Towards Measurement of Confidence in Safety Cases

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

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

***Abstract*—Safety cases capture a structured argument linking claims about the safety of a system to the evidence justifying those claims. However, arguments in safety cases tend to be predominantly qualitative. Partly, this is attributed to the lack of sufficient design and operational data necessary to measure the achievement of high-dependability goals, 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 can be systematically collected, quantitative arguments provide benefits over qualitative arguments in assessing confidence in the safety case. In this paper, we propose a basis for developing and evaluating the confidence in integrated qualitative and quantitative safety arguments. We specify a safety argument using the Goal Structuring Notation (GSN), identify and quantify uncertainties therein, and use Bayesian Networks (BNs) as a means to reason about confidence in a probabilistic way. We illustrate our approach using a fragment of a safety case for an unmanned aircraft system (UAS).**

***Keywords*-Safety; Safety-case; Uncertainty; Measurement; Bayesian Networks**

I. Introduction

A safety case provides an explicit means for justifying the safety of a system through a reasoned argument and supporting evidence. Despite the many potential advantages that a safety case can provide with respect to the explicit consideration of safety assurance, subjectivity inherent in the structure of the argument and its supporting evidence, as well as the lack of sufficient statistical data, pose a key challenge to the measurement and quantification of confidence in the overall safety case. Consequently, confidence in a safety case is often assessed by appealing to qualitative reasoning.

In this paper, we explore the challenges of measuring confidence in safety cases; in particular, we propose an approach for confidence measurement by integrating probabilistic reasoning with Bayesian Networks (BNs) [1] into safety arguments represented in the Goal Structuring Notation (GSN) [2]. An overarching motivation for this work is, eventually, to integrate it into a quantitative framework for risk analysis [3].

II. Related Work

Serious concerns exist about current safety case practices [4], highlighting the need for methods to assess that sufficient confidence can be placed in safety cases. [5] proposes an assurance approach in which a safety case comprises two complementary arguments: the safety argument documents the reasoning supporting the claims concerning the safety of the system, while an interlinked, *qualitative* confidence argument documents the reasoning as to why the confidence in that safety argument is sufficient.

Others have also recognized the need to consider uncertainties in the safety argument, albeit from the perspective of quantification, e.g., in quantifying the epistemic uncertainty in dependability arguments when assessing the confidence in claims about the probability of failure [6]; in evaluating the confidence placed in safety arguments where claims address the achievement of a desired safety integrity level [7], and the quantification of confidence in diverse argument legs to examine whether diversity in arguments improves overall confidence in a safety claim [8]. Our work is closely related to this literature through our use of a Bayesian paradigm for uncertainty modeling and assessment.

III. Proposed Approach

In our proposed approach for measuring confidence in safety cases, we first construct the safety argument using GSN: a graphical notation for representing arguments in terms of basic elements such as claims, context and evidence. A GSN argument links these elements using two main relationships: *supported by* and *in context of*, to form a goal structure. Then we build the "confidence" argument for the safety argument by quantifying the uncertainty in the latter, where applicable, by using BNs. In particular, we use BNs to measure the confidence in the claims made (and, as a consequence, in the argument) by computing the joint distribution of a set of random variables (r.v.) that represent the quantified sources of uncertainty present in, and derived from, the safety argument.

A. Example Safety Argument

Figure 1(a) shows a fragment of the safety argument for the airborne subsystem of an experimental unmanned aircraft system (UAS), being developed at NASA Ames. Through hazard analysis, we have determined that the safe functioning of the autopilot requires the correct calculation of the angle of attack of the aircraft (G1). In this paper we discuss ways to measure confidence in the argument and quantify the uncertainty in this claim.

We address G1 by arguing that (a) G1.1: the Pitot (air-data) probe provides the correct sensor values to the autopilot (b) G2.1: the specification is correct and (c) G2.2: the implementation of this specification is also correct. In turn, these claims are justified in part (using the strategies shown) by (a) E1: evidence arising from wind tunnel experiments calibrating the air-data probe (b) E2: subjective assessment of the formula used in the specification as evidenced by the outcome of a review, (c) E3: formal verification of the implementation, using a proof of correctness, and (d) 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 [5], 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 [9], 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).

B. Uncertainty in the Safety Argument

The sources of uncertainty in the argument for G1, as shown in Figure 1(a), are mainly:

(U1): Uncertainty in the sensor values is stochastic (aleatory) and is attributed, in part, to the PFD of the Pitot probe, and to any errors in converting the sensed analog values to an appropriate digital equivalent. The former is given by the variance in the PFD (or measured failure rate in the case of continuous demand) obtained, say, through statistical testing of the sensor. We assume, for the sake of simplicity, that analog to digital conversion is perfect.

(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 is itself correct and is correctly used is a source of epistemic uncertainty. Calibration of the air-data probe is experimentally performed in a wind tunnel [10]. 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 is used to compute the angle of attack is subjectively gauged by reviewing the specification against flight control theory by domain experts e.g., 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 contains a combination of several steps and related tools [11] each of which induces an uncertainty that together contribute to the overall uncertainty that the proof is perfect. For this paper, we mainly gauge (U3) via subjective judgment from the developers of the verification tools. Modeling of the sources of uncertainty in the verification chain is left for future work.

Both (U2) and (U3) are epistemic uncertainties. Additional epistemic uncertainties arise from assurance deficits [5] in the safety argument itself, and are also subjectively quantified.

(U4): Uncertainty in the sufficiency of the sub-claims is the uncertainty whether the sub-claims 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.

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

C. Measuring 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 r.v.; subsequently we characterize the overall confidence in the argument as the joint distribution of the r.v., and we use BNs 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, BNs 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[1] is available. The structure of the network encodes the assumptions of conditional independence. Thus, 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 1(b), the root node Claim Accepted (a node with only incoming arcs) models the confidence in the claim G1. The leaf nodes (nodes without

[1] 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, decreasing, or complete credibility in the latter.

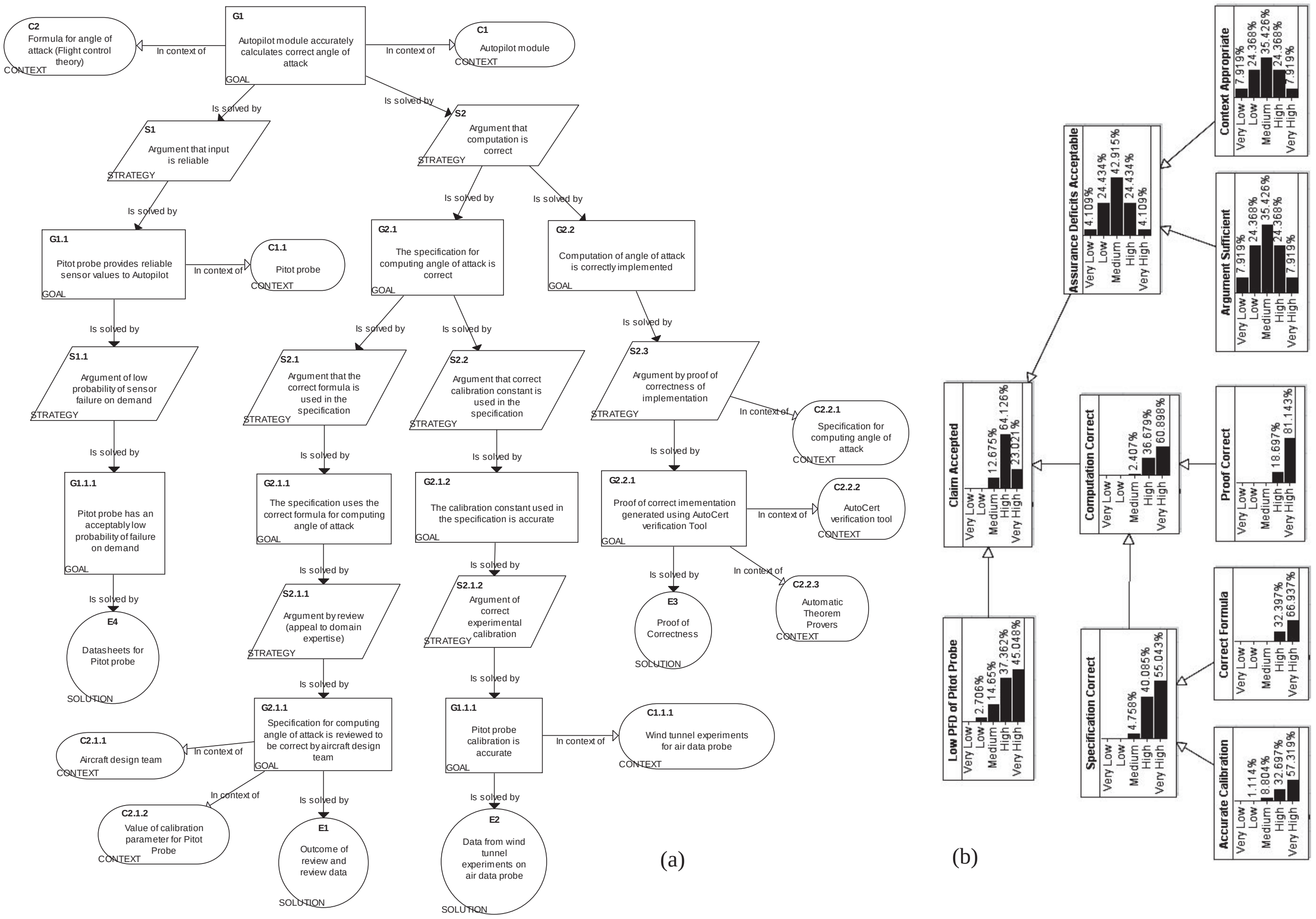

Figure 1. Integrated Safety Argumentation: (a) Fragment of the safety case of the airborne subsystem of the UAS (b) Corresponding BN, modeling sources of uncertainty and confidence in the claim G1.

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: ⟨very low, low, medium, high, very high⟩ ↔ ⟨[0, 0.2), [0.2, 0.4), [0.4, 0.6), [0.6, 0.8), [0.8, 1]⟩.

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 [12] 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 [12]. 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:

(i) $\pi(C_c) \sim \mathcal{N}_T(\mu_c, \sigma_c^2)$, where μ_c is given by the confidence measure of the experiment. In Figure 1(b), $\pi(C_c) \sim \mathcal{N}_T(0.95, 0.05)$ corresponds to the prior measure of a 95% confidence level in the calibration experiment of the air data probe.

(ii) $\pi(C_p) \sim \mathcal{N}_T(\mu_p, \sigma_p^2)$, where μ_p is given by the subjective measure of confidence in the proof. In Figure 1(b), $\pi(C_p) \sim \mathcal{N}_T(0.9, 0.01)$ would be interpreted, for instance, as there being *a priori* "very high" confidence in the proof of correctness to be supplied as evidence.

(iii) $\pi(C_{cc}|C_p, C_s) \sim \mathcal{N}_T(\mu_{cc}, \sigma_{cc}^2)$ 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. is given in a similar way. The BN, as shown in Figure 1(b), completely specifies the prior confidence in the overall argument; whereas the prior confidence to be expected in the claim, given the prior distribution of the parent r.v., is computed as $\{\text{high}\} \leftrightarrow \mathcal{N}_T(0.7257, 0.0145)$.

IV. Discussion

We have identified several challenges in quantifying confidence in a safety argument as presented; they are mainly relevant to validating the model used for quantifying confidence. First, we believe that justifying the basic BN structure and the assumptions of conditional independence could be achieved, in part, by automatically generating the BN from the GSN-based safety argument, where for each source of uncertainty identified, a corresponding node (or nodes) exists in the BN. Next, specifying leaf node probabilities and the prior CPD for the relevant intermediate/root nodes is straightforward, where empirical data is available. When only subjective judgment is available, quantifying confidence and selecting an appropriate prior distribution is problematic despite extensive research on belief elicitation methods [6].

We believe that one way to address this issue is to identify metrics using techniques such as the Goal-Question-Metric (GQM) method [13] and to correlate these metrics to confidence levels based on a defined quality model, e.g., we hypothesize that a metric such as coverage (by a safety argument) of hazards (in a hazard list) would correlate with the confidence in the sufficiency of the argument.

Finally, we require greater investigation to justify the weights used in specifying CPD requires. Assuming that the strategies used to decompose goals are viewed as being equally important, using equal weights appears to be a reasonable way forward.

V. Conclusions

Our preliminary investigation has emphasized the importance of treating assurance in an integrated way through linking qualitative safety arguments to quantitative arguments about uncertainty and confidence. This integration reaps the benefits of GSN in clearly communicating safety arguments to the many stakeholders of the safety case, while ensuring rigor in measuring confidence via probabilistic reasoning using BNs. We believe that when integrated into an engineering processes, the safety arguments in this approach will influence the development, assessment and management activities, whereas the confidence arguments will influence the level of rigor required in these activities to achieve the desired level of confidence in the safety arguments.

Acknowledgments

This work has been funded by NASA contract NNA10DE83C.

References

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

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

[3] M. Stamatelatos *et al.*, "Probabilistic risk assessment," NASA OSMA, Procedures and Guide for NASA managers and practitioners 1.1, Aug. 2002.

[4] C. H. Cave, "An Independent Review Into the Broader Issues Surrounding the Loss Of The RAF Nimrod MR2 Aircraft XV230 In Afghanistan in 2006," The Stationary Office, Tech. Rep., 2006.

[5] R. Hawkins, T. Kelly, J. Knight, and P. Graydon, "A new approach to creating clear safety arguments," in *Proc. Safety Critical Systems Symp.*, Feb. 2011.

[6] 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 Trans. Soft. Eng. (Article in Press)*, vol. PP, no. 99, 2010.

[7] R. Bloomfield, B. Littlewood, and D. Wright, "Confidence: its roles in dependability cases for risk assessment," in *Proc. 37th Int. Conf. Dependable Systems and Networks*, 2007.

[8] 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 Trans. Soft. Eng.*, vol. 33, no. 5, pp. 347–365, May 2007.

[9] C. Menon, R. Hawkins, and J. McDermid, "Interim standard of best practice on software in the context of DS 00-56 Issue 4," SSEI, University of York, Standard of Best Practice Issue 1, 2009.

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

[11] E. Denney and S. Trac, "A software safety certification tool for automatically generated guidance, navigation and control code," in *IEEE Aerospace Conf. Electronic Proc.*, Big Sky, Montana: IEEE, 2008.

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

[13] V. Basili, G. Caldiera, and D. Rombach, "Goal question metric approach," in *Encyclopedia of Soft. Eng.* John Wiley, 1994, pp. 528–532.