Towards Measurement of Confidence in Safety Cases

Ewen Denney† and Ganesh J. Pai
†*Robust Software Engineering Group*
SGT Inc., NASA Ames Research Center
Moffett Field, CA, USA
Email: {Ewen.W.Denney, ganesh.pai}@nasa.gov

Ibrahim Habli
Department of Computer Science
University of York
York, UK
Email: Ibrahim.Habli@cs.york.ac.uk

***Abstract*—Arguments in safety cases are predominantly qualitative. This is partly attributed to the lack of sufficient design and operational data necessary to measure the achievement of high-dependability targets, particularly for safety-critical functions implemented in software. The subjective nature of many forms of evidence, such as expert judgment and process maturity, also contributes to the overwhelming dependence on qualitative arguments. However, where data for quantitative measurements is systematically collected, quantitative arguments provide far more benefits over qualitative arguments, in assessing confidence in the safety case. In this paper, we propose a basis for developing and evaluating integrated qualitative and quantitative safety arguments based on the Goal Structuring Notation (GSN) and Bayesian Networks (BN). The approach we propose identifies structures within GSN-based arguments where uncertainties can be quantified. BN are then used to provide a means to reason about confidence in a probabilistic way. We illustrate our approach using a fragment of a safety case for an unmanned aerial system and conclude with some preliminary observations.**

***Keywords*-Safety case; safety; uncertainty analysis; measurement; bayesian networks**

I. Introduction

NOTE: Introduction to the context / problem:

- Why safety cases are important? - Importance of explicit reasoning, through structured arguments, within safety cases - Its inevitable that a big part of the reasoning is subjective and inductive, partly because of uncertainties in the argument and evidence (inherited from uncertainties in system development, assessment and operation) - However probabilistic reasoning is the natural way, from an engineering perspective, for addressing uncertainties and the lack of full assurance (e.g. risks at the system level are often stated in quantitative terms) - We acknowledge our inability to fully quantify assurance and propose to address this by an approach that exploits the benefits of integrating qualitative and qualitative reasoning within a safety case by using GSN and BN.

NOTE: Quantification is not for assignment of SIL / DAL and computing corresponding confidence level. Rather the goal is to compute / measure the confidence in the safety argument and use this as a basis for making decisions i.e. to accept the safety argument or reject it. The reason for this is simple: no one can make sense of SILs and DALs. The rationale behind these is not documented anywhere!

This paper is organized as follows: Section II discusses related work in the literature. In section III, we present our proposed approach for the quantification of uncertainty (confidence) in safety cases: specifically, we describe the illustrative example safety argument in section III-A, and the quantification model in section III-B. We discuss our approach in section IV and conclude with directions for future work in section V

II. Related Work

NOTE: Short review of the two main camps: pro- and anti- quantification of assurance and how this is addressed in safety standards (Safety Integrity and Assurance Levels SILs and DALs)

NOTE: Related important work: [1], [2], [3], [4], [5]

III. Proposed Approach

NOTE: A few words here on what our general approach is, for uncertainty measurement: An approach where GSN is used to construct the primary safety argument and BN is constructed to quantify confidence where enough data is available

NOTE: Also need an introduction to GSN Short intro to BN.

A. Illustrative Example

Figure 1 shows a fragment of the safety argument, described using the Goal Structuring Notation (GSN) [6] for the airborne subsystem of an experimental unmanned aerial system (UAS), being developed at NASA Ames Research Center. In the context of the overall UAS safety case and the corresponding hazard analysis, the correct functioning of the autopilot in the airborne system has been determined to be one of many functional safety requirements for mitigating certain hazards e.g., drifting outside the range-safety area. In decomposing the goal corresponding to this functional safety requirement, the correct calculation of the angle of attack of the aircraft (G1) is one sub-goal. In this paper we discuss ways to measure confidence in the argument and quantify the uncertainty in this claim.

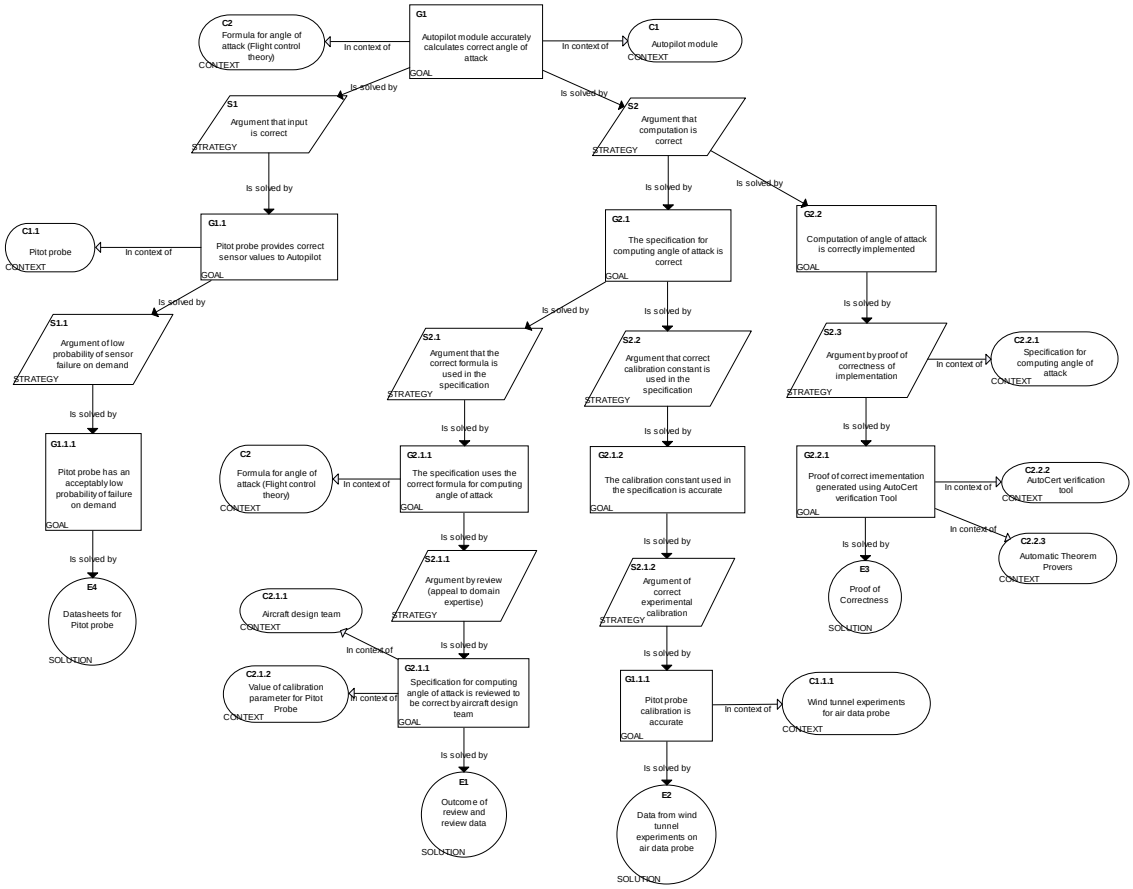

Figure 1. Fragment of the safety case of the airborne subsystem of the UAS.

As shown in the figure, we address G1 by arguing that (1) G1.1: the Pitot (air-data) probe provides the correct sensor values to the autopilot (2) G2.1: the specification is correct and (3) G2.2: the implementation of this specification is also correct. In turn, these claims are justified in part (using the strategies shown in Figure 1) by (1) E1: evidence arising from empirical data wind tunnel experiments about correctly calibrating the air-data probe (2) E2: subjective assessment of the formula used in the specification as evidenced by the outcome of a review, (3) E3: formal verification of the implementation, using a proof of correctness, and (4) E4: evidence of low probability of failure on demand (PFD) obtained from sensor datasheets.

To gauge whether G1 is to be accepted e.g., by a regulator, it is reasonable to present an additional argument to justify the sufficiency of confidence in the claim (and, as a consequence, the overall argument fragment shown). For instance, as in [2], a qualitative confidence argument may be created in which it is argued that (a) there is credible support for the inference asserted via the claims G1.1, G2.1 and G2.2 that G1 is true, (b) the assurance deficits for this asserted inference have been identified and (c) that the residual assurance deficits are acceptable. Unfortunately, although there is some guidance available on identifying where the assurance deficits lie [7], there is little guidance on *how* it may be gauged that the residual assurance deficit is acceptable. Here, the challenge for the regulator is in assessing that a qualitative argument (i.e., the confidence argument) provides sufficient confidence in another qualitative argument (i.e., the safety argument).

We believe that quantification of uncertainty and a measurement-based approach to evaluate the safety argument is an objective alternative for such decision-making. In this paper, we examine one approach towards quantifying confidence (uncertainty), and discuss the challenges therein. This approach will augment, rather than replace, qualitative arguments in a safety case.

B. Uncertainty Measurement

The sources of uncertainty in the argument for G1 (figure 1) are mainly:

(U1): Uncertainty in the sensor values is stochastic (aleatory) and it is attributed, in part, to the probability of failure of the Pitot probe, and to any errors in conversion of the sensed analog values to an appropriate digital equivalent.

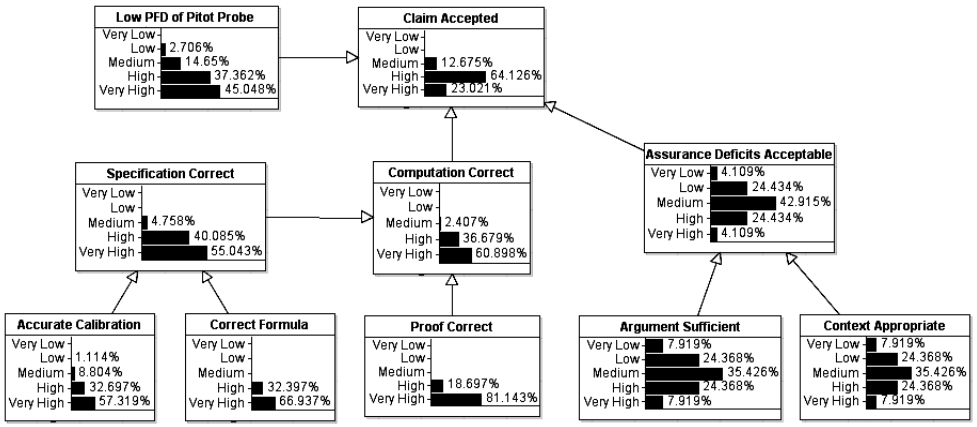

Figure 2. Bayesian network showing quantification of confidence in the claim given the sources of uncertainty in the argument fragment

In this paper, for the sake of simplicity, we assume that analog to digital conversion is perfect. This uncertainty is given by the variance in the measured failure rate (or probability of failure on demand) obtained, say, through statistical testing of the sensor and as specified on the sensor datasheets.

(U2): Uncertainty that specification is correct contains both aleatory and epistemic uncertainties: the calibration error of the Pitot probe (when the probe has not failed) is a source of aleatory uncertainty that contributes to the overall uncertainty in the correctness of the specification, whereas the uncertainty as to whether the formula for computing the angle of attack has been correctly used is the source of epistemic uncertainty. Calibration of the air-data probe is experimentally performed in a wind tunnel [8]. A confidence level can be used to effectively specify the confidence in the experiment and is obtained from statistical analysis of the corresponding empirical data. The confidence that the correct formula for computing the angle of attack is subjectively gauged by reviewing the specification against flight control theory by domain experts i.e., the aircraft design team.

(U3): Uncertainty that the implementation is correct is the uncertainty in the verification procedure i.e., the proof of correctness. The verification chain begins with parsing and pre-processing of the code, after which logical annotations are inferred based on the formal specification. Verification conditions (VC) are obtained from the annotated code by processing the latter with a verification condition generator (VCG). These are simplified, translated to TPTP [Reference needed], where various scripts then control first-order theorem provers to prove the VC from a set of axioms (some of which are static, and others are generated dynamically). Uncertainty in the proof of correctness is a combination of each of the elements involved in the verification chain. For this paper, we mainly gauge (U3) via the subjective judgment from the developers of the theorem prover, and leave the modeling of the sources of uncertainty in the verification chain, for future work.

Both (U2) and (U3) are epistemic uncertainties. Additional epistemic uncertainties arise from assurance deficits [2] in the safety argument itself.

(U4): Uncertainty in the sufficiency of the sub-claims (solutions) is the uncertainty whether the sub-claims (solutions) e.g., G1.1, G2.1, G2.2, are appropriate and sufficient to infer the parent claim (sub-claim) e.g., G1, or whether there is a need for additional sub-claims (solutions). In the context of multi-legged arguments [4], there is the additional uncertainty about the acceptability of a given argument leg, when there may be evidence to the contrary in another argument leg.

(U5): Uncertainty in the appropriateness of the context reflects on whether the context used for a claim or a strategy is appropriate.

C. Assessment of Confidence

To assess the uncertainty (confidence) in the claim G1, first we model the confidence in the claim and the sources of uncertainty (U1) - (U5) respectively as discrete random variables (r.v.); subsequently we characterize the overall confidence in the argument as the joint distribution of the r.v., and we use a Bayesian network (BN[1]) [9] to quantify this joint distribution.

A Bayesian paradigm is appropriate in this context because it permits the inclusion of both subjective and quantitative data. Additionally, BN allow us to (1) compute the joint distribution of r.v. by exploiting the conditional independence between the r.v. and (2) perform inference when evidence[2] is available. The structure of then network encodes the assumptions of conditional independence. Thus,

[1]The singular and the plural forms are given by the same acronym

[2]Note that evidence supplied in the BN is distinctly different from the evidence supplied in the safety argument itself. The former is evidence of increasing (or complete) credibility in the latter.

Table I
MAPPING R.V. STATES TO A UNIT INTERVAL

State	Interval
Very Low	[0, 0.2)
Low	[0.2, 0.4)
Medium	[0.4, 0.6)
High	[0.6, 0.8)
Very High	[0.8, 1]

the arcs represent dependencies between the r.v. and may be interpreted as correlation. Each of the r.v. has a defined set of states and an associated probability distribution over those states.

In the BN shown in figure 2, the root node Claim Accepted (a node with only incoming arcs) models the confidence in the claim G1. The leaf nodes (nodes without incoming arcs) model each of the identified sources of uncertainty e.g., the node Proof models the confidence in the solution E3: Proof of correctness, corresponding to the source of uncertainty (U3). The intermediate nodes (nodes with both incoming and outgoing arcs e.g., Computation Correct) abstract and aggregate relevant leaf nodes; additionally, they serve to reduce the complexity associated with the specification of conditional probabilities and in post-specification inference.

All the nodes in the BN have the same set of five states: ⟨very low, low, medium, high, very high⟩ which are mapped to the interval [0, 1] as shown in Table I. Such a mapping allows including confidence values that have been obtained from both quantitative data (e.g., the confidence level associated with the experimental calibration of the air data probe), and from qualitative means (e.g., the reviewer confidence in specification correctness).

The quantitative specification for each of the leaf nodes is given as a prior probability distribution over the states of the node; in particular, we use a (doubly) truncated Normal distribution [10] 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.

For intermediate nodes and the root nodes we specify a prior conditional probability distribution (CPD) in a parametric way, again using a truncated Normal distribution. Here, the mean of the distribution is the weighted average of the parent r.v. while the variance is the inverse of the sum of the weights [10]. The weights can be considered as modeling the "strength of correlation" between the r.v. In the context of a safety argument, this would be viewed as the importance assigned to the contribution of a certain source of uncertainty to the overall confidence.

Thus, if C_c, C_p, C_s and C_{cc} are the r.v. modeling the confidence in the accurate calibration of the air data probe, the correctness of the proof, the correctness of the specification, and the correct computation respectively, $\pi(X)$ is a prior distribution over a random variable X, and $\mathcal{N}_T(\mu, \sigma^2)$ is the truncated Normal distribution with mean μ and variance σ^2, we have:

- $\pi(C_c) \sim \mathcal{N}_T(\mu_c, \sigma_c^2)$, where μ_c is given by the confidence measure of the experiment. In Figure 2, $\pi(C_c) \sim \mathcal{N}_T(0.9, 0.05)$ corresponds to the prior measure of a 90% confidence level in the calibration experiment of the air data probe.
- $\pi(C_p) \sim \mathcal{N}_T(\mu_p, \sigma_p^2)$, where μ_p is given by is given by the subjective measure of confidence in the proof. In Figure 2, $\pi(C_p) \sim \mathcal{N}_T(0.9, 0.05)$ would be interpreted, for instance, that there is *a priori* "very high" confidence in the proof of correctness to be supplied as evidence.
- $\pi(C_{cc}|C_p, C_s) \sim \mathcal{N}_T(\mu_{cc}, \sigma_{cc}^2)$ is the CPD of the confidence in correct computation, given the confidence in the proof and the specification; μ_{cc} is given as $((100C_p + 100C_s)/200)$ i.e., the weighted average of the parent r.v., with each given equal weight; σ_{cc}^2 is chosen as the inverse of the sum of weights i.e., 0.005.

The specification of the priors for the rest of the r.v. in the BN (Figure 2) is given in a similar way. Once the specification of the BN is complete additional evidence of confidence may be included, if available, to examine how this modifies the confidence in the overall claim.

IV. DISCUSSION

There are several challenges when considering the quantification of uncertainty (confidence) in a safety argument as presented. The primary challenge is justifying and validating the model used for quantification. This is addressed at several levels:

where confidence is measurable e.g., through quantitative empirical data and statistical analysis, inclusion of confidence into the model is justifiable and straightforward. When considering subjective assessment of uncertainty,

1. Development of the BN Potentially automatable from the identified uncertainty sources in the GSN and the GSN structure itself. Augment the syntax of the GSN with a placeholder for an appropriate confidence measure.

2. Justification of the leaf node probabilities: Where available, from quantitative data e.g., confidence levels in statistical experiments.

3. Justification of the CPD/ CPT: Are the conditional probabilities of a node given its parents representative of the true likelihood?

4. Justification of the weights: Based on the importance given to each path in the argument chain starting from goals down to evidence.

5. Justification for the BN structure: Are the assumptions of conditional independence valid?

6. Scaling the approach to the overall argument: problems of dependence between argument legs [3]

7. Some benefits: Reasoning about the confidence in a quantitative way may explicitly highlight issues not otherwise apparent from qualitative reasoning? For example, if a high prior confidence is given to a certain evidence source, this can be called out during assessment

V. Conclusions

The conclusion goes here. this is more of the conclusion

Acknowledgments

This work has been funded by NASA contract NNA10DE83C. Opinions, findings, conclusions, and recommendations expressed in this paper are not necessarily the views of NASA.

References

[1] R. Bloomfield, B. Littlewood, and D. Wright, "Confidence: its roles in dependability cases for risk assessment," in *Proceedings of the 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks (DSN)*, 2007.

[2] R. Hawkins, T. Kelly, J. Knight, and P. Graydon, "A new approach to creating clear safety arguments," in *Proceedings of the Safety Critical Systems Symposium*, Feb. 2011.

[3] P. Bishop, R. Bloomfield, B. Littlewood, A. Povyakalo, and D. Wright, "Towards a formalism for conservative claims about the dependability of software-based systems," *IEEE Transactions on Software Engineering*, Article in Press.

[4] B. Littlewood and D. Wright, "The use of multilegged arguments to increase confidence in safety claims for software-based systems: A study based on a bbn analysis of an idealized example," *IEEE Transactions on Software Engineering*, vol. 33, no. 5, pp. 347–365, May 2007.

[5] H. Herencia-Zapana, G. Hagen, and A. Narkawicz, "Formalizing probabilistic safety claims," in *Proceedings of the 3rd NASA Formal Methods Symposium*, Apr. 2011.

[6] T. Kelly and R. Weaver, "The goal structuring notation – a safety argument notation," in *Proceedings of the Dependable Systems and Networks Workshop on Assurance Cases*, Jul. 2004.

[7] C. Menon, R. Hawkins, and J. McDermid, "Interim standard of best practice on software in the interim standard of best practice on software in the context of ds 00-56 issue 4," Software Systems Engineering Initiative, University of York, Standard of Best Practice Issue 1, 2009.

[8] C. Ippolito, "Wind tunnel calibration of the exploration aerial vehicle (eav) five-hole pitot probe," NASA Ames Research Center, Technical Report, 2006.

[9] F. Jensen, *Bayesian Networks and Decision Graphs*. Springer-Verlag, 2001.

[10] N. Fenton, M. Neil, and J. Caballero, "Using ranked nodes to model qualitative judgments in bayesian networks," *IEEE Transactions on Knowledge and Data Engineering*, vol. 19, no. 10, pp. 1420–1432, Oct. 2007.