	50%	25%	0%	25%	0%
Severity of failures	$[0, 10^{2}]$	$, 10^{4}]$	$, 10^{6}]$	$, 10^{8}]$	$, \infty)$
	25%	25%	25%	0%	25%
Safety	$[0, 10^{-8}]$	$, 10^{-6}]$	$, 10^{-4}]$	$, 10^{-2}]$	$, \infty)$
	0%	0%	25%	5%	70%

Table 4: *The node probabilities for the BBN shown in Figure 3 and the system variants defined in Table I1. We calculated Safety using the formula Safety* $=$ $(\textit{Occurrence frequency} \times \textit{Severity of failures})^{-1}$*. The probabilities for the Good and Bad variants are identical.*

Case	**Present**	**NotPresent**
From paper	89.460%	10.540%
Optimistic	89.996%	10.004%
Pessimistic	86.404%	13.596%

Table 5: *Calculated value of confidence in the claim "system S is safe" for an original example [25] and our Optimistic and Pessimistic variants.*

Optimistic and *Pessimistic*, that represent extreme assessments of confidence in the completeness of hazard identification. In both examples, we assess all factors as in the original example except confidence in the completeness of the hazard assessment. Where the original example uses 85% confidence in completeness, our Optimistic variant uses 99.9%, and our Pessimistic variant uses 0.1%. Table 5 gives the resulting confidence in system safety. It is not plausible that extreme changes in confidence in hazard analysis would produce as small a change in confidence in safety as the difference between 90% and 86% indicates. It is also implausible that anyone who completely distrusted a hazard analysis would have 86% confidence that the analyzed system is safe.

4.4.4 Other Issues

Five other issues each affect several proposals: (1) the unavailability of trustworthy source data, (2) scalability, (3) unspecified decision procedures, (4) improper treatment of GSN context, and (5) out-of-context assessment of GSN solutions.

The Unavailability of Trustworthy Source Data. We are not the first to note the lack of trustworthy sources of data on how well each kind of evidence supports each kind of claim. Surveys have established baseline figures for the reliability of components used in certain industries [45], but analogous data is not generally available for confidence in safety techniques. Nor are we the first to observe that trustworthy figures for the strength of inference rules are not generally available. Many of the selected papers call for the use of expert judgment to provide confidence figures and weights. But none of the papers provides or cites substantial evidence to show that the resulting figures are sufficiently trustworthy. Since poor input data could result in incorrect assessments of confidence, trustworthy figures are a prerequisite for confidence quantification. But no trustworthy source of the required data has been identified.

Scalability. It is not clear that all of the techniques could be applied practicably to a full-sized argument. For example, one technique requires the analyst to instantiate twelve BBN nodes, supply nine input probabilities, and populate one conditional probability table for each reasoning step in the assurance argument [25] (see Appendix L). Another technique requires the analyst to instantiate a confidence argument pattern comprising at least fifty-one GSN elements once per solution in the main argument [22, 31] (see Appendix H). Depending on the system and the complexity of the safety concept and safety evidence, an argument might have thousands of evidence citations and reasoning steps. It is not clear that techniques requiring a substantial effort for each will be feasible in practice. The selected papers present no empirical evidence of scalability.

Unspecified Decision Procedures. The motivations given in several selected papers describe using assurance arguments as part of system certification. All describe a technique for quantifying confidence in such assurance arguments. But few describe precisely how the quantified confidence should be used to make a decision about whether the system is sufficiently safe to justify putting it into service. (Ayoub et al. mention a comparison between the degree of belief and uncertainty in argument's main claim as described in Appendix A, but do not clearly state a decision procedure.) Without a clearly-specified decision procedure, the meaning of a technique's output is unclear. For example, suppose that a technique is intended solely to produce a confidence figure and that the use of that figure to make decisions is left as a separate matter. Suppose also that proponents give an example in which the technique computes 80% confidence (or 'high' on a scale that includes 'very high,' 'certain,' etc.). Without knowing the decision procedure, readers do not know whether this figure represents too little confidence to justify fielding the system, just barely enough confidence, or abundant confidence. And without knowing that, users of the technique don't know how to interpret the results and other scientists don't know how to assess the technique's efficacy.

Improper Treatment of GSN Context. The Goal Structuring Notation includes *context* elements [3]. Of the twelve selected techniques, two specify how GSN context elements affect confidence [28, 29] (see

Appendices C and E). The example of the first of these argues that "assurance deficits [are] acceptable" because the "argument [is] sufficient" and the "context [is] appropriate" [28]. The precise semantics of GSN context are unclear [46]. But suppose both that inappropriate context causes assurance deficits and that the evidence is from an entirely different system so that context is completely incompatible. Because the premises are equally weighted and the conclusion is one of three premises for the conclusion "claim [should be] accepted," the analysis might show that the claim should be accepted despite the argument's irrelevance.

The other technique specifies that context asserted at a goal should be factored into the confidence computation as if it was a premise [29]. This is inconsistent with normative guidance on GSN [46]. Moreover, this raises the question of how analysts should evaluate confidence in the truth of the example context statements given in the GSN standard such as "Operating Role and Context," "Control System Definition," and "SIL Guidelines and Processes" [3].

Out-of-Context Assessment of GSN Solutions. The Goal Structuring Notation also includes *solution* elements [3]. In three of the twelve selected techniques analysts assess confidence in GSN solution elements out of context [18,19,32] (see Appendices A, D, and J). For example, in one, the analyst "makes his/her assessments on the sufficiency and insufficiency of evidence nodes" [18]. A goal supported by a single solution then inherits the assessment of that solution. But a thing is evidence only by virtue of having been cited in support of a claim [47]. Moreover, an item of evidence might support one claim better than another. Opinions about the strength of evidence make sense only when applied to evidence–claim pairs as is done in other selected techniques [22].

5 Other Potential Arguments for Quantification

In Section 4.3, we identified evidence from the selected papers that supports the hypothesis that quantified confidence is a fit basis for release-to-service decisions. In Section 4.4, we present our argument against that hypothesis. A proponent of quantifying confidence might respond to this criticism by challenging the criticism or weakening the hypothesis. In this section, we examine options for doing so.

5.1 Overlooking the Counterexamples

A proponent of quantifying confidence might claim that a given model produces results that are *generally* accurate even if it produces implausible results in some 'corner' cases. This is plausible: a technique *might* be imperfect yet sufficiently accurate in an acceptably large proportion of cases that its use should be recommended over alternatives. But showing this would require empirical studies of how accurately typical analysts assess typical arguments using the proposed techniques.

One possible response to the counterexamples we present in Section 4.4 and Appendices A–L is to declare our choices of assessments, weights, or functions implausible. But doing so risks committing the *No True Scotsman* fallacy [48]. Our counterexamples reuse assessments, weights, and function choices from the original examples so as to ensure the plausibility of our selections.

5.2 There Is No Alternative

A proponent might propose that the need to assess confidence justifies using one of the available quantified confidence techniques even though its efficacy is not firmly established. But a quantitative technique that is no more effective than qualitative techniques might be riskier: some readers might come away with a greater impression of the trustworthiness of the quantitative analysis than they would a qualitative analysis about which equally little was known. To use quantities of unknown quality is to risk committing the fallacy of *overprecision* [49]. That is, the appearance of precision given by quantities might lead readers to put undue trust in confidence assessments.

5.3 Quantification is Systematic

A proponent of quantifying confidence might hypothesize that because systematic methods are more likely to give reproducible results than ad hoc approaches, quantitative confidence analyses should be preferred to qualitative analyses. A systematic method *might* yield more trustworthy results than an unsystematic method. But even if qualitative analyses were completely unsystematic—and they aren't [5]—being systematic does not guarantee producing a trustworthy result.

6 Possible Research Directions

While the proposed quantified confidence techniques are not known to be effective and might have flaws, they might nonetheless have value. To assess this value, researchers might consider several lines of inquiry:

1. *Demonstrating repeatability.* If representative subjects given an argument and told to apply a given confidence assessment technique to it do not achieve similar results, the technique is not measuring confidence. (The inverse is not true: repeatable assessments of confidence need not be accurate measures of confidence. The technique of always reporting medium confidence is repeatable but does not meaningfully measure confidence.)

2. *Demonstrating insensitivity to expected variance.* Subjects given details of a system and told to construct a safety argument will construct different arguments. If applying a confidence assessment technique to different but acceptable arguments for the same system yields substantially different results, the technique might be too sensitive to expected, insignificant variance to produce trustworthy confidence figures.

3. *Demonstrating predictive power.* Accidents and incidents are rare and safety arguments are generally confidential. But if researchers compile a substantial historical record, they could study it. If assessed confidence in safety is not inversely correlated with accidents and incidents, the technique is not measuring confidence.

7 Other Work

Following the method described in Section 3, we selected and reviewed fifteen papers describing twelve techniques for quantifying confidence in assurance arguments. But there are other works that discuss quantified confidence and safety arguments.

7.1 Littlewood and Wright

Littlewood and Wright model an argument using Bayesian Belief Networks to explore the gain in confidence resulting from the addition of an additional type of evidence [50]. While the paper uses a BBN to model confidence, it does not propose this as a general technique with which safety analysts should assess assurance arguments.

7.2 Wu and Kelly

Wu and Kelly propose a procedure for using Bayesian Belief Networks to analyze the architecture of software for safety-critical systems [51]. They do not propose using BBNs to analyze confidence in the safety argument. Rather, they propose developing BBNs to analyze the probability of certain architectural events and citing this as evidence of the correctness of the architecture.

8 Conclusion

The burden of proving that an assurance argument confidence quantification technique produces trustworthy assessments lies with its proposers. But the published papers concerning such techniques do not provide or reference strong empirical evidence for the hypothesis that typical safety analysts using the technique will produce confidence assessments that are sufficiently trustworthy to serve as a basis for the decision to accept or reject a critical system based on its assurance argument.

Nonetheless, we investigated the proposed techniques we were aware of. For each technique, we attempted to replicate the authors' examples and derive counterexamples that produce implausible results. We found a counterexample for each technique for which there was a compelling worked example we could reproduce.

We do not claim that our results show that the probability theories underlying the proposed techniques are in error. However, the proposed uses of those theories—the techniques we review—are, at best, imperfect. Without research that provides strong, direct evidence that the resulting confidence assessments are trustworthy, there is no plausible justification for relying on one of these techniques in making decisions about which critical systems to deploy or continue to operate.

References

1. U.S. Chemical Safety and Hazard Investigation Board: Chevron Richmond Refinery Pipe Rupture and Fire. Regulatory Report 2012-03-I-CA, CSB, 2014. URL `http://www.csb.gov/chevron-refinery-fire/`.

2. A-P-T Research, Inc.: Safety Case Workshop. Electronic document, 2014. URL `http://www.apt-research.com/news/2014-01-15_SafetyCaseWorkshop/T-13-00600%20Safety%20Case%20Workshop%20Findings.pdf`.

3. Attwood, K.; et al.: *GSN Community Standard Version 1*. Origin Consulting Limited, York, UK, November 2011. URL `http://www.goalstructuringnotation.info/documents/GSN_Standard.pdf`.

4. Graydon, P.; Knight, J.; and Green, M.: Certification and Safety Cases. *Proceedings of the 28th International Systems Safety Conference (ISSC)*, Minneapolis, Minnesota, USA, August–September 2010. URL `http://www.cs.virginia.edu/~pjg2e/documents/2010.issc.pdf`.

5. Kelly, T. P.: Reviewing Assurance Arguments — A Step-By-Step Approach. *Proceedings of the Workshop on Assurance Cases for Security — The Metrics Challenge*, Edinburgh, UK, July 2007. URL `http://www-users.cs.york.ac.uk/~tpk/pubs.html`.

6. Goodenough, J. B.; Weinstock, C. B.; and Klein, A. Z.: Eliminative Induction: A Basis for Arguing System Confidence. *Proceedings of the International Conference on Software Engineering (ICSE)*, San Francisco, CA, USA, May 2013. URL `http://www.sei.cmu.edu/library/assets/whitepapers/icse13nier-id10-p-15833-preprint.pdf`.

7. Hawkins, R.; Kelly, T.; Knight, J.; and Graydon, P.: A New Approach to Creating Clear Safety Arguments. *Advances in Systems Safety: Proceedings of the 19th Safety-Critical Systems Symposium (SSS)*, Southampton, UK, February 2011, pp. 3–23. URL http://www.cs.virginia.edu/~jck/publications/SSS.2011.safety.cases.pdf.

8. NOR-STA. URL http://www.nor-sta.eu/en, last accessed 2015-12-09.

9. Graydon, P. J.: Uncertainty and Confidence in Safety Logic. *Proceedings of the 31st International System Safety Conference (ISSC)*, Boston, MA, USA, August 2013. URL http://www.mrtc.mdh.se/index.php?choice=publications&id=3349.

10. Littlewood, B.: Dependability Assessment of Software-based Systems: State of the Art. *Proceedings of the 27th International Conference on Software Engineering (ICSE)*, May 2005, pp. 6–7. URL http://dx.doi.org/10.1109/ICSE.2005.1553530.

11. Lord Robens; Beeby, G. H.; Pike, M.; Robinson, S. A.; Shaw, A.; Windeyer, B. W.; Wood, J. C.; Wake, M.; and Neale, C.: *Report of the Committee on Safety and Health at Work*. Her Majesty's Stationery Office, London, UK, July 1972.

12. Moran, R.; et al.: Workplace Safety and Health: Multiple Challenges Lengthen OSHA's Standard Setting. Report to Congressional Requesters GAO-12-330, U.S. Government Accountability Office, Washington, DC, USA, April 2012. URL http://www.gao.gov/assets/590/589825.pdf.

13. Cullen, W. D.: *The Public Inquiry Into the Piper Alpha Disaster*, vol. 2. Her Majesty's Stationery Office, 1990.

14. Cullen, PC, T. R. H. L.: *The Ladbroke Grove Rail Inquiry: Part 2 Report*. HSE Books, Norwich, UK, 2001. URL http://www.railwaysarchive.co.uk/documents/HSE_Lad_Cullen002.pdf.

15. Interim Defence Standard 00-56, Issue 5: *Safety Management Requirements for Defence Systems — Part 1: Requirements and Guidance*. (U.K.) Ministry of Defence, Glasgow, UK, February 2014.

16. ISO 26262: *Road vehicles — Functional safety*. International Organization for Standardization, 2011.

17. Center for Devices and Radioogical Health (CDRH): *Infusion Pumps Total Product Life Cycle: Guidance for Industry and FDA Staff*. US Food and Drug Administration (FDA), December 2014. URL http://www.fda.gov/downloads/medicaldevices/

deviceregulationandguidance/guidancedocuments/ucm209337.pdf.

18. Ayoub, A.; Chang, J.; Sokolsky, O.; and Lee, I.: Assessing the Overall Sufficiency of Safety Arguments. *Assuring the Safety of Systems: Proceedings of the Twenty-first Safety-Critical Systems Symposium (SSS)*, Bristol, UK, February 2013. URL http://scsc.org.uk/p119.

19. Duan, L.; Rayadurgam, S.; Heimdahl, M. P. E.; Sokolsky, O.; and Lee, I.: Representing Confidence in Assurance Case Evidence. *Proceedings of the 3rd International Workshop on Assurance Cases for Safety-intensive Systems (ASSURE)*, Delft, The Netherlands, September 2015, pp. 15–26. URL http://dx.doi.org/10.1007/978-3-319-24249-1_2.

20. Hobbs, C.; and Lloyd, M.: The Application of Bayesian Belief Networks to Assurance Case Preparation. *Achieving Systems Safety: Proceedings of the 20th Safety-Critical Systems Symposium (SSS)*, Bristol, UK, February 2012, pp. 159–176. URL http://dx.doi.org/10.1007/978-1-4471-2494-8_12.

21. Nair, S.; Walkinshaw, N.; and Kelly, T.: Quantifying Uncertainty in Safety Cases Using Evidential Reasoning. *Proceedings of the Workshop on Next Generation of System Assurance Approaches for Safety-Critical Systems (SASSUR)*, Florence, Italy, September 2014, pp. 413–418. URL http://dx.doi.org/10.1007/978-3-319-10557-4_45.

22. Nair, S.; Walkinshaw, N.; Kelly, T.; and de La Vara, J. L.: An Evidential Reasoning Approach for Assessing Confidence in Safety Evidence. *Proceedings of the 26th IEEE International Symposium on Software Reliability Engineering (ISSRE)*, Washington, DC, USA, November 2015.

23. The SERENE Partners: The SERENE Method Manual. Task Report SERENE/5.3/CSR/3053/R/1, The SafEty and Risk Evaluation using bayesian NEts (SERENE) project, 1999. URL http://www.eecs.qmul.ac.uk/~norman/papers/serene.pdf.

24. Zeng, F.; Lu, M.; and Zhong, D.: Using D-S Evidence Theory to Evaluation of Confidence in Safety Case. *Journal of Theoretical and Applied Information Technology*, vol. 47, no. 1, January 2013, pp. 184–189. URL http://www.jatit.org/volumes/Vol47No1/22Vol47No1.pdf.

25. Zhao, X.; Zhang, D.; Lu, M.; and Zeng, F.: A New Approach to Assessment of Confidence in Assurance Cases. *Proceedings of the 31st*

International Conference on Computer Safety, Reliability, and Security (SAFECOMP) Workshops, September 2012, pp. 79–91. URL http://dx.doi.org/10.1007/978-3-642-33675-1.

26. Cyra, L.; and Gorski, J.: Supporing Expert Assessment of Argument Structures in Trust Cases. *Proceedings of the 9th International Probabilistic Safety Assessment and Management Conference (PSAM)*, Hong Kong, China, 2008, pp. 1–9. URL http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.163.1409&rep=rep1&type=pdf.

27. Cyra, L.; and Górski, J.: Support for Argument Structures Review and Assessment. *Reliability Engineering and System Safety*, vol. 96, no. 1, 2011, pp. 26–37. URL http://dx.doi.org/10.1016/j.ress.2010.06.027, special Issue on Safecomp 2008.

28. Denney, E.; Pai, G.; and Habli, I.: Towards Measurement of Confidence in Safety Cases. *Proceedings of the 5th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement (ESEM)*, Banff, Alberta, Canada, September 2011. URL http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20110016239.pdf.

29. Guiochet, J.; Hoang, Q. A. D.; and Kaâniche, M.: A Model for Safety Case Confidence Assessment. *Proceedings of the International Conference on Computer Safety, Reliability, & Security (SAFECOMP)*, Delft, The Netherlands, September 2015. URL http://dx.doi.org/10.1007/978-3-319-24255-223.

30. Guo, B.: Knowledge Representation and Uncertainty Management: Applying Bayesian Belief Networks to a Safety Assessment Expert System. *Proceedings of the International Conference Natural Language Processing and Knowledge Engineering*, Beijing, China, October 2003, pp. 114–119. URL http://dx.doi.org/10.1109/NLPKE.2003.1275879.

31. Nair, S.; Walkinshaw, N.; Kelly, T.; and de La Vara, J. L.: An Evidential Reasoning Approach for Assessing Confidence in Safety Evidence. Technical Report 2014-17, Simula Research Laboratory, Fornebu, Norway, November 2014. URL https://www.simula.no/publications/evidential-reasoning-approach-assessing-confidence-safety-evidence.

32. Yamamoto, S.: Assuring Security through Attribute GSN. *Proceedings of the 5th International Conference on IT Convergence and Security (ICITCS)*, August 2015, pp. 1–5. URL http://dx.doi.org/10.1109/ICITCS.2015.7292954.

33. Johnson, R. H.: Some Reflections on the Informal Logic Initiative. *Studies in Logic, Grammar and Rhetoric*, vol. 16, no. 29, 2009, pp. 17–46. URL http://logika.uwb.edu.pl/studies/index.php?page=search&vol=29.

34. Toulmin, S. E.: *The Uses of Argument*. Cambridge University Press, New York, NY, USA, updated ed., 2003.

35. Walton, D.; Reed, C.; and Macagno, F.: *Argumentation Schemes*. Cambridge University Press, Cambridge, UK, 2008.

36. Hawthorne, J.: Inductive Logic. *Stanford Encyclopedia of Philosophy*, 2012. URL http://plato.stanford.edu/entries/logic-inductive/#2.1.

37. Keynes, J. M.: *A Treatise on Probability*. Macmillan and Co., Ltd., St. Martin's St., London, UK, 1921.

38. Shafer, G.: *A Mathematical Theory of Evidence*. Princeton University Press, 1976.

39. Yang, J.-B.; and Xu, D.-L.: On the Evidential Reasoning Algorithm for Multiple Attribute Decision Analysis Under Uncertainty. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 32, no. 3, May 2002, pp. 289–304. URL http://dx.doi.org/10.1109/TSMCA.2002.802746.

40. Cohen, L. J.: *The Probable and the Provable*, vol. 10 of *Clarendon Library of Logic and Philosophy*. Clarendon Press, Oxford, UK, 1977.

41. Jøsang, A.; Pope, S.; and Daniel, M.: Conditional Deduction Under Uncertainty. *Proceedings of the 8th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU)*, Barcelona, Spain, 2005, pp. 471–478. URL http://folk.uio.no/josang/papers/JPD2005-ECSQARU.pdf.

42. Jøsang, A.: A Logic for Uncertain Probabilities. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, vol. 9, no. 3, June 2001, pp. 279–311. URL http://eprints.qut.edu.au/7204/1/Jos2001-IJUFKS.pdf.

43. Strigini, L.: Formalism and Judgement in Assurance Cases. *Proceedings of the Workshop on Assurance Cases: Best Practices, Possible Obstacles, and Future Opportunities*, Florence, Italy, June 2004. URL http://openaccess.city.ac.uk/279/2/LS_formal_judgement_v06.pdf.

44. Nichols, A. L.; and Maner, J. K.: The Good-Subject Effect: Investigating Participant Demand Characteristics. *Journal of General Psychology*, vol. 135, no. 2, Apr. 2008, pp. 151–166.

45. OREDA: Offshore & Onshore Reliability Data. Web page: https://www.oreda.com/.

46. Graydon, P. J.: Towards a Clearer Understanding of Context and Its Role in Assurance Argument Confidence. *Proceedings of the International Conference on Computer Safety, Reliability and Security (SAFECOMP)*, Florence, Italy, September 2014. URL http://dx.doi.org/10.1007/978-3-319-10506-2_10.

47. Graydon, P. J.; and Holloway, C. M.: "Evidence" Under a Magnifying Glass: Thoughts on Safety Argument Epistemology. *Proceedings of the IET System Safety and Cyber Security Conference*, Bristol, UK, October 2015. URL http://www.researchgate.net/publication/280247123_Evidence_Under_a_Magnifying_Glass_Thoughts_on_Safety_Argument_Epistemology.

48. Richardson, J.; Smith, A.; and Meaden, S.: Your Logical Fallacy is No True Scotsman. Web page. URL https://yourlogicalfallacyis.com/no-true-scotsman, last accessed 14 December 2015.

49. Fallacy Files: Overprecision. Web page. URL http://www.fallacyfiles.org/fakeprec.html, last accessed 15 December 2015.

50. Littlewood, B.; and Wright, D.: 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. *IEEE Transactions on Software Engineering*, vol. 33, no. 5, May 2007, pp. 347–365. URL http://dx.doi.org/10.1109/TSE.2007.1002.

51. Wu, W.; and Kelly, T.: Combining Bayesian Belief Networks and the Goal Structuring Notation to Support Architectural Reasoning About Safety. *Proceedings of the 26th International Conference on Computer Safety, Reliability, and Security (SAFECOMP)*, Nuremberg, Germany, September 2007, pp. 172–186. URL http://dx.doi.org/10.1007/978-3-540-75101-4_17.

52. Park, R. E.; Goethert, W. B.; and Florac, W. A.: Goal-Driven Software Measurement: A Guidebook. Handbook CMU/SEI-96-HB-002, Software Engineering Institute, Pittsburgh, PA, USA, August 1996. URL http://www.sei.cmu.edu/reports/96hb002.pdf.

53. Fenton, N. E.; Neil, M.; and Caballero, J. G.: Using Ranked Nodes to Model Qualitative Judgments in Bayesian Networks. *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 10, October 2007, pp. 1420–1432. URL http://dx.doi.org/10.1109/TKDE.2007.1068.

54. Jøsang, A.: Artificial Reasoning with Subjective Logic. *Proceedings of the 2nd Australian Workshop on Commonsense Reasoning (AWCR)*, Australian Computer Society, Perth, Australia, December 1997. URL `http://folk.uio.no/josang/papers/Jos1997-AWCR.pdf`.

55. Jøsang, A.; Hayward, R.; and Pope, S.: Trust Network Analysis with Subjective Logic. *Proceedings of the 29th Australasian Computer Science Conference (ACSC)*, Australian Computer Society, 2006, pp. 85–94.

56. Duan, L.: Personal communication. October 2015.

57. Adelard: ASCAD: Adelard Safety Case Development Manual. Electonic document, London, UK, 1988.

58. Runeson, P.; and Höst, M.: Guidelines for conducting and reporting case study research in software engineering. *Empirical Software Engineering*, vol. 14, 2008, pp. 131–164. URL `http://dx.doi.org/10.1007/s10664-008-9102-8`.

Appendix A

Ayoub, Chang, Sokolsky, and Lee

Ayoub, Chang, Sokolsky, and Lee propose to compute the sufficiency of safety arguments using Dempster–Shafer theory [18].

A.1 Proposed Technique

In the proposed technique, the analyst uses an existing confidence argument and a procedure not specified in the paper "to make his/her assessments on the sufficiency and insufficiency of the evidence nodes" [18]. The analyst then uses aggregation rules to assess claims, working from evidence toward the main safety claim. While the paper describes this process as "automatic," the analyst must make judgements about evidence, select aggregation rules, and provide weights and scaling factors.

Illustrative Example. The paper illustrates the technique with the example depicted in Figure A1 and presents calculations for several situations, including one where "the expert opinion is that the justification J1 covers only 80% of the cases." Table A1 gives the evidence sufficiency figures from the paper.

Claim Supported by a Single Piece of Evidence. When a claim is supported by exactly one piece of evidence, such as G2 in Figure A1, it is assessed the same as its supporting evidence.

Claim Supported by an Alternative Argument. The paper refers to arguments such as that supporting G3 in Figure A1 "where more than one independent support of the common conclusion is provided" as *alternative arguments* [18]. Analysts calculate the sufficiency ($m_c(S)$) and insufficiency ($m_c(I)$) of the such arguments from premises a and b using the following formulae:

$$m_{c=a\oplus b}(S) = \frac{m_a(S)m_b(S) + m_a(S)m_b(U) + m_a(U)m_b(S)}{1 - (m_a(S) * m_b(I) + m_a(I)m_b(S))} \tag{A1}$$

$$m_{c=a\oplus b}(I) = \frac{m_a(I)m_b(I) + m_a(I)m_b(U) + m_a(U)m_b(I)}{1 - (m_a(S) * m_b(I) + m_a(I)m_b(S))} \tag{A2}$$

The uncertainty is the remaining mass:

$$m_c(U) = 1 - (m_c(S) + m_c(I)) \tag{A3}$$

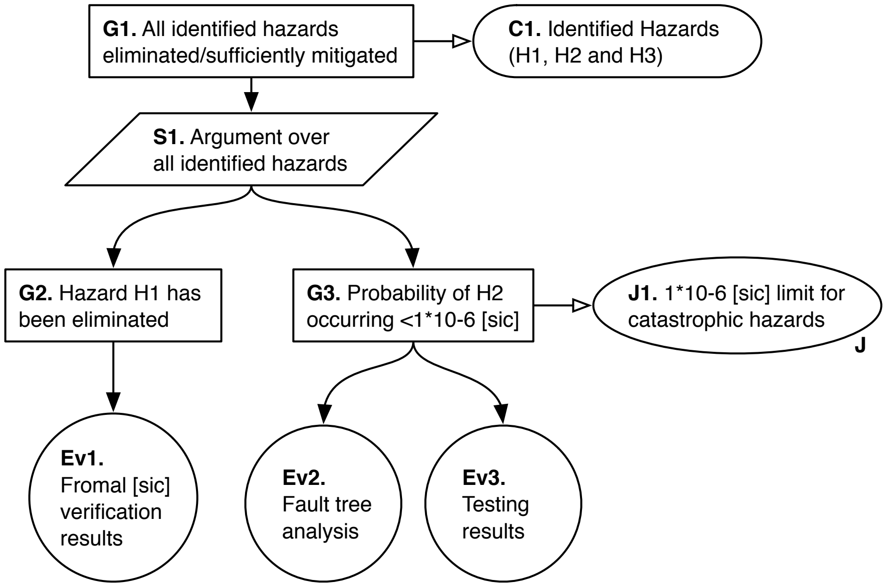

G1. All identified hazards eliminated/sufficiently mitigated
C1. Identified Hazards (H1, H2 and H3)
S1. Argument over all identified hazards
G2. Hazard H1 has been eliminated
G3. Probability of H2 occurring <1*10-6 [sic]
J1. 1*10-6 [sic] limit for catastrophic hazards
J
Ev1. Fromal [sic] verification results
Ev2. Fault tree analysis
Ev3. Testing results

Figure A1: *An example safety argument rendered in the Goal Structuring Notation (GSN) [3] and taken from [18].*

Evidence	Sufficient	Insufficient	Uncertainty
Ev1	0.800	0.100	0.100
Ev2	0.700	0.100	0.200
Ev3	0.500	0.200	0.300
G1	0.807	0.106	0.087
G2	0.800	0.100	0.100
G3	0.815	0.111	0.074
S1	0.807	0.106	0.087

Table A1: *Sufficiency of Figure A1's nodes, taken from [18].*

Claim Supported by a Disjoint Argument. The paper refers to arguments such as that supporting S1 in Figure A1 "where the supporting nodes provide complementary support for the conclusion" as *disjoint arguments* [18]. Analysts calculate the sufficiency ($m_c(S)$) and insufficiency ($m_c(I)$) of such arguments from premises $1 \ldots n$ using the following formulae:

$$m_c(S) = \frac{\sum_{i=1}^{n} w_i * m_i(S)}{\sum_{i=1}^{n} w_i} \tag{A4}$$

$$m_c(I) = \frac{\sum_{i=1}^{n} w_i * m_i(I)}{\sum_{i=1}^{n} w_i} \tag{A5}$$

The paper does not specify how to obtain the weights $w_1 \ldots w_n$ generally but uses equal weights in the argument supporting S1.

Test for Argument Sufficiency. While the paper does not explicitly define a test for argument acceptance, the authors reject an alternative version of the argument in Figure A1 because its "degree of uncertainty is very high relative to the degree of belief of the argument sufficiency."

A.2 Replication

We replicated the base case of the paper's example. H1 and H2 are the only identified hazards and the sufficiency and insufficiency of G1 are the same as that of S1. Using Equations A1–A5 and the assessments of Ev1, Ev2, and Ev3 given in Table A1, we calculated the assessments of G1–G3 and S1 shown in Table A1. Our results do not match the authors' figures. The authors appear to have used a variant of the G3 assessment (labeled as Case 3 of G2 in the paper) as input to the base-case assessment of S1 and G1. In any case, since G1's sufficiency (0.807) is much higher than its uncertainty (0.087), this argument is acceptable per the technique.

A.3 Hypotheses and Evidence

The paper does not present an empirical evaluation of the efficacy of the proposed technique [18]. Nevertheless, the authors' "preliminary experience of applying the proposed method has revealed that the assessing mechanism yields the expected benefits in guiding the safety argument reviewer and helping him/her to reduce the effect of the confirmation bias mindset." The paper speculates that the latter effect will result from asking the analyst's opinion on both sufficiency and insufficiency.

A.4 Counterargument

To show that confidence computed using the proposed technique is not an appropriate basis for making release to service decisions in all cases, we use four variants of the example presented in the paper: *Many Hazards*, *Undermined Evidence 1*, *Undermined Evidence 2*, and *Counterevidence*. The examples collectively illustrate two problems with the proposed technique: (1) that it allows evidence of the mitigation of one hazard to hide a lack of evidence of the mitigation of another or even evidence showing that another is inadequately mitigated, and (2) that it gives different results depending on arbitrary choices about hazard scope.

Many Hazards Example. While some systems might have only two hazards, other systems will have dozens. In our Many Hazards variant of the paper's example, the system addresses twenty-one hazards. We assume that the first twenty of these have been eliminated in design like H1 and so replace G2 and Ev1 in Figure A1 with G2.1–G2.20 and Ev1.1–Ev1.20, respectively. Figure A2 shows the resulting argument.

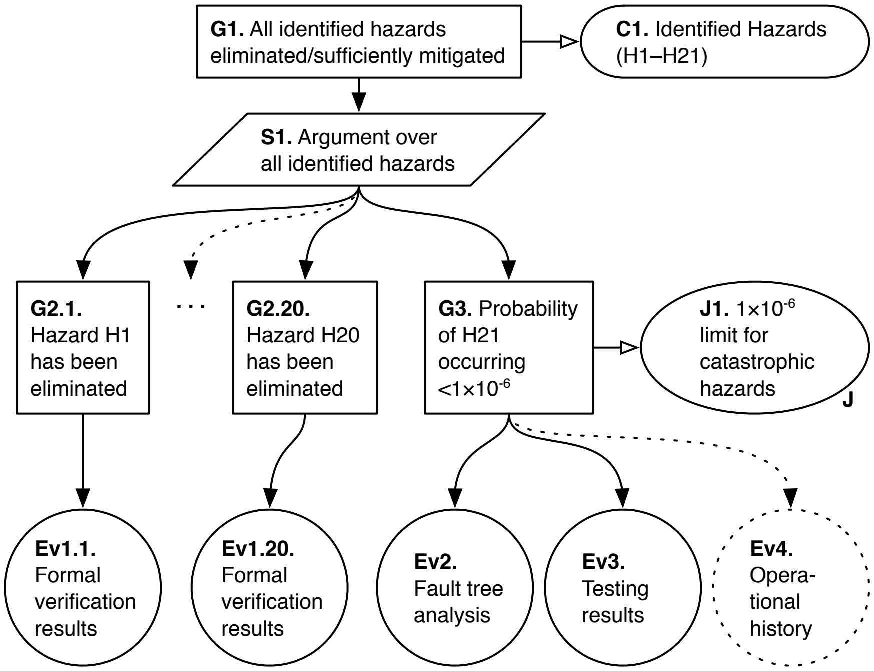

G1. All identified hazards eliminated/sufficiently mitigated
C1. Identified Hazards (H1–H21)
S1. Argument over all identified hazards
G2.1. Hazard H1 has been eliminated
...
G2.20. Hazard H20 has been eliminated
G3. Probability of H21 occurring <1×10-6
J1. 1×10-6 limit for catastrophic hazards
J
Ev1.1. Formal verification results
Ev1.20. Formal verification results
Ev2. Fault tree analysis
Ev3. Testing results
Ev4. Opera-tional history

Figure A2: *Our Many Hazards (with G2.2–G2.19 and Ev1.2–Ev.19 but not Ev4), Undermined Evidence 1 (with G2.2–G2.19 and Ev1.2–Ev.19 but not Ev4), Undermined Evidence 2 (without G2.2–G2.19, Ev1.2–Ev.19 or Ev4) and Counter Evidence (with G2.2–G2.19, Ev1.2–Ev.19, and Ev4) variants of the example safety argument shown in Figure A1.*

Undermined Evidence 1 and 2 Examples. Suppose that the evidence regarding hazard H20 is insufficient: the premises to be verified were entered incorrectly into the theorem prover, so the proof does not show what it was meant to show. Our Undermined Evidence 1 example is identical to Many Hazards except that $m_{\mathsf{Ev1.20}}(S) = 0.050$ and $m_{\mathsf{Ev1.20}}(I) = 0.900$. Our Undermined Evidence 2 example is identical to Many Hazards except that it lacks hazards H2–H19 because the definition of hazard H1 has been expanded to include these.

Counterevidence Example. It is possible that operational history will confirm or deny safety claims. Our Counterevidence example is identical to Many Hazards except that goal G3 is additionally supported by a solution Ev4 citing operational history. The sufficiency of G3 is given by:

$$m_{\mathsf{G3}}(S) = (m_{\mathsf{Ev2}}(S) \oplus m_{\mathsf{Ev3}}(S)) \oplus m_{\mathsf{Ev4}}(S) \tag{A6}$$

Similarly, the insufficiency of G3 is given by repeated applications of Equation A2. We suppose that history shows that hazard H21 has been occurring at a rate far exceeding the limit specified in G3 and J1 and thus that $m_{\mathsf{Ev4}}(S) = 0.050$ and $m_{\mathsf{Ev4}}(I) = 0.900$.

Element(s)	Variant	Suff.	Insuff.	Uncert.
Ev1.1	All	0.800	0.100	0.100
Ev1.2–19	All but Undermined Evidence 2	0.800	0.100	0.100
Ev1.20	Many Hazards, Counterevidence	0.800	0.100	0.100
	Undermined Evidence 1 and 2	0.050	0.900	0.050
Ev2	All	0.700	0.100	0.200
Ev3	All	0.500	0.200	0.300
Ev4	Counterevidence	0.050	0.900	0.050
G3	All but Counterevidence	0.815	0.111	0.074
	Counterevidence	0.326	0.660	0.014
S1, G1	Many Hazards	0.801	0.101	0.099
	Undermined Evidence 1	0.765	0.139	0.096
	Undermined Evidence 2	0.079	0.053	0.868
	Counterevidence	0.777	0.127	0.096

Table A2: *The sufficiency, insufficiency, and uncertainty of the evidence, Hazard H21 mitigation claim, and main safety claim in the example arguments shown in Figure A2.*

Analysis of examples. Table A2 gives the calculated assessment of G1 for all four variants. It is not implausible that $m_{\mathsf{G1}}(S)$ is 0.801 in the Many Hazards example. But the proposed technique computes similar values of $m_{\mathsf{G1}}(S)$ for the Undermined Evidence 1 and Counterevidence variants. Because an uncertainty of 0.096–0.099 is not very high relative to a sufficiency of 0.765–0.801, the proposed technique shows these three example arguments to be acceptable. But it is not plausible that the system is acceptable either despite a lack of substantial evidence that one of its hazards is mitigated or in the presence of evidence that one is not mitigated. These examples show that it might be dangerous to base release-to-service decisions solely on the calculated argument sufficiency: an analyst who did so might decide that the lack of evidence or presence of counter evidence does not indicate a problem that must be addressed.

The Undermined Evidence 1 and 2 examples demonstrate that calculated confidence is highly sensitive to arbitrary decisions about hazard scope. In both of these examples, the same dangerous system states are mitigated in the same way. The only difference between them is that the state that is defined as hazards H1–H19 in Undermined Evidence 1 is defined as the single hazard H1 in Undermined Evidence 2. While the technique assesses the safety argument to be acceptable in the former case, it finds the safety argument unacceptable in the latter. It is not plausible than an arbitrary difference in hazard scope should change our confidence in system safety.

Note. It might have been more appropriate to represent undermined evidence of the mitigation of hazard H20 as $m_{\mathsf{Ev1.20}}(S) = 0.050$, $m_{\mathsf{Ev1.20}}(I)$ $= 0.050$, and $m_{\mathsf{Ev1.20}}(U) = 0.900$ rather than 0.050, 0.900, and 0.050, respectively. Had we done so, $m_{\mathsf{G1}}(S) = 0.765$, $m_{\mathsf{G1}}(I) = 0.098$, and $m_{\mathsf{G1}}(U) = 0.137$ in the Undermined Evidence 1 case and 0.079, 0.012, and 0.908, respectively, in the Undermined Evidence 2 case. Our choice to represent undermined evidence as insufficiency rather than uncertainty does not substantially affect our counterexamples.

Appendix B

Cyra and Górski

Cyra and Górski propose to evaluate confidence in assurance cases (which they call *trust cases*) using Dempster–Shafer theory [26, 27].

B.1 Proposed Technique

In the proposed technique, assurance arguments are composed of *claims* supported by *arguments*. Each argument has *premises* and, in some cases, a *warrant*. Each premise might be another claim (supported by further argument), a *fact* (associated with a reference to the evidence), or an *assumption*. The analyst rates each assumption, fact, and warrant using Jøsang's opinion triangle [27, 41]. On one axis of the triangle, decisions are rated on the scale:

Acceptable	$\mathit{Dec}(s) = 3/3$
Tolerable	$\mathit{Dec}(s) = 2/3$
Opposable	$\mathit{Dec}(s) = 1/3$
Rejectable	$\mathit{Dec}(s) = 0/3$

On the other axis, confidence in those decisions is rated on the scale:

For sure	$\mathit{Conf}(s) = 5/5$
With very high confidence	$\mathit{Conf}(s) = 4/5$
With high confidence	$\mathit{Conf}(s) = 3/5$
With low confidence	$\mathit{Conf}(s) = 2/5$
With very low confidence	$\mathit{Conf}(s) = 1/5$
Lack of confidence	$\mathit{Conf}(s) = 0/5$

The analyst converts these assessments into Dempster–Shafer belief and plausibility values and computes confidence in the conclusion of each reasoning step by applying one of four different sets of formulae depending on the type of argument [27].

Complementary Arguments. The paper defines arguments in which "each of the premises 'covers' part of the conclusion" as *complementary arguments* (*C-args*) [27]. The analyst assessing a complementary argument converts his or her assessment of the warrant and premises into belief $\mathit{Bel}(s)$ and plausibility $\mathit{Pl}(s)$ values using

$$\mathit{Bel}(s) = \mathit{Conf}'(s) \cdot \mathit{Dec}'(s) \tag{B1}$$

and

$$\mathit{Pl}(s) = 1 - \mathit{Conf}'(s) \cdot \left(1 - \mathit{Dec}'(s)\right) \tag{B2}$$

where

$$\left\langle \mathit{Conf}'(s), \mathit{Dec}'(s) \right\rangle = s_C \left(\left\langle \mathit{Conf}(s), \mathit{Dec}(s) \right\rangle \right) \tag{B3}$$

and $s_C : [0,1] \times [0,1] \rightarrow [0,1] \times [0,1]$ is a scaling function for use in complementary arguments. The paper provides a graphic representation of the effect of s_C on confidence but not on the decision. The analyst then combines the converted values of the premises $a_1 \ldots a_n$ and warrant w using

$$Bel(c) = Bel(w) \cdot \frac{\sum_{i=1}^{n} k_i Bel\left(a_i\right)}{\sum_{i=1}^{n} k_i} \tag{B4}$$

and

$$Pl(c) = 1 - Bel(w) \cdot \left(1 - \frac{\sum_{i=1}^{n} k_i Pl(a_i)}{\sum_{i=1}^{n} k_i}\right) \tag{B5}$$

The paper does not specify how to obtain the weights k_i. The analyst then converts the combined values back into decision and confidence figures using

$$Dec'\left(c\right) = \begin{cases} \frac{Bel(c)}{Bel(c)+1-Pl(c)} & Bel\left(c\right) + 1 - Pl\left(c\right) \neq 0 \\ 1 & Bel\left(c\right) + 1 - Pl\left(c\right) = 0 \end{cases} \tag{B6}$$

$$Conf'\left(c\right) = Bel\left(c\right) + 1 - Pl\left(c\right) \tag{B7}$$

and

$$\langle Conf(c), Dec(c) \rangle = s_C^{-1}\left(\langle Conf'(c), Dec'(c) \rangle\right) \tag{B8}$$

Necessary and Sufficient Condition List Arguments. An analyst assessing a *necessary and sufficient condition list* argument (*NSC-arg*) converts decision and confidence assessments into belief and plausibility using Equations B1 and B2 and an equation similar to Equation B3 except using a different, unspecified scaling function s_{NSC}. The analyst then combines the belief and plausibility figures for the premises $a_1 \ldots a_n$ and warrant w using

$$Bel(c) = Bel(w) \cdot \prod_{i=1}^{n} Bel(a_i) \tag{B9}$$

and

$$Pl(c) = 1 - Bel(w) \cdot \left(1 - \prod_{i=1}^{n} Pl(a_i)\right) \tag{B10}$$

Again, the combined belief and plausibility values can be converted back to decision and confidence figures using Equations B6 and B7 and an equation similar similar to Equation B8 using the scaling function s_{NSC}.

Conversion to Assessment Scales. Analysts and tools convert computed decision and confidence values into their linguistic equivalents by choosing the nearest linguistic value.

Test for Argument Sufficiency. The paper does not specify how to use computed results to determine whether a system is sufficiently safe or secure to be put into service.

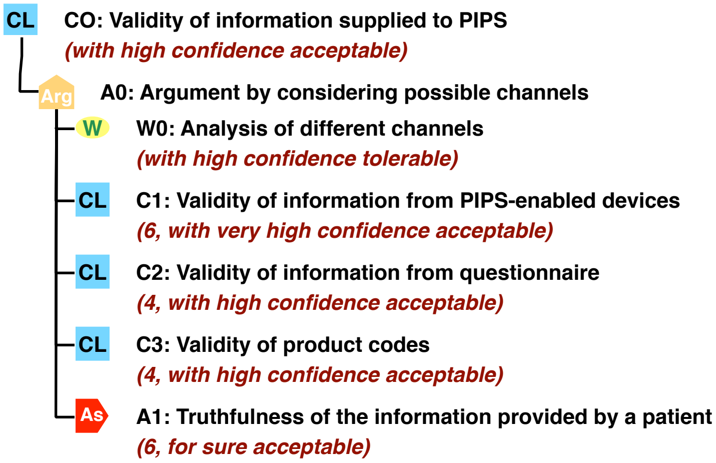

CL
CO: Validity of information supplied to PIPS
(with high confidence acceptable)
Arg
A0: Argument by considering possible channels
W
W0: Analysis of different channels
(with high confidence tolerable)
CL
C1: Validity of information from PIPS-enabled devices
(6, with very high confidence acceptable)
CL
C2: Validity of information from questionnaire
(4, with high confidence acceptable)
CL
C3: Validity of product codes
(4, with high confidence acceptable)
As
A1: Truthfulness of the information provided by a patient
(6, for sure acceptable)

Figure B1: *An example trust argument taken from [27]. PIPS stands for the name of the example system.*

Illustrative Example. The papers give several examples about the Personalised Information Platform for Health and Life Services (PIPS) system argument to illustrate the proposed technique [27]. Figure B1 depicts the paper's argument for that system.

B.2 Replication

Since the papers do not provide the scaling functions, we were unable to replicate the calculations in the paper's example by hand. However, the scaling functions are built into the latest version of the authors' tool, now called Nor-Sta [8]. The authors graciously granted us access to their tool, and with it we were able to replicate their calculations.

B.3 Hypotheses and Evidence

The objectives for the proposed technique are "(1) assessment of the compelling power of an argument and (2) communication of the result of such an assessment to relevant recipients" [27]. To assess the technique, the authors performed an "experimental calibration and validation of the aggregation mechanism."

Experimental Method. The paper reports,

> All scaling functions were calculated using the data obtained in the experiment. 31 students studying for the Master's degree in information technologies (the last year) were selected for the experiment. The participants had a good background in logic and mathematics and they also attended a two-hour lecture about trust cases.

> The participants were divided into three groups. Each group was supposed to apply one of the aggregation rules: A-rule, C-rule or NSC-rule . . . .
>
> Each participant was provided with five simple trust cases composed of a claim, an argument strategy, a warrant and premises (in the case of C-rule and NSC-rule) or a claim with a few argument strategies (in the case of A-rule). . . .
>
> The experiment participants were asked to assess the warrant and, in the case of C-rule to assign weights to the premises. Then, assuming the pre-defined assessments of each premise (in the case of C-rule and NSC-rule) or the assessments assigned to each of the argument strategies (in the case of A-rule) the participants were asked to give their assessment of the conclusion using the Assessment Triangle.

From the resulting data, the authors derived the four scaling functions and assessed both

> *Consistency of assessments*, measured by calculating the root-mean-square value of the difference between the first and repeated assessment (by the same participant) of the same conclusion with the same assessments assigned to the premises

and

> *Accuracy of assessments*, measured by calculating the root-mean-square value of the difference between a participant's assessment and the result of application of the aggregation rule.

Experimental Results. The experimental results show that individual students assess arguments somewhat inconsistently: the root-mean-squared value of consistency ranged between 0.62 of a linguistic scale category in the case of C-rule decisions and 1.03 of a category in the case of A-rule confidence. Individual assessments varied from what application of the proposed technique would predict by about as much: the root-mean-squared value of accuracy ranged between 0.66 of a category in the case of NSC-rule decisions and 1.10 of a category in the case of C-rule confidence.

Our Analysis. The authors present one of only two only empirical assessments of the efficacy of a quantified confidence technique we are aware of. But there are threats to the validity of the experiment that limit its value as evidence of the technique's efficacy. Two stand out: (1) the problem of using the same data to define and assess a technique, and (2) the representativeness of a small student sample.

The danger in using the same data to both define and assess a technique is that the technique might be effective only for the specimen

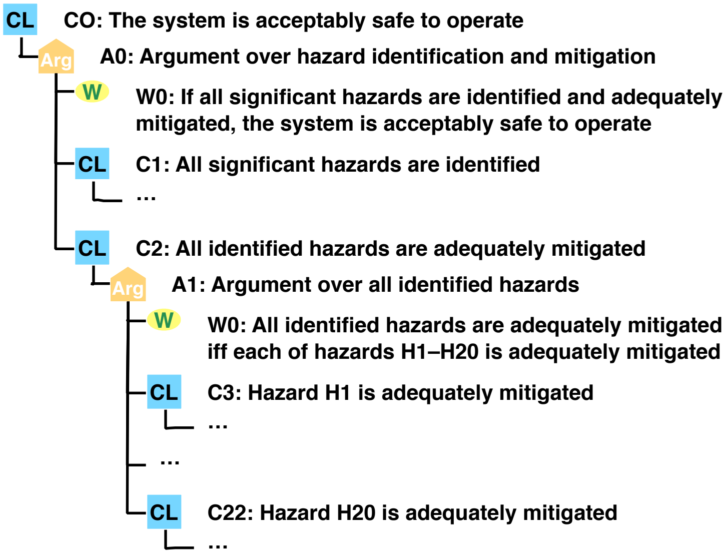

CL
CO: The system is acceptably safe to operate
Arg
A0: Argument over hazard identification and mitigation
W
W0: If all significant hazards are identified and adequately mitigated, the system is acceptably safe to operate
CL
C1: All significant hazards are identified
...
CL
C2: All identified hazards are adequately mitigated
Arg
A1: Argument over all identified hazards
W
W0: All identified hazards are adequately mitigated iff each of hazards H1–H20 is adequately mitigated
CL
C3: Hazard H1 is adequately mitigated
...
...
CL
C22: Hazard H20 is adequately mitigated
...

Figure B2: *A typical top-level safety argument in the style of the proposed technique [26, 27].*

problem from which it was derived. This problem can be addressed by using an approach such as *n-fold cross validation*. But the paper does not mention whether or how the authors addressed this problem.

It is not clear that the reasoning a small number of students used to assess a small number of simple example arguments can be taken as the norm for reasoning about confidence in all assurance cases. Students and seasoned security professionals might assess arguments differently. The five selected arguments of each type might not be representative of all arguments. And it is possible that people generally assess the specimen arguments incorrectly.

B.4 Counterargument

Since our expertise is in safety, not security, we use a safety argument to assess the proposed technique. Figure B2 depicts a typical top-level safety argument. A0 is a sufficient condition list argument (SC-arg) and A1 is a necessary and sufficient condition list argument (NSC-arg). Table B1 shows assessments for five variants of the argument in Figure B2, which we identify as *Optimistic*, *Counterevidence*, *Unsupported*, *Imperfect A* and *Imperfect B*. In all five variants, we are sure about the warrants for both arguments and have very high confidence in the completeness of hazard identification. We use the five examples to demonstrate two implausible features of the proposed technique: (1) it amplifies small doubts about each hazard's mitigation into large doubt about system safety, and (2) it produces similar confidence estimates for systems that should be regarded differently by stakeholders.

Node	Opt.	Cntrev.	Unsupp.	Impf. A	Impf. B
W0	5/5, 3/3				
W1	5/5, 3/3				
C1	4/5, 3/3				
C3	4/5, 3/3	4/5, 1/3	1/5, 3/3	4/5, 3/3	
C4–C7	5/5, 3/3			5/5, 3/3	
C8–C19	5/5, 3/3			4/5, 3/3	3/5, 3/3
C20–C22	5/5, 3/3			3/5, 3/3	
C2	4/5, 3/3	4/5, 1/3	1/5, 3/3	2/5, 3/3	1/5, 3/3
C0	4/5, 3/3	0/5, —	1/5, 3/3	1/5, 3/3	0/5, —

Table B1: *Assessments of the confidence and decision (Conf(s), Res(c)) for the Optimistic (Opt.), Counterevidence (Cntrev.), Unsupported (Unsupp.), Imperfect A (Impf. A), and Imperfect B (Impf. B) variants of the argument shown in Figure B2.*

Optimistic Example. In the Optimistic variant, we have very high confidence that hazard H1 is adequately mitigated and are sure that all other hazards are adequately mitigated. This is unrealistic: typical safety evidence does not entail adequate management of a hazard. But the result is that the safety claim is with very high confidence acceptable.

Counterevidence Example. In the Counterevidence variant, we are completely sure about the mitigation of all hazards except H1: we assess the claim that H1 is mitigated as with very high confidence opposable. That is, we have counterevidence against the claim that H1 is adequately managed. The result is a complete lack of confidence in the safety claim.

Unsupported Example. In the Unsupported variant, we are also completely sure about the mitigation of all hazards accept H1. In this example, we find the claim that H1 is adequately mitigated to be with very low confidence acceptable. That is, the claim that H1 is (nearly) unsupported. The result is that the safety claim is with very low confidence acceptable.

Imperfect A and B Examples. In the Realist A and Realist B variants, we are completely sure that some hazards are adequately mitigated (four in each case), have very high confidence that others are adequately mitigated (thirteen in Realist A and one in Realist B), and have only high confidence that the remaining hazards are adequately mitigated (three in Realist A and fifteen in Realist B). Since the Nor-Sta tool does not give $Conf(c)$ and $Res(c)$ numerically, we use two Realist variants to illustrate the difference needed to push an assessment from one linguistic category to the next. In any case, the result is that either the safety claim is with very low confidence acceptable (Realist A) or complete lack of confidence in the safety claim (Realist B).

Analysis of Examples. These examples demonstrate two implausible features of the proposed technique. First, the technique amplifies small doubts about how well each hazard has been mitigated into large doubt about the safety of the system. The Realist A case illustrates this effect: if we have high confidence or better in the mitigation of twenty hazards—very high confidence or certainty in the mitigation of all but three—we still have very low confidence in system safety. This effect is the result of multiplying together the assessor's beliefs in the adequacy of the mitigation of each hazard as specified by Equation B9 and will worsen as the number of hazards increases.

Second, the technique produces similar confidence estimations for (i) systems for which we have realistic doubts, such as the Imperfect A and B variants; (ii) systems for which we have little evidence about the mitigation of one hazard, such as the Unsupported variant; and (iii) systems in which we have reasonable evidence to show that one hazard isn't adequately mitigated, such as the Counterevidence variant. It is not plausible that confidence in our Imperfect A, Imperfect B, Counterevidence, and Undermined variants should be similar, as this means that any test of confidence sufficiency that would cause us to reject the latter two arguments would also cause us to reject the former two.

Appendix C

Denney, Pai, and Habli

Denney, Pai, and Habli propose to compute uncertainty in safety claims by constructing Bayesian Belief Networks (BBNs) that roughly mirror the structure of a safety argument [28].

C.1 Proposed Technique

Experts enumerate each source of uncertainty in the argument and create a BBN node representing the absence of each on the following scale:

Very Low	0–20%	probability
Low	20–40%	probability
Medium	40–60%	probability
High	60–80%	probability
Very High	80–100%	probability

Experts provide the value for each leaf node "from both quantitative data ... and from qualitative means." They use a truncated "normal distribution whose mean is the prior belief (or measure) of confidence and the variance is picked so as to appropriately represent the confidence in this prior itself." The authors note that "when only subjective judgment is available, quantifying confidence and selecting an appropriate prior distribution is problematic" but "believe that one way to address this issue is to identify metrics using techniques such as the Goal-Question-Metric" method [52].

For non-leaf nodes, the technique requires the analyst to "specify a prior conditional probability distribution ... in a parametric way, again using a truncated Normal distribution. Here, the mean of the distribution is the weighted average of the parent [random variable] while the variance is the inverse of the sum of the weights [53]."

Illustrative Example. Figure C1 shows the structure of the BBN with which the paper illustrates the technique.

Test for Argument Sufficiency. The paper states that part of the purpose of assessing confidence in an argument is to gauge whether its main claim should be accepted but does not specify a test for whether a given computed probability justifies putting a system into service.

C.2 Replication

The paper depicts the example BBN's structure and gives a histogram for each node but not the means and variances of leaf nodes or the

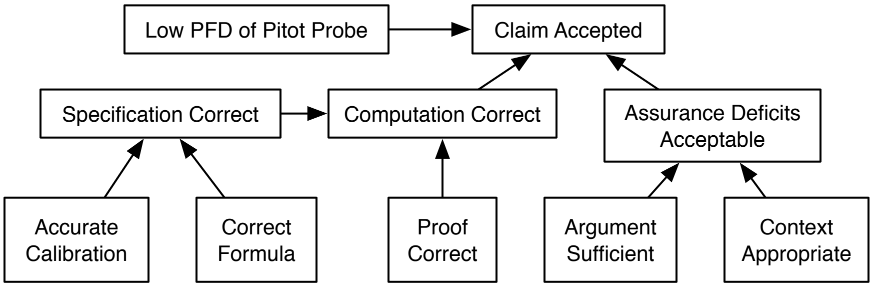


Figure C1: *An example BBN for part of a safety argument, taken from [28]. For compactness and consistency with other BBNs in this paper, we represent nodes as rectangles. 'PFD' stands for probability of failure on demand.*

Node	**Value**
Accurate Calibration	Mean=0.95, Variance=0.05
Correct Formula	Mean=0.85, Variance=0.01
Proof Correct	Mean=0.90, Variance=0.01
Argument Sufficient	Mean=0.50, Variance=0.05
Context Appropriate	Mean=0.50, Variance=0.05
Low PFD of Pitot Probe	Mean=0.85, Variance=0.05
Specification Correct	Accurate Calibration and Correct Formula both weighted 100
Computation Correct	Specification Correct and Proof Correct both weighted 100
Assurance Deficits Acceptable	Argument Sufficient and Context Appropriate both weighted 100
Claim Accepted	Low PFD of Pitot Probe and Assurance Deficits Acceptable weighted 100, Computation Correct weighted 200

Table C1: *Parameters for normal distributions of nodes shown in Figure C1.*

weights for non-leaf nodes. By trial and error, we arrived at parameters which approximate the example in the paper. We used ranked nodes to replicate the structure shown in Figure C1. For leaf nodes, we used the truncated normal distribution and the mean and variance values given in Table C1. For non-leaf nodes, we used the weights shown in Table C1 and our tool's weighted mean function with the variance $\sigma^2 = (\sum_{w:\text{weights}} w)^{-1}$. Table C2 gives the resulting distributions. The paper gives the distribution to thousandths of a percent, and our calculated values agree exactly. Our non-leaf nodes differ from those in the paper by as much as 0.9%, indicating that our weights differ slightly.

Node	Very Low		Low		Medium		High		Very High	
	R.E.	F.P.	R.E.	F.P.	R.E.	F.P.	R.E.	F.P.	R.E.	F.P.
Accurate Calibration (AC)			1.1%	1.1%	8.8%	8.8%	32.7%	32.7%	57.3%	57.3%
Correct Formula (CF)							32.4%	32.4%	66.9%	66.9%
Proof Correct (PC)							18.7%	18.7%	81.1%	81.1%
Argument Sufficient (AS)	7.9%	7.9%	24.4%	24.4%	35.4%	35.4%	24.4%	24.4%	7.9%	7.9%
Context Appropriate (CA)	7.9%	7.9%	24.4%	24.4%	35.4%	35.4%	24.4%	24.4%	7.9%	7.9%
Low PFD of Pitot Probe (LPFD)			2.7%	2.7%	14.7%	14.7%	37.4%	37.4%	45.0%	45.0%
Specification Correct (SC)					4.8%	4.8%	39.5%	40.0%	55.7%	55.0%
Computation Correct (CC)					2.4%	2.4%	35.8%	36.7%	61.8%	60.9%
Assurance Deficits Acceptable (ADA)	4.1%	4.1%	24.4%	24.4%	42.9%	42.9%	24.4%	24.4%	4.1%	4.1%
Claim Accepted (Top)					12.5%	12.7%	64.0%	64.1%	23.3%	23.0%

Table C2: *Calculated distributions of the nodes shown in Figure C1. We include both the figures from our reverse engineering (R.E.) and the original figures from the paper (F.P.)*

	Recreated Original	**Observed Failures**
Very Low	—	—
Low	—	5.3%
Medium	12.5%	59.8%
High	64.0%	34.5%
Very High	23.3%	—

Table C3: *Confidence in Figure C1's Claim Accepted.*

C.3 Hypotheses and Evidence

The authors motivate their work, in part, by noting that "subjectivity inherent in the structure of the argument and its supporting evidence ... pose[s] a key challenge to the measurement and quantification of confidence in the overall safety case." However, since the proposed technique derives the BBN's structure from that of the assurance argument, it could not possibly address this concern.

The paper hypothesizes that "where data for quantitative measurements can be systematically collected, quantitative arguments provide benefits over qualitative arguments in assessing confidence in the safety case." But data might be collected systematically (i.e., according to a defined procedure) and yet be inaccurate. The paper neither provides nor cites empirical evidence that assessors using figures from a given source produce more accurate assessments of argument confidence than qualitative assessments would provide.

The paper concludes that "linking qualitative safety arguments to quantitative arguments about uncertainty and confidence ... ensur[es] rigor in measuring confidence via probabilistic reasoning using [BBNs]." While the reader might infer that a rigorous measure of confidence provides a trustworthy result, the paper neither presents nor cites substantial empirical evidence for this hypothesis.

C.4 Counterargument

We created a Observed Failures variant of our recreation of the original example given in the paper. The premise of the variant is that the pitot tube is observed to fail more often than the limit represented by 'low' probability of failure on demand. We represent this by setting the the mean of Low PFD of Pitot Tube to 0 and its variance to 0.0001. Table C3 gives the results. While the correct functioning of the pitot tube is critical in this system, the result of the observed failures is to turn High confidence in the safety claim into Medium confidence. The paper does not define Medium confidence in terms of the effect it should on a decision to release a system into service or allow it to continue operation. But our understanding of 'medium confidence' is incompatible with a system that depends on a critical part that is known to fail too often.

The figures in Table C3 might be slightly different had we used slightly different weights. But the problem illustrated by the Observed Failures would remain. That is because confidence in the main safety claim is defined as the weighted mean of confidence in the reliability of the pitot tube and the correctness of the software and argument.

Appendix D

Duan, Rayadurgam, Heimdahl, Sokolsky, and Lee

Duan, Rayadurgam, Heimdahl, Sokolsky, and Lee propose using the beta distribution and Jøsang's opinion triangle to model uncertainty in assurance case evidence [19].

D.1 Proposed Technique

The proposed technique assumes the existence of a safety argument. Analysts express their opinion about "evidence nodes" in this argument in terms of degree of belief (b), disbelief (d), and uncertainty (u). Belief, disbelief, and uncertainty in claims depending on that evidence is then calculated using the beta distribution. For example, the authors relate Jøsang's *consensus operator* for opinions $\pi_A = b_A, d_A, u_A$ and $\pi_B = b_B, d_B, u_B$ in cases where $u_A + u_B - u_A u_B \neq 0$ [54]:

$$b_{A,B} = \frac{b_A u_B + b_B u_A}{u_A + u_B - u_A u_B} \tag{D1}$$

$$d_{A,B} = \frac{d_A u_B + d_B u_A}{u_A + u_B - u_A u_B} \tag{D2}$$

$$u_{A,B} = \frac{u_A u_B}{u_A + u_B - u_A u_B} \tag{D3}$$

Illustrative Example. The paper provides an illustrative example of a partial assurance argument for an airport backscatter x-ray machine [19]. Figure D1 depicts this argument. Table D1 gives the example opinions about the evidence cited by the argument. The example models the joint support of Sn1 and Sn2 for G2 and of Sn3 and Sn4 for G3 using the consensus operator. It models the joint support of G2 and G3 for G1 using an unspecified *logical or* operator. The calculated values are then mapped onto opinions using Jøsang's opinion triangle [42].

Test for Argument Sufficiency. The paper does not specify how to use calculated opinions to make a decision about whether an assurance argument justifies putting a system into service.

D.2 Replication

The paper does not specify how to convert between opinions and beta distributions save that this should be done "based on Jøsang's work" [19]. Jøsang's recent work (e.g., [42,55]) uses the following equations, in which a represents *atomicity*:

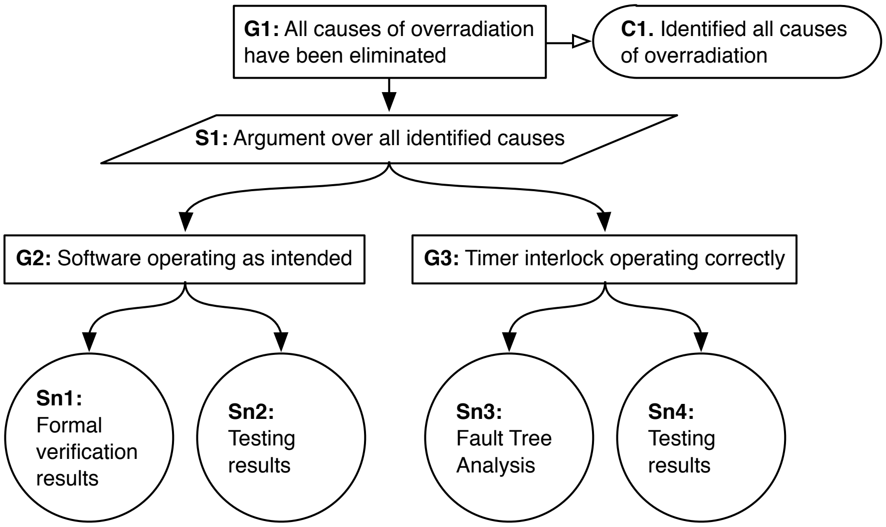

G1: All causes of overradiation have been eliminated
C1. Identified all causes of overradiation
S1: Argument over all identified causes
G2: Software operating as intended
G3: Timer interlock operating correctly
Sn1: Formal verification results
Sn2: Testing results
Sn3: Fault Tree Analysis
Sn4: Testing results

Figure D1: *An example GSN assurance argument about an airport x-ray machine taken from [19].*

Element	Opinion values	Beta parameters	
Sn1	$\pi(0.70, 0.20, 0.10)$	$\alpha = 8.00$	$\beta = 2.00$
Sn2	$\pi(0.50, 0.20, 0.30)$	$\alpha = 2.67$	$\beta = 1.67$
Sn3	$\pi(0.30, 0.50, 0.20)$	$\alpha = 2.50$	$\beta = 3.50$
Sn4	$\pi(0.90, 0.05, 0.05)$	$\alpha = 19.00$	$\beta = 2.00$
G1	Not given	$\alpha = 40.29$	$\beta = 2.38$
G2	Not given	$\alpha = 9.75$	$\beta = 3.75$
G3	Not given	$\alpha = 21.25$	$\beta = 4.75$

Table D1: *Opinions about elements shown in Figure D1, taken from [19]. Opinions are given as $\pi(b, d, u)$ where b represents the degree of belief, d the degree of disbelief, and u the uncertainty, all on the scale $[0, 1]$.*

$$\alpha = r + 2a \tag{D4}$$

$$\beta = s + 2\,(1 - a) \tag{D5}$$

$$r = 2b_x/u_x \tag{D6}$$

$$s = 2d_x/u_x \tag{D7}$$

$$1 = b_x + d_x + u_x \tag{D8}$$

Combining Equations D4–D8 yields:

$$\alpha = \left(\frac{2b_x}{u_x}\right) + 2a \tag{D9}$$

$$\beta = \left(\frac{2d_x}{u_x}\right) + 2\,(1 - a) \tag{D10}$$

But these equations are not compatible with the opinions and beta parameters shown in Table D1. For example, applying Equations D9 and D10 yield $\alpha_{\mathsf{Sn1}} = 15.0$ and $\beta_{\mathsf{Sn1}} = 5.00$. Personal communication with the authors revealed that the paper uses an earlier version of Jøsang's equations [54, 56]:

$$\alpha = r + 1 \tag{D11}$$

$$\beta = s + 1 \tag{D12}$$

$$r_p = b_p / i_p \tag{D13}$$

$$s_p = d_p / i_p \tag{D14}$$

$$1 = b + d + i \tag{D15}$$

In this earlier model—and in contrast with the paper—Jøsang uses i (ignorance) in place of u (uncertainty) and omits a (atomicity) entirely. The authors use the earlier versions of these equations because these were the first versions they encountered rather than to obtain some perceived benefit [56]. In any case, combining Equations D11–D15 yields

$$\alpha = \frac{b}{i} + 1 \tag{D16}$$

$$\beta = \frac{d}{i} + 1 \tag{D17}$$

Using these equations, we arrive at the beta distributions shown in Table D1, save for β_{Sn1}, which the authors agree should be 3.0, not 2.0 [56]. While the authors do not present a logical-or operator, Jøsang defines such an operator as [42]

$$b_{x \vee y} = b_x + b_y - b_x b_y \tag{D18}$$

$$d_{x \vee y} = d_x d_y \tag{D19}$$

$$u_{x \vee y} = d_x u_y + u_x d_y + u_x u_y \tag{D20}$$

Using Equations D1–D3 and D11–D20, substituting in the erroneous value of β_{Sn1}, and rounding the figures after calculating each conclusion as the authors did [56], we duplicated the results shown in Table D1.

D.3 Hypotheses and Evidence

The paper hypothesizes that "the use of a distribution to represent confidence in assurance cases makes intuitive sense and . . . the beta distribution is the most appropriate one to use" [19]. It supports this hypothesis by arguing that the beta distribution is "more versatile" and "allows for a better representation of human opinion" than competing distributions such as the doubly-truncated normal distribution used in some BBN tools. This is because "there are a couple of shapes that the truncated normal cannot approximate, such as a curve where the mean and mode are not equal (a skewed Gaussian), a true uniform distribution, or a true power distribution." The paper presents no evidence that it is necessary to approximate those distributions to produce a trustworthy figure for confidence in any assurance argument that might arise in practice.

The paper concludes that the proposed "novel use of subjective logic and the beta distribution to represent confidence will be of great benefit to assurance case evaluation and review." This is because the proposed technique offers "a more intuitive way to capture the confidence and uncertainty associated with evidence and compute the confidence and uncertainty associated with claims relying on that evidence" than confidence arguments [7], BBN-based approaches [28, 53], and Dempster–Shafer theory [18]. The paper offers no evidence that the result is intuitive to the intended safety case readership. More importantly, it offers no evidence to support the implied claim that the proposed technique *is* a way to (correctly) assess confidence and uncertainty in safety claims.

D.4 Counterargument

The proposed technique obscures the impact of conflicting evidence and incorrectly ascribes confidence to items of evidence. Moreover, the original example fails to account for confidence in an important property of evidence because the argument asserts that property as context.

D.4.1 Conflicting Evidence

One obvious question to ask is what the proposed technique makes of disconfirming evidence. Here, we focus on the software leg of Figure D1, i.e. G2, Sn1, and Sn2. Table D1 gives the opinions about Sn1 and Sn2 from the paper, which are repeated in Table D2. From these, we calculate the belief, disbelief, and uncertainty in G2 using Equations D1–D3. Table D2 shows the results. To demonstrate the consensus operator's effect on disconfirming evidence, we define the *Conflict* and *More Evidence* variants of the paper's original example. We use these examples to show how confirming evidence overwhelms disconfirming evidence in cases where it should not.

Conflict Example. In the Conflict example, we posit complete source-code-to-specification refinement proofs and a qualified complier. This is stronger than the static analysis used in the original example, but the proof tools might still have undiscovered defects or be configured or used improperly. We further suppose that a unit test run in a simulated environment fails, resulting in strong but imperfect disconfirming evidence. As shown in Table D2, we assign 95% belief to the static analysis evidence and 95% disbelief to the testing evidence. The result is an even 49%–49% balance of belief and disbelief in the software correctness claim (G2) and 90% belief in the correctness of the system safety claim (G1).

Node	Original version			Conflict			More Evidence		
	b	d	u	b	d	u	b	d	u
Ev1	0.70	0.20	0.10	0.95	0.00	0.05	0.95	0.00	0.05
Ev2	0.50	0.20	0.30	0.00	0.95	0.05	0.00	0.95	0.05
Ev5	Not applicable						0.60	0.00	0.40
G2	0.70	0.22	0.08	0.49	0.49	0.03	0.51	0.47	0.02
G1	0.94	0.03	0.02	0.90	0.07	0.03	0.91	0.07	0.02

Table D2: *Computed confidence in a software claim for three variants of the argument shown in Figure D1: the authors' original and our Conflict and More Evidence variants.*

More Evidence Example. In the More Evidence example, we add a third piece of evidence, namely source code inspections. This third piece of evidence is much weaker—60% belief, 40% uncertainty—but reveals no problem with the source code. Table D2 shows the results. In the More Evidence alternative, the weak third item of evidence tips the balance in favor of the software behavior claim: 51% belief in the software correctness claim and 91% belief in the system safety claim.

Analysis of Examples. It is possible (and perhaps likely) that the symptoms illustrated by the Conflict and More Evidence scenarios reflect a problem that could be revealed by testing but not by static analysis or source code inspection. For example, a compiler option might have been set incorrectly, resulting in an executable binary that is unsafe to use. Engineers should investigate failing test cases to determine the cause and its impact on safety. Quantifying confidence using this technique might be dangerous if it is used to justify forgoing such investigation.

D.4.2 Confidence in Evidence

Analysts using the proposed technique assess evidence in the abstract. But a thing is evidence only by virtue of having been cited in support of a claim and might lend more support to some claims than to others [47]. The opinions in Table D1 make sense only if applied to evidence–claim pairs, e.g. the analyst's opinion of Sn1's support for G2 is $\pi(0.7, 0.2, 0.1)$.

D.4.3 Confidence in the Hazard Analysis

The original example asserts the completeness of the hazard analysis in context element C1 in Figure D1. In GSN, propositions should appear as goals, not context [3, 46]. Because the completeness of the hazard analysis appears as context, confidence in completeness is excluded from the calculated confidence and uncertainty in the main safety claim G1 despite the obvious impact that it would have on safety.

Appendix E

Guiochet, Hoang, and Kaâniche

Guiochet, Hoang, and Kaâniche propose a technique for assessing confidence in an existing assurance case using Dempster–Shafer theory [29].

E.1 Proposed Technique.

The proposed technique models confidence in a conclusion as the strength of belief that the conclusion is true (Equation E1), defines uncertainty as the lack of belief that the conclusion is either true or false (Equation E2), and rules out disbelief (which presumably should not be included in the argument):

$$m(A) = Bel(A) = g(A) \in [0, 1] \tag{E1}$$

$$m(A, \bar{A}) = 1 - g(A) \in [0, 1] \tag{E2}$$

$$m(\bar{A}) = 0 \tag{E3}$$

Conclusion Supported by a Single Premise. The paper defines confidence in a goal A supported by a solution or goal B as

$$g(A) = p * g(B) \tag{E4}$$

where p represents confidence in the inference. The paper does not specify how to obtain values of p.

Alternative Arguments. The paper defines *alternative arguments* as arguments where either of two premises B or C support conclusion A. The technique models these using the *noisy-or* function:

$$g(A) = p * g(B) + q * g(C) - g(B) * g(C) * p * q \tag{E5}$$

The paper does not specify how to obtain the weighting factors p and q.

Complementary Arguments. The paper defines arguments where “a set of solutions or subgoals are required simultaneously for supporting the main goal” as *complementary arguments.* The proposed technique models confidence in complementary arguments using a variant of the *leaky-noisy-and* function. The paper defines this function in terms of the truth table given in Table E1 and the following equation for v:

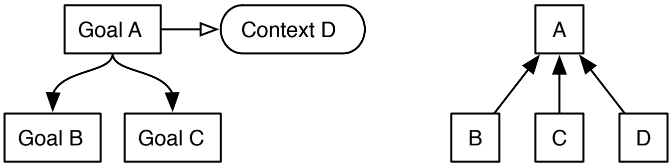
$$v = \frac{p + q}{2} \tag{E6}$$

The paper then generalizes to a conclusion X based on n premises:

$$v = \frac{1}{n} \sum_{i=1}^{n} p_i \tag{E7}$$

$g(B)$	1		0	
$g(C)$	1	0	1	0
$g(A)$	v	$v.(1-q)$	$v.(1-p)$	0

Table E1: *Truth table for conclusion C supported through complementary argument by premises B and C [29]. The value of v is given by Equation E6.*



(a) *GSN argument* **(b)** *BBN model*

Figure E1: *Representing a complementary argument with two sub-goals and a context element (a) as a BBN (b). Taken from [29].*

$$g(X|\bar{Y}_1,\ldots,\bar{Y}_k) = v.\prod_{i=1}^{k}(1-pi) \tag{E8}$$

For complementary arguments with context, the paper proposes modeling the conclusion as the noisy-and of the two premises and the context as shown in Figure E1. The paper does not specify how to obtain the needed weighting parameters.

Test for Argument Sufficiency. The paper does not specify how to use the computed likelihood of a main safety claim to decide whether to release a safety-critical system into service.

E.2 Replication

The paper does not present a worked example that we could replicate.

E.3 Hypotheses and Evidence

The paper asserts that "providing quantitative estimation of confidence is an interesting approach to manage complexity of arguments" [29]. This implies that a benefit of quantification is reduced need for careful reading of the argument. (The act of computing confidence does not change the argument structure. To reduce the difficulty of assessing complex arguments, the task of quantitative confidence assessment must replace a more difficult task. Presumably, that task is reading and understanding the argument.) But arguments have purposes other than supporting release-to-service decisions. For example, safety arguments communicate the system safety rationale to those who, years on, will maintain or modify the system.

The paper notes that one of the central features of the technique, “the constraint $m(X, \bar{X}) = 0$, brings the main benefit of letting [analysts] use mathematical tools, such as BBN.” This constraint might facilitate tool reuse, but the paper presents no evidence that the resulting computations are useful. The paper refers to “an experiment on a real case study of a rehabilitation robot,” citing a Ph.D. thesis available only in French, but do not relate the study method or results or connect these to a specific claim about the technique. Instead, the authors explicitly limit their conclusions to computational feasibility: “in this paper, we focus only on the feasibility of a quantitative estimation of confidence.” But a feasible technique must also be accurate if it is to be fit for use.

E.4 Counterargument

Since the paper does not present a worked example that we could replicate, we could not attempt to falsify an efficacy hypothesis by finding plausible alternatives that yield implausible results. But we note two problems with the proposed technique in addition to the usual problem of obtaining the required probability factors, weights, and leak factors.

Incompleteness and Inconsistency. The proposed combination rule for complementary arguments is both incomplete and inconsistent. The paper does not define the result when belief in the premises is neither 0 nor 1. Furthermore, Equation E8 is incompatible with the truth table given in Table E1. We assume that the authors meant n rather than k and p_i rather than pi. But even so, Equation E8 does not produce the value for two false premises shown in Table E1 unless $p_i = 1$ for some i.

Context is Not a Proposition. The proposed technique models the impact of GSN context elements as though they were propositions, but they are not. While the precise meaning and function of GSN context elements are not clear [46], the notion of belief in the truth or falsehood of context is incompatible with example context elements given in the GSN standard [3]. For example, it is meaningless to speak of the degree of belief in the truth of the phrases “Operating Role and Context,” “Control System Definition,” or “SIL Guidelines and Processes” because none of these are propositions.

Appendix F

Guo

Guo proposes representing safety arguments based on conformance to standards as Bayesian Belief Networks (BBNs) [30].

F.1 Proposed Technique

The paper does not define a specific structure for the networks or recommend a specific formula or method for populating node probability tables beyond the recommendation to "gathe[r] data from objective and subjective sources, such as the historical dependability of similar system, the competence of the developer, the methods and process deployed." The paper does not specify how to use the computed probability of a safety claim to decide whether to release a safety-critical system into service. However, it does provide an example BBN for calculating the quality of a safety requirements specification. Figure F1 presents that example.

F.2 Replication

The paper does not provide node probabilities for the example given in Figure F1. We were thus unable to replicate the example.

F.3 Hypotheses and Evidence

The paper reasons that because "BBNs are based on Bayes' theory and have a sound mathematical background," they "provide a quantitative figure for evaluating safety from probabilistic meaning." This reasoning is too vague to accept: without knowing what properties are meant and how these relate to the technique's fitness for use, a reader cannot assess either the premise that BBNs have these properties or that having these properties makes the technique fit for use. The paper continues,

> The graphic representation of BBNs is intuitive in nature . . . . By using BBNs, safety assessment hence becomes more intuitive and understandable, thereby more efficiently and effectively.

The paper neither specifies what form of representation (e.g., a prose text accomplishment summary, a prose text safety argument, or a graphical safety argument) this hypothesis compares BBNs to nor identifies evidence supporting this hypothesis. The authors conclude,

> We build an initial prototype BBN for systematic safety assessment to a certain phase of life cycle of IEC 61508. Some factors, such as staff, budget, complexity of problem, time

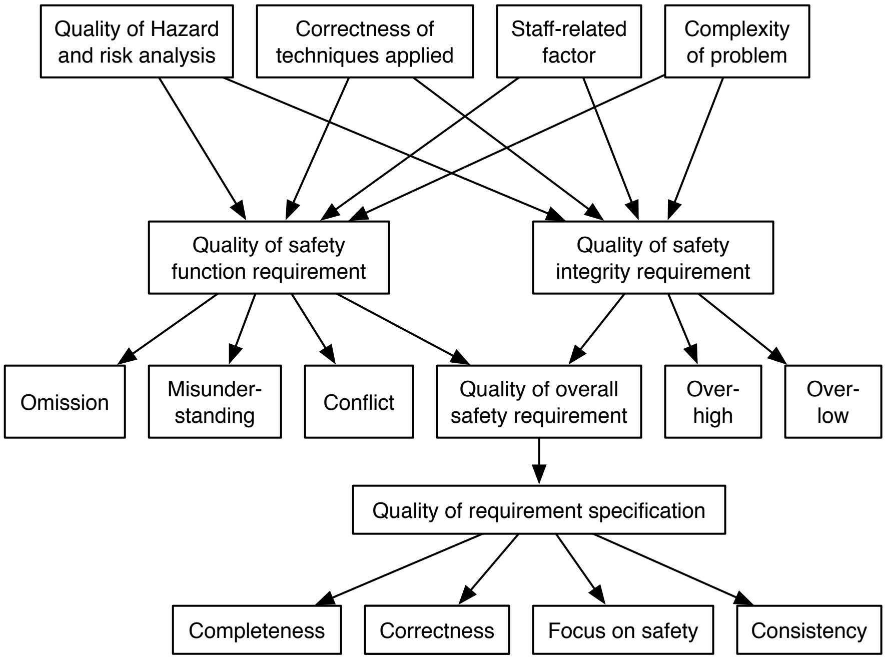

Quality of Hazard and risk analysis
Correctness of techniques applied
Staff-related factor
Complexity of problem
Quality of safety function requirement
Quality of safety integrity requirement
Omission
Misunder-standing
Conflict
Quality of overall safety requirement
Over-high
Over-low
Quality of requirement specification
Completeness
Correctness
Focus on safety
Consistency

Figure F1: *"The initial BBN topology for systematically assessing overall safety requirement" from [30]. For consistency with other BBNs in this paper, we represent nodes as rectangles rather than ovals.*

> requirement, and development hardware environment, have been carefully thought about in this BBN prototype. The novel points of this BBN include improved transparency, objective and quantitative assessment, changing purely subjective views about conformance and etc.

Again, it is not clear what the paper is comparing BBNs against. In any case, the paper presents no empirical evidence to show that a BBN-based quantitative assessment of a safety case provides a trustworthy basis for deciding whether to release a system into service.

F.4 Counterargument

We could not create a counterexample similar to the BBN shown in Figure F1 because that example is itself implausible for two reasons. First, it is meaningless to describe the "quality of [a] safety integrity requirement" as "over-high." Second, the difference between the Quality of overall safety requirement and Quality of requirement specification nodes is unclear. A requirement specification is not its requirements, much less a single requirement, and might contain requirements that are not safety-related. But if the latter node was meant to represent these broader concerns, it would depend on nodes representing them. It does not.

Appendix G

Hobbs and Lloyd

Hobbs and Lloyd propose representing assurance arguments as Bayesian Belief Networks (BBNs) [20].

G.1 Proposed Technique

The paper proposes representing an assurance cases as "a BBN where each leaf ... represents elementary evidence and the network's structure represents the argument." The authors refer to the SafEty and Risk Evaluation using bayesian NEts (SERENE) method manual [23] for its patterns but suggest that the *noisy-or* and *noisy-and* functions might model assurance case reasoning better than simple *or* and *and* functions.

The Noisy-Or Function. The paper defines noisy-or as

$$p(X|Y_1,\dots,Y_N) = 1-(1-k)\prod_{i:Y_i}(1-p_i) \tag{G1}$$

where X is the child (conclusion) node, $Y_1 \dots Y_N$ are the N parent (premise) nodes, and k is the level of confidence in X when all of the parent nodes are false ($k = p(\bar{Y}_1\ \dots, \bar{Y}_N)$).

The Noisy-And Function. The paper defines noisy-and as

$$p(X|Y_1,\dots,Y_N) = (1-k)\prod_{i:Y_i}(1-p_i) \tag{G2}$$

where k is a leakage factor such that $0.0 \leq k \leq 1.0$. It does not identify a means for analysts to obtain the leakage factors and weights that the noisy-or and noisy-and functions require.

Test for Argument Sufficiency. The paper does not specify what probability of truth in a safety conclusion should be taken as sufficient to justify releasing a system into service. But it does note that

> when the Bayesian calculation is first performed, it is probable that the final result will not be to the liking of the analyst: the level of assurance may be too low. In this case a sensitivity calculation may be performed to find where the 'biggest bang can be achieved for the buck'.

Illustrative Example. The paper provides an example BBN to illustrate the proposed technique. Figure G1 shows this example. Table G1 and Table G2 give the example node probabilities.

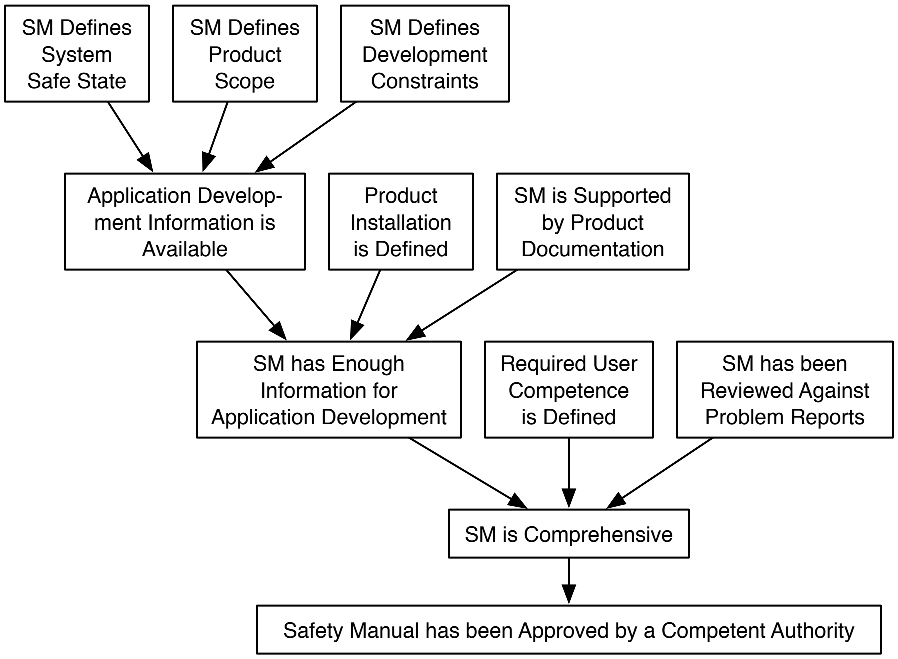

SM Defines System Safe State
SM Defines Product Scope
SM Defines Development Constraints
Application Development Information is Available
Product Installation is Defined
SM is Supported by Product Documentation
SM has Enough Information for Application Development
Required User Competence is Defined
SM has been Reviewed Against Problem Reports
SM is Comprehensive
Safety Manual has been Approved by a Competent Authority

Figure G1: *An example BBN taken from [20]. For compactness and consistency with other BBNs in this paper, we represent nodes as rectangles. The abbreviation SM stands for safety manual.*

G.2 Hypotheses and Evidence

The paper hypothesizes that a benefit of using BBNs to show how evidence substantiates dependability claims is that it "provides fewer opportunities for flaws in the argument or inappropriate reliance on untrustworthy evidence." But it neither specifies what this hypothesis compares BBNs to—e.g., prose text arguments or arguments rendered in Goal Structuring Notation (GSN) [3] or Claims-Argument-Evidence notation [57]—nor provides supporting evidence. The paper concludes,

> Preparing an assurance case for a complex system, particularly one containing sophisticated software, by means of a BBN allows activities to be separated cleanly: 'how should we argue about the evidence?' and 'what evidence should we present?'. It prevents unstructured 'dumping' of material into a document by forcing the role of each piece of material to be considered individually. We have found that auditors to whom we have presented these cases have welcomed the structured argument. In summary, applying a BBN to represent an assurance case, when backed by a suitable computation tool, appears to be a flexible and powerful technique.

This efficacy hypothesis is unclear. Would auditors find a BBN more or less accessible than, for example, an argument presented in GSN? Does

Node	**False**	**True**
SM has been reviewed against problem reports	0.2	0.8
Required user competence is defined	0.2	0.8
SM has been approved by competent authority	0.0	1.0
SM is supported by product documentation	0.1	0.9
Product installation is defined	0.6	0.4
SM defines development constraints	0.2	0.8
SM defines system safe state	0.0	1.0
SM defines product scope	0.2	0.8
Application development information is available	See Table G2	
SM has enough information for application development	Not given	
SM is comprehensive	Not given	

Table G1: *Example node probability inputs for the BBN shown in Figure G1, taken from [20].*

SM defines development constraints	**SM defines product scope**	**SM defines system safe state**	**False**	**True**
False	False	False	0.952	0.048
False	False	True	0.760	0.240
False	True	False	0.920	0.080
False	True	True	0.600	0.400
True	False	False	0.880	0.120
True	False	True	0.400	0.600
True	True	False	0.800	0.200
True	True	True	0.000	1.000

Table G2: *Example truth table for Application development information is available in the BBN shown in Figure G1, taken from [20].*

being 'powerful' mean that the conclusion reached by BBN analysis is sufficiently trustworthy to use as the basis for a decision to release a system to service? In any case, the paper presents no evidence of efficacy other the the authors' anecdotal experience.

G.3 Replication

The paper provides observations for all leaf nodes and a conditional probability table for the Application Development Information is Available node. But it does not provide conditional probability tables for the other non-leaf nodes. While we could not replicate the complete example, we were able to implement the Application development information is available and supporting nodes in a BBN tool. The result is that Application development information is available is 19.04% false and 80.96% true.

G.4 Counterargument

We do not understand the meanings of the propositions each node represents well enough to formulate plausible alternatives to the original example. For example, we do not know how true Application development information is available must be if development is to proceed. The example permits this to be up to 20% true even if the safety manual providing that information does not define a safe state. It can be up to 60% true if the manual does not define product scope. These difficulties preclude generating an example that is close enough to the original to clearly be an application of the proposed technique.

Appendix H

Nair, Walkinshaw, Kelly, and de la Vara

Nair, Walkinshaw, Kelly, and de la Vara propose to use Evidential Reasoning (ER) to calculate belief in a safety claim [21,22,31].

H.1 Proposed Technique

One paper describes an approach based on expert assessment of the solution elements in Goal Structuring Notation (GSN) arguments [21]:

> Through a series of generic and specific questions about the solution, the expert will set out their assessment (ranging from a scale of 0–5) and their confidence (a quantified value of confidence level e.g., in percentage) in the satisfaction of the claim. ER will then propagate these beliefs through the GSN structure to yield an overall assessment of the system.

That paper leaves both further development and assessment and of the technique to future work.

A later paper describes both (a) a pattern for representing the portion of a confidence argument [7] related to a single evidence assertion in the safety argument and (b) a tool for instantiating that pattern and performing the ER calculations needed to compute overall confidence [22]. Figure H1 and Figure H2 present part of that pattern. The paper refers to a paper on Evidential Reasoning for definitions of the equations used to combine beliefs [39].

Representing Belief using ER. In the work the authors' papers reference, Evidential Reasoning defines a set of N *evaluation grades* $H = \{H_1, H_2, \dots H_N\}$ [39]. An *attribute* y is defined in terms of "L basic attributes $e_i\,(i = 1, \dots, L)$." The "*weights* of the attributes are given by $\omega = \{\omega_1 \omega_2 \dots \omega_i \dots \omega_L\}$ where ω_i is the relative weight of the ith basic attribute (e_i) with $0 \leq \omega_i \leq 1$." An *assessment* "$e_i\,(i = 1, \dots, L)$ of an alternative may be ... represented as

$$S(e_i) = \{(H_n, \beta_{n,i}), n = 1, \dots, N\} \qquad i = 1, \dots, L \tag{H1}$$

where $\beta_{n,i} \geq 0$, $\sum_{n=1}^{N} \beta_{n,i} \leq 1$, and $\beta_{n,i}$ denotes a degree of belief." The probability mass associated with the belief in attribute e_i is split into $N+2$ categories. The first N parts of the mass, $m_{n,i}$, each represent one of the evaluation grades. Another part, $\bar{m}_{H,i}$ represents the "part of the remaining probability mass that is not yet assigned to individual grades due to the fact that attribute i (denoted by e_i) only plays one part in the assessment relative to its weight." The final part, $\tilde{m}_{H,i}$, represents

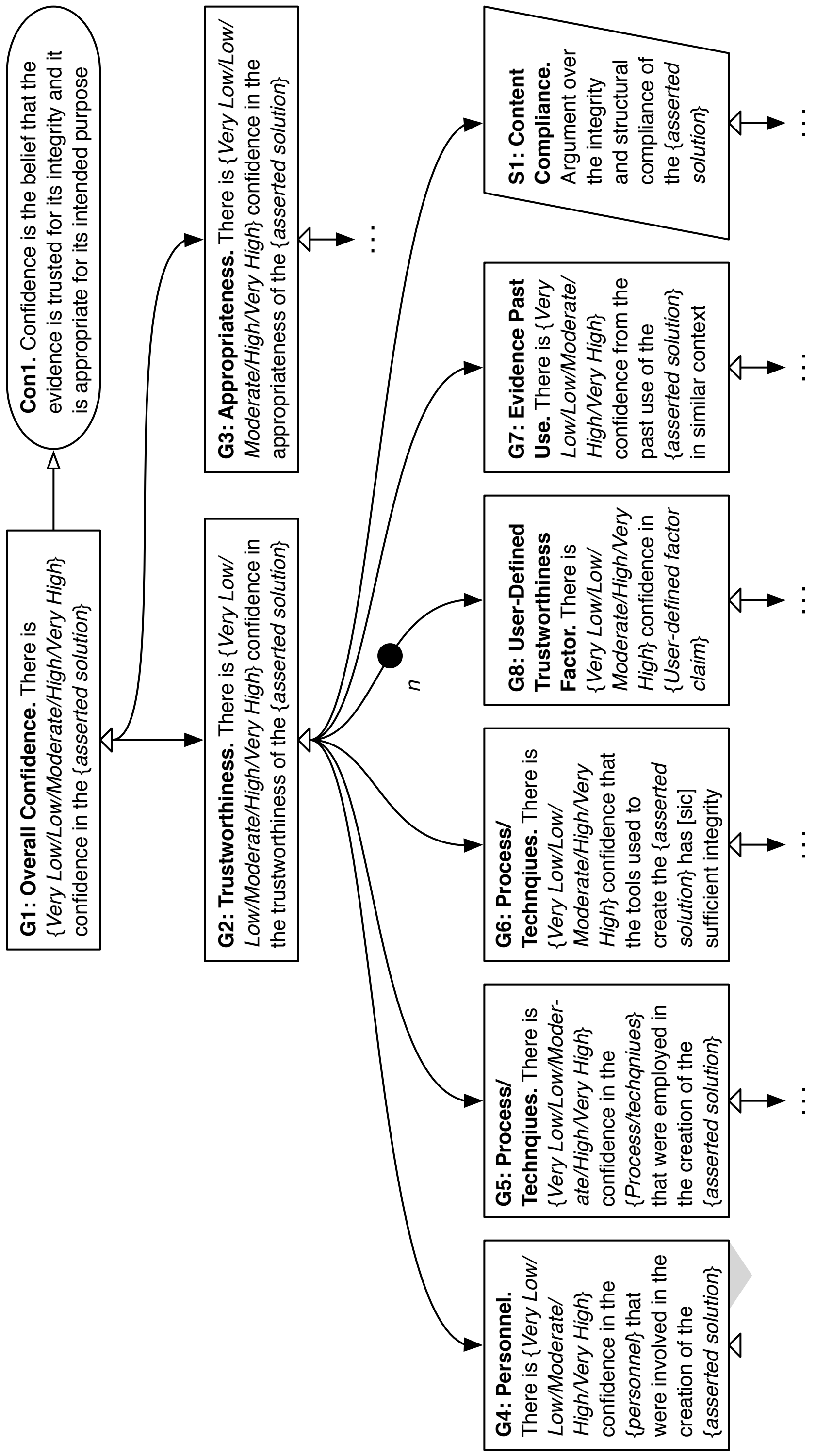

Con1. Confidence is the belief that the evidence is trusted for its integrity and it is appropriate for its intended purpose
G1: Overall Confidence. There is {Very Low/Low/Moderate/High/Very High} confidence in the {asserted solution}
G2: Trustworthiness. There is {Very Low/Low/Moderate/High/Very High} confidence in the trustworthiness of the {asserted solution}
G3: Appropriateness. There is {Very Low/Low/Moderate/High/Very High} confidence in the appropriateness of the {asserted solution}
...
n
G4: Personnel. There is {Very Low/Low/Moderate/High/Very High} confidence in the {personnel} that were involved in the creation of the {asserted solution}
G5: Process/Technqiues. There is {Very Low/Low/Moderate/High/Very High} confidence in the {Process/techqniues} that were employed in the creation of the {asserted solution}
G6: Process/Technqiues. There is {Very Low/Low/Moderate/High/Very High} confidence that the tools used to create the {asserted solution} has [sic] sufficient integrity
G8: User-Defined Trustworthiness Factor. There is {Very Low/Low/Moderate/High/Very High} confidence in {User-defined factor claim}
G7: Evidence Past Use. There is {Very Low/Low/Moderate/High/Very High} confidence from the past use of the {asserted solution} in similar context
S1: Content Compliance. Argument over the integrity and structural compliance of the {asserted solution}
...
...
...
...
...

Figure H1: *The proposed GSN pattern for confidence in a GSN solution solving a goal [22]. The complete pattern contains thirty-six mandatory, four optional, and four repeat-as-necessary elements. Instantiating the pattern and citing evidence requires a minimum of fifty-one GSN elements.*

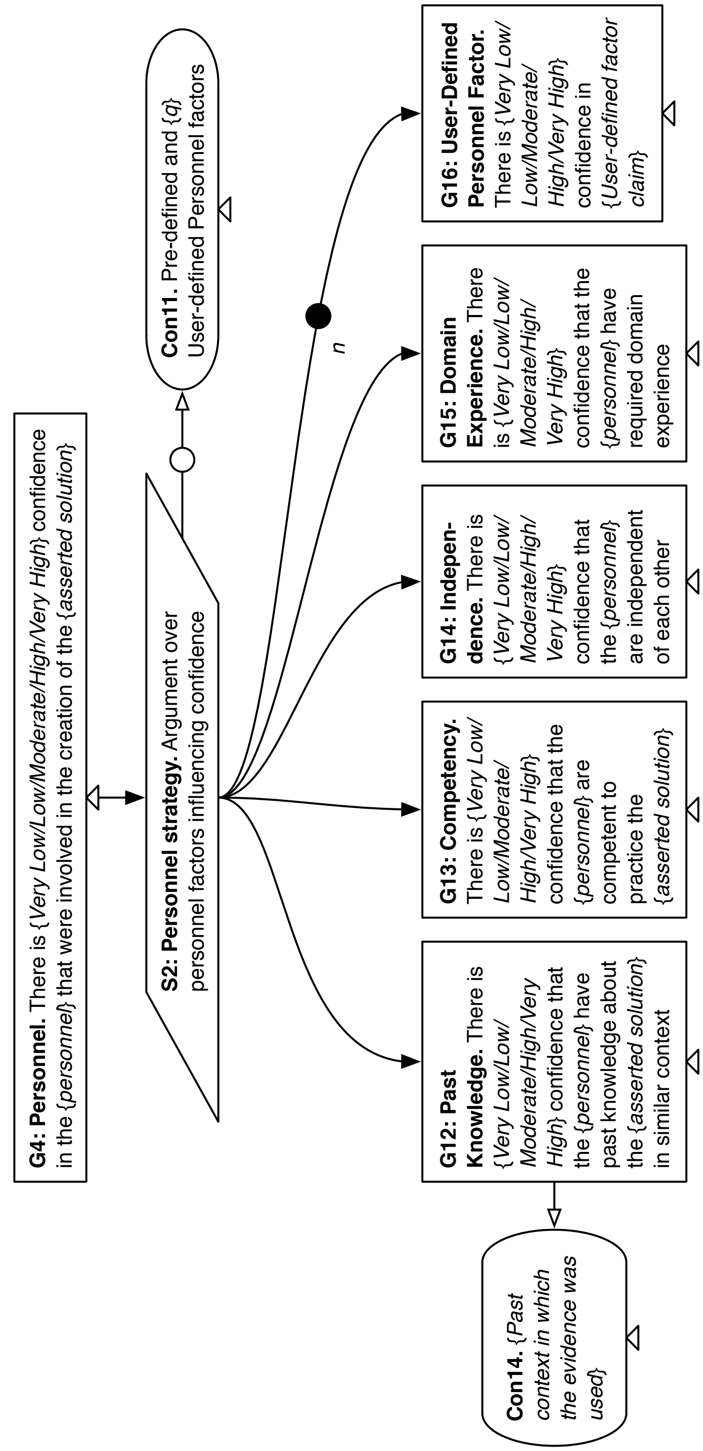

G4: Personnel. There is {Very Low/Low/Moderate/High/Very High} confidence in the {personnel} that were involved in the creation of the {asserted solution}
S2: Personnel strategy. Argument over personnel factors influencing confidence
Con11. Pre-defined and {q} User-defined Personnel factors
n
G12: Past Knowledge. There is {Very Low/Low/Moderate/High/Very High} confidence that the {personnel} have past knowledge about the {asserted solution} in similar context
G13: Competency. There is {Very Low/Low/Moderate/High/Very High} confidence that the {personnel} are competent to practice the {asserted solution}
G14: Independence. There is {Very Low/Low/Moderate/High/Very High} confidence that the {personnel} are independent of each other
G15: Domain Experience. There is {Very Low/Low/Moderate/High/Very High} confidence that the {personnel} have required domain experience
G16: User-Defined Personnel Factor. There is {Very Low/Low/Moderate/High/Very High} confidence in {User-defined factor claim}
Con14. {Past context in which the evidence was used}

Figure H2: *The pattern begun in Figure H1, continued. For brevity, we do not present the entirety of the pattern.*

the "part of the remaining probability mass unassigned to individual grades, which is caused due to the incompleteness in the assessment." The probability mass can be calculated from beliefs and weights:

$$m_{n,i} = \omega_i \beta_{n,i} \qquad n = 1, \ldots, N \tag{H2}$$

$$\bar{m}_{H,i} = 1 - \omega_i \tag{H3}$$

$$\tilde{m}_{H,i} = \omega_i \left(1 - \sum_{n=1}^{N} \beta_{n,i}\right) \tag{H4}$$

Combining Belief Masses Using ER. The probability masses corresponding to the assessments $S(e_i)$ and $S(e_j)$ for attributes e_i and e_j can be combined into a joint mass:

$$m_{n,i\oplus j} = K_{i\oplus j} \left(m_{n,i} m_{n,j} + m_{H,i} m_{n,j} + m_{n,i} m_{H,j}\right) \tag{H5}$$

$$\tilde{m}_{H,i\oplus j} = K_{i\oplus j} \left(\tilde{m}_{H,i} \tilde{m}_{H,j} + \bar{m}_{H,i} \tilde{m}_{H,j} + \tilde{m}_{H,i} \bar{m}_{H,j}\right) \tag{H6}$$

$$\bar{m}_{H,i\oplus j} = K_{i\oplus j} \bar{m}_{H,i} \bar{m}_{H,j} \tag{H7}$$

$$m_{H,i} = \bar{m}_{H,i} + \tilde{m}_{H,i} \tag{H8}$$

$$K_{i\oplus j} = \left(1 - \sum_{t=1}^{N} \sum_{\substack{l = 1 \\ l \neq t}}^{N} m_{t,i} m_{l,j}\right)^{-1} \tag{H9}$$

When there are more than two basic attributes, analysts apply Equations H5–H9 recursively, first calculating the joint mass for attributes one and two, then applying the equations again to join the resulting joint mass with the mass for attribute three, and so on. After obtaining the joint mass for y, an analyst can calculate belief in y using the following:

$$\beta_{n,y} = \frac{m_{n,y}}{1 - \bar{m}_{H,y}} \tag{H10}$$

$$\beta_{H,y} = \frac{\tilde{m}_{H,y}}{1 - \bar{m}_{H,y}} \tag{H11}$$

Representing Confidence in Safety Arguments. The authors instantiate evaluation grades for safety arguments as

H_1	Very low
H_2	Low
H_3	Moderate
H_4	High
H_5	Very high

In the proposed use of Evidential Reasoning, attribute y is a GSN goal and the basic attributes are supporting goals or evidence [22]. The papers do not specify how to obtain the necessary weights.

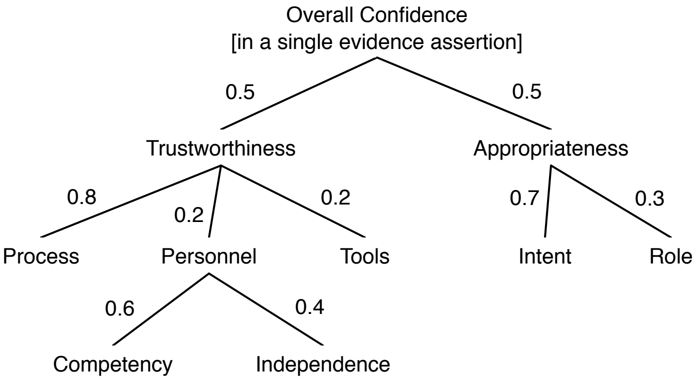


Figure H3: *Example weights taken from [22]. We note that this example does not directly correspond to the proposed confidence argument pattern (Figure H1).*

Question	Answer	Confidence	Belief distribution				
			β_1	β_2	β_3	β_4	β_5
Q1	Yes	80%	0.0	0.0	0.0	0.8	0.0
Q2	Absolutely	100%	0.0	0.0	0.0	0.0	1.0
Q3	Maybe	50%	0.0	0.0	0.5	0.0	0.0
Q4	Yes	80%	0.0	0.0	0.0	0.8	0.0

Table H1: *The assessments for the partial example given in the paper [22].*

Test for Argument Sufficiency. The paper does not specify how to use computed confidence to determine whether a system is safe enough to put into service.

Illustrative Example. One paper provides a partial example set of weights and an example calculation of trustworthiness in a hazard log [22]. Figure H3 presents the example weights, which do not correspond to either the confidence argument pattern shown in Figure H1 or the weights shown in Figure H3. The paper's presentation of the example calculation begins by defining four checklist questions that the argument assessor would answer:

Q1. "If independence is required, is the person doing the verification different than the one responsible for developing the hazard log?

Q2. "Is the manager to whom the team reports identified so that it can be confirmed that the requirements on independence are met?

Q3. "Are there records of attendance/participation in hazard identification workshops/exercises of the personnel that include the name, organisation and role?

Q4. "Is there information on the competency of the personnel?"

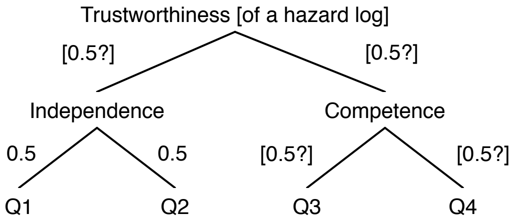


Figure H4: *Example weights for reasoning about the trustworthiness of a hazard log taken from [22]. The figures in brackets are not presented in the paper; we arrived at them by reverse engineering.*

These questions are answered using a different Likert scale:

H_1	Definitely not
H_2	No
H_3	Maybe
H_4	Yes
H_5	Absolutely

Table H1 gives the assessments for each question. Figure H4 gives the weights with which these are combined.

H.2 Replication

Using the figures and structure presented in Table H1 and Figure H4, we reproduced the calculations shown in the papers. We did not use the authors' experimental EviCA tool because the authors declined our request for access to it. Using spreadsheet software, we calculated the belief in the trustworthiness of the hazard log to be $\beta_1 = 0.000$, $\beta_2 = 0.000$, $\beta_3 = 0.099$, $\beta_4 = 0.452$, and $\beta_5 = 0.281$, with uncertainty $\beta_H = 0.168$. These figures match those in the paper. The paper gives the following direction for turning these belief figures into a qualitative description:

> To represent the assessor's confidence in the trustworthiness of the hazard log, we treat the final distribution of the assessment as a five point Likert-scale: *Very Low*, *Low*, *Medium*, *High*, and *Very High*. ... To best represent the assessor's confidence, in our approach we use the median of the distribution. In the above distribution, the median is 0.099, which corresponds to *Medium* in the Likert scale. ... To represent the assessor's confidence in their assessment, we quantify confidence from a scale of 0–100%, with intervals: 0–20% *Very Low*, 20–40% *Low*, 40–60% *Medium*, 60–80% *High*, and 80–100% *Very High*.

We note that following this procedure would always result in an assessment of *Medium* trustworthiness regardless of the calculations. It might be supposed that the authors meant the median of the non-zero elements of the distribution. But that interpretation would yield an assessment of *High*, not *Medium*, trustworthiness in this case.

H.3 Hypotheses and Evidence

One paper about the proposed technique points out that it is based on a theory with desirable properties, contends that tool support will make it feasible, and reports a survey-based assessment [22].

Properties of ER. One paper, citing Yang and Xu [39], reports that the Evidential Reasoning approach

> obeys certain desirable axioms that ensure the following:
>
> 1. If none of the basic attributes for y is assessed at a grade H_n, then $\beta_{n,y} = 0$.
> 2. If all of its basic attributes are assessed to a grade H_n then $\beta_{n,y} = 1$.
> 3. If all of the basic attributes are completely assessed to a subset of evaluation grades then y should be completely assessed to the same subset of grades.
> 4. If, for a basic attribute z, $\sum_{i=0}^{n} \beta_{i,z} < 1$, then the same holds for y: $\sum_{i=0}^{n} \beta_{i,y} < 1$ [22].

These properties, while desirable, are not sufficient to show that applying the proposed technique will result in a trustworthy assessment of confidence. Rules that would produce different assessments of confidence also satisfy these criteria. For example, one might define the combination y of assessments of premises a and b as the categorical weaker of the two,

$$\beta_{n,y} = \min(\beta_{n,a}, \beta_{n,b}) \tag{H12}$$

or the categorical mean of the two,

$$\beta_{n,y} = \frac{\beta_{n,a} + \beta_{n,b}}{2} \tag{H13}$$

Tool Support Makes the Approach Feasible. The proposed technique requires instantiating the large pattern shown in Figure H1 many times. The paper argues that tool support makes this practical by generating argument graphics from user input about assessments, context, user-specific confidence factors, etc. given through dialog boxes. But gathering and supplying that data might itself be a large expense. Moreover, the paper presents no evidence to show that resulting large argument would have practical value to any reader. If the pattern's purpose is to facilitate calculating the confidence in a claim rooted directly in evidence, and if confidence figures actually encapsulate confidence, why render this confidence argument fragment graphically at all?

Evaluation by Survey. The authors evaluated the proposed technique and the EviCA tool by surveying twenty-one safety experts that were personally selected by the authors or recruited through a social networking website (LinkedIn). The subjects viewed a presentation and then responded to survey questions. When asked whether "use of the approach will lead to more accurate safety evidence assessments," four strongly agreed, nine agreed, and eight neither agreed nor disagreed.

It is not clear that the reported survey provides much support for the hypothesis that the technique is a trustworthy basis for making release-to-service decisions. This is in part due to weaknesses that affect survey research generally: surveys are affected by *response biases*, including those related to *demand characteristics*. For example, survey respondents sometimes feel compelled to be 'good subjects' by giving responses they think support the researchers' hypothesis [44]. Researchers can limit response bias through measures such as deliberately hiding the hypothesis from subjects, but the papers describe no such controls.

The other main threat to validity is that respondents are not known to be capable of accurately assessing efficacy. Efficacy might depend on subtle matters of degree such as analysts' skill in choosing weights and making judgments. If experimental and observational assessments of efficacy and its contributory factors were available to respondents, their opinions might synthesize this evidence. Without such data, it is not clear that participants' opinions on efficacy are a measure of efficacy.

H.4 Counterargument

The example presented in the papers does not seem to be an application of the proposed technique as defined, but rather a demonstration of the ER equations the authors adopt from Yang and Xu. Nevertheless, a variant of the example serves as a counterexample. Suppose (a) that two safety engineers are known, by long history, to be hopelessly incompetent to practice, (b) that they work independently (and that a manager has documented this) and (c) that there is record of them attending training. The answers to questions Q1–Q4 above might plausibly be *absolutely, with very high confidence.* (The engineers' history of incompetence is "information on the competency of the personnel" even if it shows incompetence.) Using the proposed technique, one might conclude with very high confidence that the trustworthiness of the hazard log that one of the incompetent engineers compiles and the other verifies is very high. This is not plausible.

Appendix I

The SERENE Partners

The partners of the Safety and Risk Evaluation Using Bayesian Nets (SERENE) project propose using Bayesian Belief Networks (BBNs) to model system safety and compute risk [23].

I.1 Proposed Technique

The SERENE manual defines a number of *idioms* (patterns) that the analyst uses to model the system in question. The analyst populates the node probability tables—the manual does not specify where the analyst should get the needed data—and uses a BBN tool to perform the computations.

I.2 Replication

The SERENE manual defines and gives examples of several idioms. Figure I1 depicts the manual's instantiation of the *definition/synthesis* idiom for safety. The manual does not specify the function or conditional probability for the Safety node, but we presume that safety decreases as the product of failure frequency and severity increases. We were able to instantiate the model in a BBN tool using continuous interval nodes.

I.3 Hypotheses and Evidence

The SERENE manual asserts that the SERENE method provides:

1. "Improved communication of safety"
2. "Greater focus on the properties which lead to safety"
3. "A basis for empirical validation of the beliefs of safety experts and of the rationale for existing standards"

The manual does not specify the means of communication it compares the proposed technique to. But similar conclusions might be made about non-quantitative safety arguments in comparison to having no representation of what each system's safety most depends on. The manual further asserts that the use of BBNs provides two benefits:

1. "The uncertainty associated with the causes of safety can be included in the model and the uncertainty about the overall system safety is explicit"
2. "The safety achieved can be quantified, allowing alternative safety strategies to be compared"

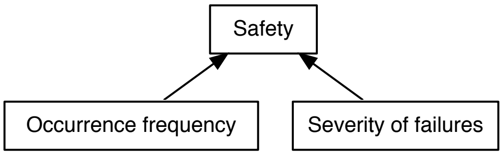


Figure I1: *A BBN from [23] instantiating the definition/synthesis idiom for safety. For consistency with Figure L1, we represent BBN nodes using rectangles.*

Variant	**Failure modes** ($<$ **Frequency**, **Severity** $>$)	**Risk**
Good	$\langle 10^{-9}, 10^{9}\rangle$, $\langle 10^{-9}, 10^{5}\rangle$, $\langle 10^{-7}, 10^{3}\rangle$, $\langle 10^{-3}, 10^{1}\rangle$	1×10^{1}
Bad	$\langle 10^{-9}, 10^{1}\rangle$, $\langle 10^{-9}, 10^{3}\rangle$, $\langle 10^{-7}, 10^{5}\rangle$, $\langle 10^{-3}, 10^{9}\rangle$	1×10^{6}

Table I1: *The failure modes and total risk of system variants to be modeled using the argument in Figure I1. Risk is computed using Equation I1.*

Again, the first conclusion could be made about any representation of uncertainty, including lists of defeaters [6, 7, 9]. The manual provides example BBNs but no empirical evidence that the computed results actually represent the "safety achieved."

I.4 Counterargument

Our counterexample comprises two variants of a system, *Good* and *Bad*. We define these so that they give rise to maximally distinct levels of risk yet are evaluated identically according the idiom shown in Figure I1.

The Good and Bad Examples. The Good and Bad examples each have four failure modes as shown in Table I1. Both use the same sets of four frequency values and four severity values, but in the Good system the most frequent failures are the least severe and vice-versa, while the opposite is true in the Bad system. The risk column gives the system risk as computed using a typical expected-value formulation:

$$\text{Risk} = \sum_{i:\{\text{Failure modes}\}} \text{Frequency}_i \times \text{Severity}_i \tag{I1}$$

Analysis of Examples. It is easy to see that the Good system is superior to Bad. The calculated risk bears this out. But both systems would be modeled identically by the BBN shown in Figure I1. Table I2 gives the node probabilities that result if we use the equation

$$\mathsf{Safety} = (\mathsf{Occurrence\ frequency} \times \mathsf{Severity\ of\ failures})^{-1} \tag{I2}$$

The technique calculates the same value of Safety for both the Good and Bad variants. This is not only implausible but dangerous: any

Node	**Distribution**				
Occurrence frequency	$[0, 10^{-8}]$	$, 10^{-6}]$	$, 10^{-4}]$	$, 10^{-2}]$	$, 1]$
	50%	25%	0%	25%	0%
Severity of failures	$[0, 10^{2}]$	$, 10^{4}]$	$, 10^{6}]$	$, 10^{8}]$	$, \infty)$
	25%	25%	25%	0%	25%
Safety	$[0, 10^{-8}]$	$, 10^{-6}]$	$, 10^{-4}]$	$, 10^{-2}]$	$, \infty)$
	0%	0%	25%	5%	70%

Table I2: *The node probabilities for the BBN shown in Figure I1 and the system variants defined in Table I1. We calculated Safety using the formula* $\textit{Safety} = (\textit{Occurrence frequency} \times \textit{Severity of failures})^{-1}$. *The probabilities for the Good and Bad variants are identical.*

decision procedure based on the idiom calculation that would accept the Good system as sufficiently safe would accept the Bad system also.

It might be argued that we should have used an equation other than Equation I2, that we have inappropriately combined information about multiple failure modes into the Occurrence frequency and Severity of failures nodes, or that the manual used the word ‘safety’ where it meant ‘risk.’ But this counterexample survives such objections. The problem that the counterexample reveals is the result of first reasoning separately about the frequency and severity of system failure and then reasoning about the resulting effect on safety or risk. To avoid this problem, the idiom could first reason separately about the risk contribution from each failure mode (thus pairing only related frequencies and severities) and then reason about the aggregate effect of those contributions.

Appendix J

Yamamoto

Yamamoto proposes annotating assurance arguments recorded in the Goal Structuring Notation (GSN) with attributes and using these to evaluate architectures "for quality claims, such as security and safety" [32].

J.1 Proposed Technique

In the proposed technique, analysts annotate GSN solution elements with attributes and then use rules to assess goals starting from the bottom of the goal structure and working upwards.

Scale for Annotations. The technique rates confidence on a five-point Likert scale:

-2	Strongly unsatisfied
-1	Unsatisfied
0	Unknown
1	Satisfied
2	Strongly satisfied

Assessing Goals Supported by a Solution. Goals supported by a single solution inherit the attribute of that solution. The paper does not specify how to combine attributes when a goal is supported by more than one solution.

Assessing Goals Supported by Subgoals. When a goal is supported through a strategy by k premises, the analyst annotates the strategy with a unit vector of weights for the premises. These weights are denoted $Q_1 \ldots Q_k$. The analyst then calculates the attribute P of the conclusion goal using the formula

$$P = \sum_{i=1,k} Q_i * R_i \tag{J1}$$

where $R_1 \ldots R_k$ are the attributes of the premises. (The paper defines P as $\sum_{i=1,k} Q_i * R_i * W_i$, where $\sum_{i=1,k} W_i = 1$, but does not define W_i or use it in its illustrative example [32]. Since W_i and Q_i are both unit weighting vectors, we eliminate W_i from our discussion for simplicity.)

The paper does not specify how to propagate attributes when goals directly solve other goals as is explicitly permitted in GSN [3].

Test for Argument Sufficiency. The paper does not specify how to use the computed values to decide whether an architecture is acceptable. Since the technique is proposed as a means of assessing architectures, not complete systems, it is not surprising that the paper also does not specify how to use computed values to decide whether a system is sufficiently safe to put into service.

Illustrative Example. The paper illustrates the proposed technique using an example assurance argument for a local-area network (LAN) device management system (LDMS) [32]. Figure J1 depicts this example. The figures in angle brackets are attributes (P) in the case of goals and solutions and weights ($Q_1 \ldots Q_k$) in the case of strategies.

J.2 Replication

Given the goal structure and attributes provided in the paper and reproduced in Figure J1, we were able to reproduce the author's calculations.

J.3 Hypotheses and Evidence

The paper describes its example as a "case study" but does not contain information usually found in case study reports such as descriptions of the study method, the data collected, and how these data relate to the questions at issue [58]. Nevertheless, it concludes that

> the case study on the LDMS was executed to evaluate the effectiveness of the GSN attribute method .... The result showed the propagation from evidences to the top claim in GSN is easy and traceable. This showed that the effectiveness of the attribute propagation method. Although the evaluation was only executed for one example, it is clear the easiness on the GSN attribute propagation can be derived for other applications. ... Discussions based on the case study showed the effectiveness and appropriateness of the proposed method to resolve security and safety issues simultaneously.

J.4 Counterargument

The paper's example argues that the LDMS is safe because each of its major components is safe and that a monitor component is safe because "abnormal events are defined." This is not plausible: it is not clear what it means for a component to be safe in isolation, the argument presents no evidence that component interactions do not produce unsafe behavior, and defining abnormal events is only the first step to detecting and mitigating them. Since any variants of this argument would be equally implausible, we used a different argument to assess the proposed technique. Figure J2 gives our specimen safety argument, which is for an unspecified system with 2–10 hazards. Our counterexample uses three variants of this argument: *Optimistic*, *Poor Evidence 1*, and *Poor Evidence 2*. These examples show two problems with the proposed technique: (a) it allows evidence for the mitigation of some hazards to mask

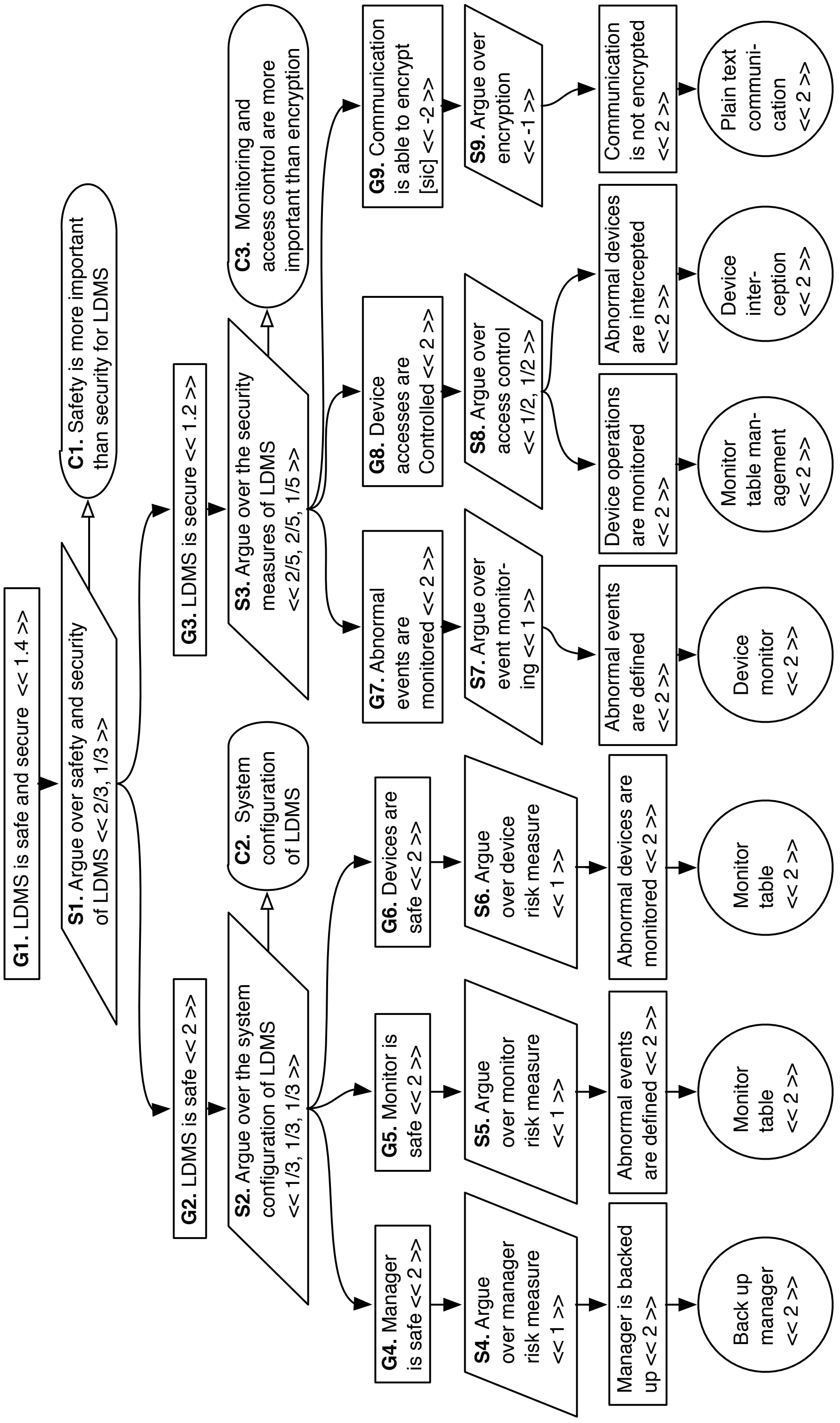

G1. LDMS is safe and secure << 1.4 >>
S1. Argue over safety and security of LDMS <<2/3, 1/3 >>
C1. Safety is more important than security for LDMS
G2. LDMS is safe << 2 >>
G3. LDMS is secure << 1.2 >>
S2. Argue over the system configuration of LDMS << 1/3, 1/3, 1/3 >>
C2. System configuration of LDMS
S3. Argue over the security measures of LDMS << 2/5, 2/5, 1/5 >>
C3. Monitoring and access control are more important than encryption
G4. Manager is safe << 2 >>
G5. Monitor is safe << 2 >>
G6. Devices are safe << 2 >>
G7. Abnormal events are monitored << 2 >>
G8. Device accesses are Controlled << 2 >>
G9. Communication is able to encrypt [sic] << -2 >>
S4. Argue over manager risk measure << 1 >>
S5. Argue over monitor risk measure << 1 >>
S6. Argue over device risk measure << 1 >>
S7. Argue over event monitoring << 1 >>
S8. Argue over access control << 1/2, 1/2 >>
S9. Argue over encryption << -1 >>
Manager is backed up << 2 >>
Abnormal events are defined << 2 >>
Abnormal devices are monitored << 2 >>
Abnormal events are defined << 2 >>
Device operations are monitored << 2 >>
Abnormal devices are intercepted << 2 >>
Communication is not encrypted << 2 >>
Back up manager << 2 >>
Monitor table << 2 >>
Monitor table << 2 >>
Device monitor << 2 >>
Monitor table management << 2 >>
Device interception << 2 >>
Plain text communication << 2 >>

Figure J1: *An example annotated GSN assurance argument taken from [32].*

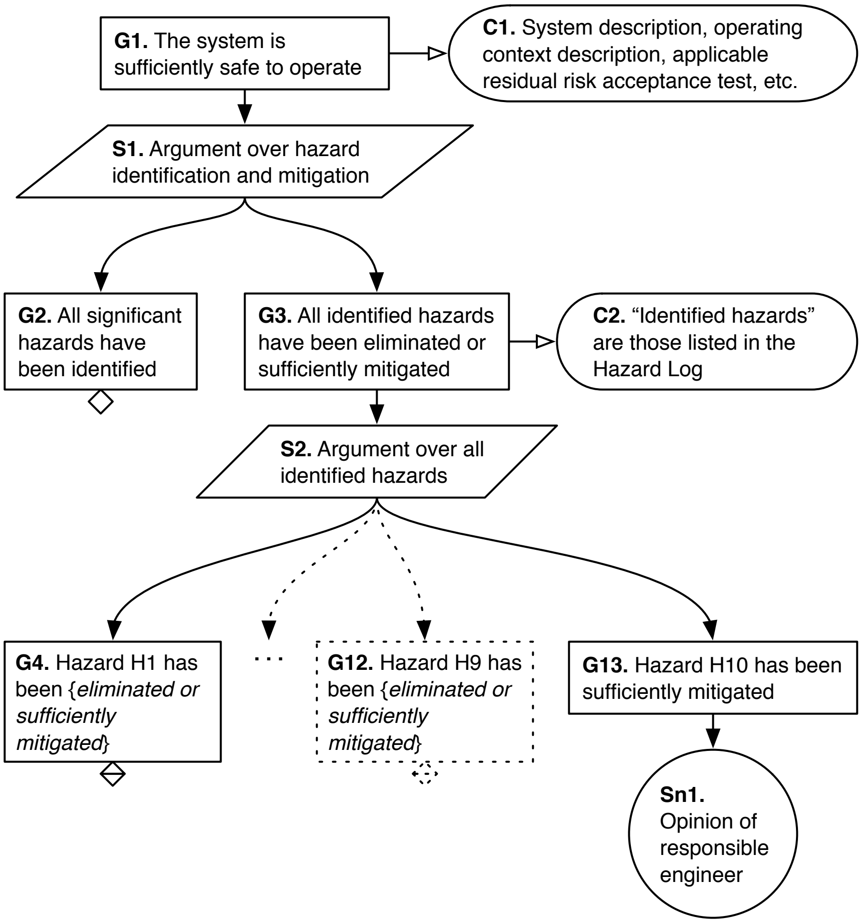

G1. The system is sufficiently safe to operate
C1. System description, operating context description, applicable residual risk acceptance test, etc.
S1. Argument over hazard identification and mitigation
G2. All significant hazards have been identified
G3. All identified hazards have been eliminated or sufficiently mitigated
C2. "Identified hazards" are those listed in the Hazard Log
S2. Argument over all identified hazards
G4. Hazard H1 has been {eliminated or sufficiently mitigated}
...
G12. Hazard H9 has been {eliminated or sufficiently mitigated}
G13. Hazard H10 has been sufficiently mitigated
Sn1. Opinion of responsible engineer

Figure J2: *The specimen safety argument we created and used to assess the proposed technique. The Optimistic and Poor Evidence 1 examples have all elements, but the Poor Evidence 2 example lacks goals G5–G13.*

poor information about the mitigation of others, and (b) assessments depend on the arbitrary scope of hazards.

Optimistic Example. In the original example, the assessment of every solution that indirectly supports the claim of safety is 2 (strongly satisfied). In our Optimistic example, we correspondingly give a value of 2 to G2, Sn1, and all instantiations of G4. Since each component in the original example had equal weight, we use equal weight for S1 and S2. Using the technique's propagation rules, we calculate that G3 and G1 also have value 2. These values are optimistic—evidence of safety is rarely, if ever, perfect—but mirror the original example, which assesses the claim that the "LDMS is safe" as 2 (strongly satisfied).

Poor Evidence 1 and 2 Examples. Our Poor Evidence 1 and Poor Evidence 2 variants of the Optimistic example suppose that we have very

Element	**Optimistic**	**Poor Evidence 1**	**Poor Evidence 2**
G2, G4	2.00		
G5–G13	2.00		N/A
Sn1, G14	2.00	-1.00	
S1, S2	Equal weights		
G3	2.00	1.70	0.50
G1	2.00	1.85	1.25

Table J1: *Attribute values for the Optimistic, Poor Evidence 1, and Poor Evidence 2 variants of the safety argument in Figure J2.*

poor evidence to show that one of the hazards is adequately mitigated. Figure J2 shows that the only evidence cited in support of the claim that hazard H10 has been sufficiently mitigated is the opinion of the responsible engineer (Sn1). Suppose that the engineer's opinion is that the hazard probably isn't mitigated. We thus assess Sn1 as -1 (unsatisfied) for both examples. The difference between the Optimistic and Poor Evidence 1 examples is the assessment of Sn1. The only difference between the Poor Evidence 1 and 2 examples is that in the latter, safety engineers have defined hazard H1 so that its scope encompasses the system states defined by hazards H1–H9 in the former.

Analysis of Examples. Table J1 gives the calculated attributes of the main safety claim, G1. In the Optimistic case, the attribute of G1 is 2 (strongly satisfied). But in the Poor Evidence 1 case, the attribute of G1 is 1.85. That is, we should be strongly satisfied that the system is safe even though all that we know about one of its hazards is that, in the responsible engineer's opinion, it is insufficiently mitigated. These examples show that the proposed technique allows evidence for the mitigation of some hazards to mask poor information about the mitigation of others.

The Poor Evidence 1 and Poor Evidence 2 examples differ only in the scope of the hazard definitions: the system states defined as hazards H1–H9 in the former are defined as hazard H1 in the latter. The same hazardous system states are mitigated in the same way as shown by the same evidence. Yet the calculated attribute for the main safety claim G1 is 1.85 (strongly satisfied) in the former case and 1.25 (satisfied) in the latter case. These examples show that the proposed technique produces assessments that depend on arbitrary hazard scope. But it is not plausible than an arbitrary difference in hazard scope should change our confidence in system safety.

Appendix K

Zeng, Lu, and Zhong

Zeng, Lu, and Zhong propose using Dempster–Shafer theory to assess the confidence in arguments recorded in the Goal Structuring Notation (GSN) [3, 24].

K.1 Proposed Technique

The proposed technique defines confidence on a 5-point Likert scale and uses an improved Dempster's rule to assess confidence in conclusions.

Confidence Scale. The technique defines confidence in an assurance argument claim as a probability mass distributed over six confidence and uncertainty categories:

A_1	Very low confidence
A_2	Low confidence
A_3	Medium confidence
A_4	High confidence
A_5	Very high confidence
θ	Unknown confidence / uncertainty in assessment

Source of Data. Assessments of claims based directly on evidence "can be obtained by [using the] Delphi Method." Weights (w_i) for each subgoal's contribution to a parent goal "can be obtained by [using the] Analytic Hierarchy Process Method." Weights sum to 1 for each conclusion.

Test for Argument Sufficiency. The paper does not describe how to use the calculated values to decide whether an argument is sufficient to justify fielding a system.

Illustrative Example. Figure K1 depicts the example argument given in the paper. Table K1 gives the belief masses for the goals in Figure K1. Table K2 gives the weights for the inferences depicted in Figure K1.

K.2 Replication

The equations in the paper contain typos. Moreover, the paper uses multiple assessments of the confidence in some premises to illustrate various scenarios and it is not clear which assessments were used to derive which assessments of confidence in the conclusion. However, we were able to reverse-engineer combination rules that roughly correspond with those given in the paper and produce the computed masses shown in the paper.

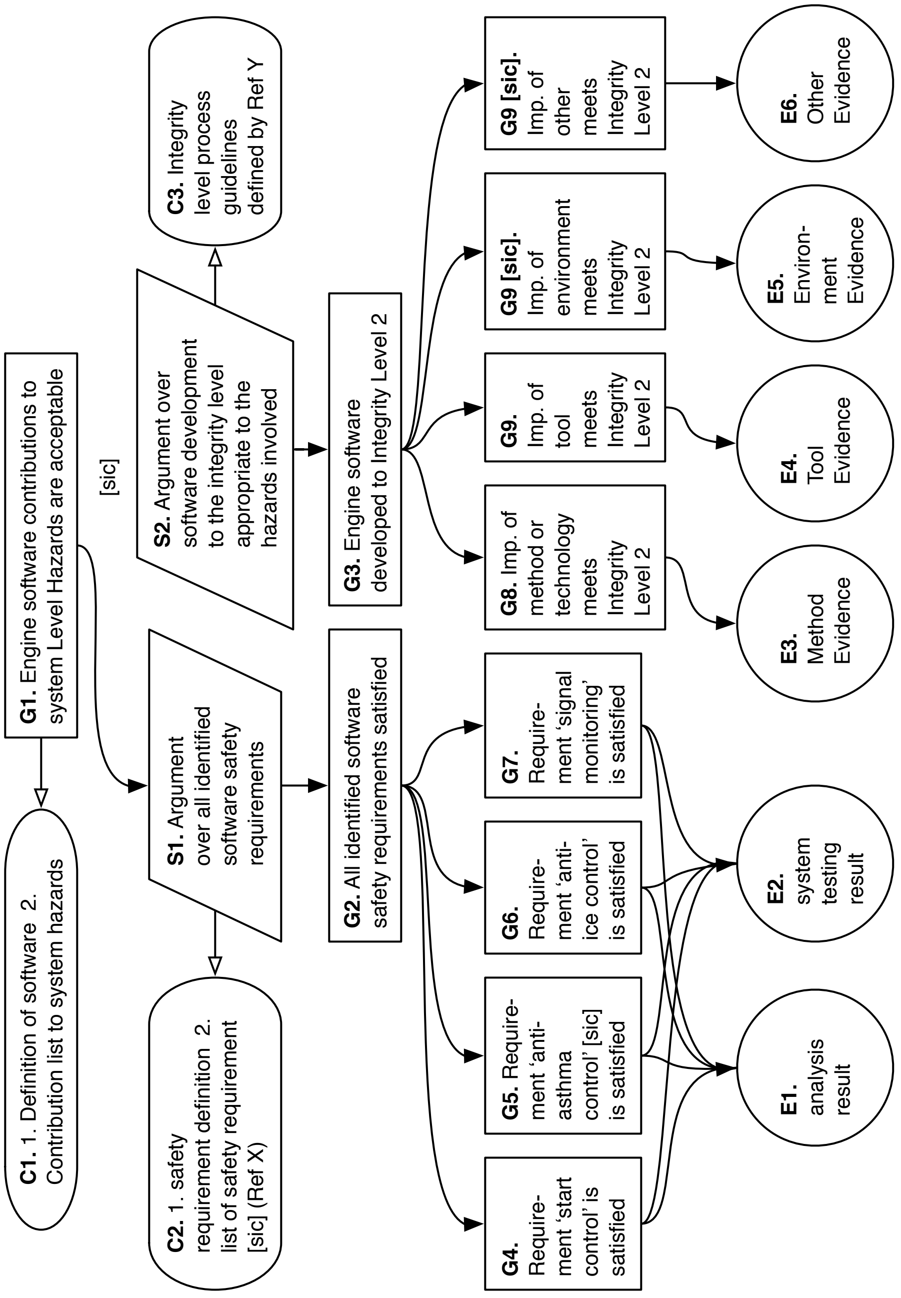

C1. 1. Definition of software 2. Contribution list to system hazards
G1. Engine software contributions to system Level Hazards are acceptable
[sic]
C2. 1. safety requirement definition 2. list of safety requirement [sic] (Ref X)
S1. Argument over all identified software safety requirements
S2. Argument over software development to the integrity level appropriate to the hazards involved
C3. Integrity level process guidelines defined by Ref Y
G2. All identified software safety requirements satisfied
G3. Engine software developed to Integrity Level 2
G4. Require-ment 'start control' is satisfied
G5. Require-ment 'anti-asthma control' [sic] is satisfied
G6. Require-ment 'anti-ice control' is satisfied
G7. Require-ment 'signal monitoring' is satisfied
G8. Imp. of method or technology meets Integrity Level 2
G9. Imp. of tool meets Integrity Level 2
G9 [sic]. Imp. of environment meets Integrity Level 2
G9 [sic]. Imp. of other meets Integrity Level 2
E1. analysis result
E2. system testing result
E3. Method Evidence
E4. Tool Evidence
E5. Environ-ment Evidence
E6. Other Evidence

Figure K1: *An example "fragment of the safety case of engine software," taken from [24] and depicted in GSN [3]. We assume that (i) the second and third goal G9 are meant to be goals G10 and G11, (ii) strategy S1 should really appear between goal G2 and supporting goals G4–G7, (iii) strategy S2 should appear between goal G3 and supporting goals G8–G10, and (iv) goal G1 is supported by goals G2 and G3.*

	Lower ← Confidence → Higher					Unknown
Goal	$(\mathbf{A_1})$	$(\mathbf{A_2})$	$(\mathbf{A_3})$	$(\mathbf{A_4})$	$(\mathbf{A_5})$	(θ)
G4	0.03	0.10	0.48	0.34	0.05	0.00
G5	0.10	0.10	0.35	0.37	0.05	0.03
G6	0.02	0.10	0.46	0.36	0.05	0.01
G7	0.01	0.10	0.53	0.29	0.05	0.02
G8	0.00	0.05	0.31	0.47	0.17	0.00
G9	0.00	0.05	0.23	0.54	0.18	0.00
G10	0.00	0.05	0.28	0.58	0.08	0.01
G11	0.00	0.05	0.33	0.55	0.05	0.02

Table K1: *Original belief probability masses $m_4 \dots m_7$ for the goals directly supported by evidence (G4–G11) in Figure K1 [24].*

Premise	Claim	Weight (ω)
G2	G1	0.731
G3	G1	0.269
G4	G2	0.250
G5	G2	0.250
G6	G2	0.312
G7	G2	0.188
G8	G3	0.336
G9	G3	0.234
G10	G3	0.257
G11	G3	0.173

Table K2: *Weights for inferences in Figure K1 [24].*

The Improved Dempster's Rule. Equation K1 and its supporting equations give the part of the combination of masses m_1 and m_2 that correspond to confidence level A_i in terms of the masses of the premises (m_1 and m_2) and the weights of each premise toward the conclusion (ω_i). Equation K2 and its supporting equations give the part of the combination of masses that corresponds to unknown confidence (θ). When more than two premises support a conclusion, the analyst applies Equations K1 and K2 applied repeatedly (to combine the first two, and then the result with the third, and so on) and the maximum weight to be used in Equations K3 and K4 is the maximum of all n premises.

$$(m_1 \oplus m_2)(A_i) = \frac{m'_1(A_i)\, m'_2(A_i) + m'_1(\theta)\, m'_2(A_i) + m'_1(A_i)\, m'_2(\theta)}{(m_1 \oplus m_2)''(\theta) + \sum_i (m_1 \oplus m_2)''(A_i)} \tag{K1}$$

$$(m_1 \oplus m_2)(\theta) = \frac{m'_1(\theta)\, m'_2(\theta)}{(m_1 \oplus m_2)''(\theta) + \sum_i (m_1 \oplus m_2)''(A_i)} \tag{K2}$$

$$m'_i(A_k) = \frac{\omega_i}{\max(\{\omega_1, \omega_2, \dots \omega_n\})} m(A_i) \tag{K3}$$

$$m_i'(\theta) = \frac{\omega_i}{\max(\{\omega_1, \omega_2, \ldots \omega_n\})} m\,(\theta) + \left(1 - \frac{\omega_i}{\max(\{\omega_1, \omega_2, \ldots \omega_n\})}\right) \quad \text{(K4)}$$

Using these and the belief mass and weight figures given in the paper, we were able to compute the same belief mass m_1 given in the paper and shown in Table K4. Given the belief masses and weights used in the example, medium (A_3) confidence in safety is not implausible.

K.3 Hypotheses and Evidence

The paper presents an illustrative "experimental example" [24] but gives no hypothesis to be tested, describes no independent or dependent variables, and discusses no threats to validity. The authors observe that

> the belief values which are high before combination become higher after combination, the belief values which are low before combination become lower after combination, [and] the uncertainty become lower and lower during the combination process .... In whole process, we can see G2 ... is the more degree contributions to the confidence in safety case, which results are consistent with common sense. Therefore considered that [Dempster–Shafer] evidence theory in the application of qualitative assessment is valid, and can better solve uncertainty of assessment result ....
>
> It is proved that [Dempster–Shafer] evidence theory has advantages in evaluating confidence in safety case which has some uncertainty. The usage of [Dempster–Shafer] evidence theory reduced the effect of the uncertainty, improved the precision and the validity of the evaluation, and reduced the blindness and the subjectivity of the evaluation of confidence in safety case.

It is not clear what the proposed technique (or Dempster–Shafer theory more broadly) is meant to be an improvement over. The paper describes no study that compares it to any alternative.

K.4 Counterargument

We created two variants of the original example, *Optimistic* and *Missing Evidence*, to illustrate that the proposed technique allows evidence that some safety requirements are met to overcome a lack of evidence that others are met. In the original example and our variants, goals G4–G7 in Figure K1 represent claims, backed by both "analysis" and "system testing," that the system satisfies each of its four safety requirements. In the original example, the bulk of the belief mass in each is spread over medium (A_3) and high (A_4) confidence.

	Lower ← Confidence → Higher					**Unknown**
Goal	**(A_1)**	**(A_2)**	**(A_3)**	**(A_4)**	**(A_5)**	**(θ)**
G4–G6 (O)	0.01	0.02	0.04	0.08	0.83	0.02
G4–G6 (ME)	0.01	0.02	0.04	0.08	0.83	0.02
G7 (O)	0.01	0.02	0.04	0.08	0.83	0.02
G7 (ME)	0.83	0.08	0.04	0.02	0.01	0.02

Table K3: *Belief probability masses $m_4 \ldots m_7$ for the Optimistic (O) and Missing Evidence (ME) variants of goals G4–G7 in Figure K1.*

Example	**Lower ← Confidence → Higher**					**Unknown**
Variant	**(A_1)**	**(A_2)**	**(A_3)**	**(A_4)**	**(A_5)**	**(θ)**
Original	0.003	0.022	0.576	0.391	0.008	0.001
Optimistic	0.000	0.001	0.002	0.008	0.988	0.001
Missing Evidence	0.003	0.002	0.005	0.015	0.973	0.001

Table K4: *Calculated belief mass m_1 for goal G1 in Figure K1. Since the computed confidence in a claim is defined as the confidence category in which the highest proportion of the belief mass lies, confidence in G1 is medium (A_3) in the original example given in the paper and very high (A_5) in both our Optimistic and Missing Evidence variants.*

Optimistic Example. In our Optimistic example, we put the bulk of our belief mass in the very high (A_5) confidence category.

Missing Evidence Example. In our Missing Evidence example, we assume very high confidence in verification and validation for three requirements, but very low confidence for the fourth. This simulates missing (or undermined) evidence of the satisfaction of the fourth requirement.

Analysis of Examples. Table K3 gives the belief mass figures for our Optimistic and Missing Evidence examples. Table K4 gives the resulting belief mass in the system safety proposition (goal G1). Since in both cases the vast bulk of the calculated belief mass lies in the very high (A_5) confidence category, we conclude that in both the Optimistic and Missing Evidence variants the proposed technique assesses very high confidence in the main safety claim. It is implausible to have nearly identical confidence in the safety of both examples. It is also implausible to have very high confidence in the safety of a system when there is only very low confidence that one of its four safety requirements is met.

Appendix L

Zhao, Zhang, Lu, and Zeng

Zhao, Zhang, Lu, and Zeng propose to calculate confidence using Bayesian Belief Networks (BBNs) [25].

L.1 Proposed Technique

In the proposed technique, an analyst assesses confidence by:

1. Interpreting a given assurance argument as a sequence of Toulmin arguments [34]
2. Creating a BBN by instantiating a given pattern for each instance
3. Supplying the requisite probabilities

Test for Argument Sufficiency. The paper does not describe how to use the calculated values to decide whether an argument justifies putting a safety-critical system into service.

Illustrative Example. The paper illustrates the proposed technique by applying it to "the typical safety argument." Figure L1 shows its example of the proposed BBN pattern. Tables L1 and L2 give the probabilities for each node.

L.2 Replication

We had little difficulty reproducing the example. Using the given figures and node probability tables (NPTs), we also calculated 89% confidence in node N11 "Justified Claim 'system S is safe.'"

L.3 Hypotheses and Evidence

The paper describes its example as a "simplified case study" but does not present information usually found in case study reports such as descriptions of the study method, the data collected, and how these data relate to the questions at issue [58]. On the basis of their example, the authors cautiously deem their approach "a potentially helpful way towards the measurement of confidence in assurance cases" [25].

L.4 Counterargument

We created two variants of the original example, namely *Optimistic* and *Pessimistic*, that represent extreme assessments of confidence in the completeness of hazard identification. We use these examples to illustrate that the proposed technique is not sensitive enough to this critical factor.

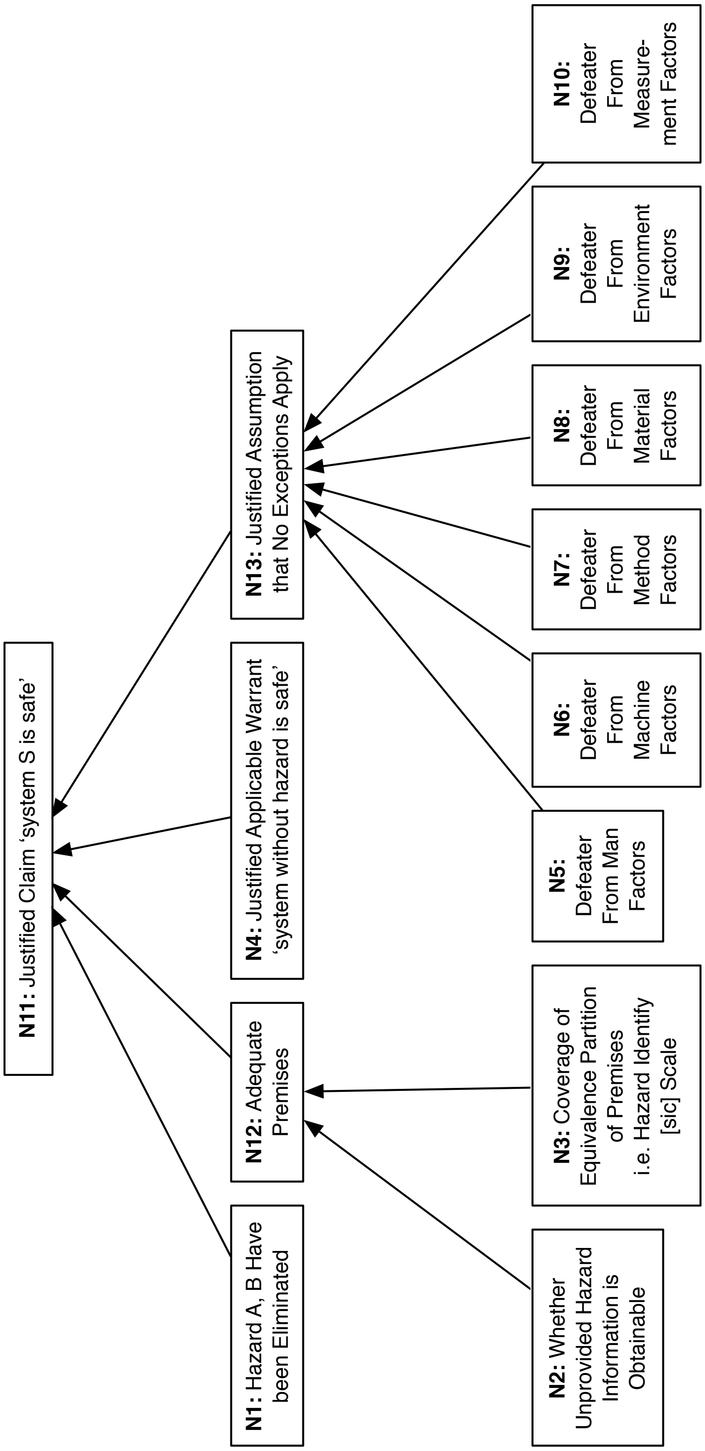

N11: Justified Claim ‘system S is safe’
N1: Hazard A, B Have been Eliminated
N12: Adequate Premises
N4: Justified Applicable Warrant ‘system without hazard is safe’
N13: Justified Assumption that No Exceptions Apply
N2: Whether Unprovided Hazard Information is Obtainable
N3: Coverage of Equivalence Partition of Premises i.e. Hazard Identify [sic] Scale
N5: Defeater From Man Factors
N6: Defeater From Machine Factors
N7: Defeater From Method Factors
N8: Defeater From Material Factors
N9: Defeater From Environment Factors
N10: Defeater From Measure-ment Factors

Figure L1: *An example “basic BBN for the typical safety argument,” taken from [25]. For compactness, we represent BBNs nodes as rectangles.*

Node	**Value**	**Original**	**Optimistic**	**Pessimistic**
N1: Hazard A, B Have been Eliminated	Present	90%	90%	90%
	NotPresent	10%	10%	10%
N2: Whether Unprovided Hazard Information . . .	Complete	85%	85%	85%
	NotComplete	15%	15%	15%
N3: Coverage of Equivalence Partition . . .	Complete	85%	**99.9**%	**0.1**%
	NotComplete	15%	**0.1**%	**99.9**%
N4: . . . Warrant 'system without hazard is safe'	Complete	100%	100%	100%
	NotComplete	0%	0%	0%
N5–10: Defeater from {Man, Machine, . . .} Factors	Complete	0%	0%	0%
	NotComplete	100%	100%	100%
N11: Justified Claim 'system S is safe'	Present	If N1, N12, N4, and N13 are Present		
	NotPresent	Otherwise		
N12: Adequate Premises	See Table L2.			
N13: Justified Assumption that No Exceptions Apply	Present	If N6–N10 are NotPresent		
	NotPresent	Otherwise		

Table L1: *Node values for the example shown in Figure L1.*

N3	Complete		NotComplete	
N2	Obtainable	NotObtainable	Obtainable	NotObtainable
Present	1	1	0.8	1
NotPresent	0	0	0.2	0

Table L2: *Node probability table for N12 in Figure L1, taken from [25]. The paper notes that this conditional probability table is "very arbitrary" and "in practice it should be discussed by experts and stakeholders" [25].*

Case	**Present**	**NotPresent**
From paper	89.460%	10.540%
Optimistic	89.996%	10.004%
Pessimistic	86.404%	13.596%

Table L3: *Calculated value of N11 ("Justified Claim 'system S is safe'") for the original version presented in [25] and our Optimistic and Pessimistic variants.*

Optimistic Example. Node N3 ("Coverage of Equivalence Partition of Premises") represents confidence in the completeness of the hazard analysis (i.e., whether all significant hazards were identified). While the original example shows 85% confidence in N3, our Optimistic variant uses 99.9%, leaving all other inputs and CPTs unchanged. This represents extreme optimism in the completeness of the hazard identification.

Pessimistic Example. Our Pessimistic example is identical to the original and Optimistic examples except that we assess the confidence in N3 as 0.1%. This represents extreme pessimism in the completeness of the hazard identification.

Analysis of Examples. Table L3 gives the calculated confidence in safety for all three examples. It is not plausible that extreme changes in confidence in hazard analysis would produce as small a change in confidence in safety as the difference between 90% and 86% indicates. It is also implausible that anyone who completely distrusted a hazard analysis would have 86% confidence that the analyzed system is safe.

REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704–0188*

1. REPORT DATE *(DD-MM-YYYY)* 01-05-2016	**2. REPORT TYPE** Technical Memorandum	**3. DATES COVERED** *(From - To)* September–December 2015

4. TITLE AND SUBTITLE An Investigation of Proposed Techniques for Quantifying Confidence in Assurance Arguments	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Graydon, Patrick J. Holloway, C. Michael	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER 999182.02.50.07.02
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) NASA Langley Research Center Hampton, Virginia 23681-2199	**8. PERFORMING ORGANIZATION REPORT NUMBER** L–20670
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) National Aeronautics and Space Administration Washington, DC 20546-0001	**10. SPONSOR/MONITOR'S ACRONYM(S)** NASA
	11. SPONSOR/MONITOR'S REPORT NUMBER(S) NASA/TM–2016–219195

12. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified-Unlimited
Subject Category 03
Availability: NASA CASI (443) 757-5802

13. SUPPLEMENTARY NOTES

An electronic version can be found at http://ntrs.nasa.gov.

14. ABSTRACT

The use of safety cases in certification raises the question of assurance argument sufficiency and the issue of confidence (or uncertainty) in the argument's claims. Some researchers propose to model confidence quantitatively and to calculate confidence in argument conclusions. We know of little evidence to suggest that any proposed technique would deliver trustworthy results when implemented by system safety practitioners. Proponents do not usually assess the efficacy of their techniques through controlled experiment or historical study. Instead, they present an illustrative example where the calculation delivers a plausible result. In this paper, we review current proposals, claims made about them, and evidence advanced in favor of them. We then show that proposed techniques can deliver implausible results in some cases. We conclude that quantitative confidence techniques require further validation before they should be recommended as part of the basis for deciding whether an assurance argument justifies fielding a critical system.

15. SUBJECT TERMS

safety case, assurance argument, confidence, uncertainty

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			STI Help Desk (email: help@sti.nasa.gov)
U	U	U	UU	96	**19b. TELEPHONE NUMBER** ***(Include area code)*** (443) 757-5802