A Framework for Assessing Safety Argumentation Confidence

Rui Wang, Jérémie Guiochet, Gilles Motet

A Framework for Assessing Safety Argumentation Confidence

Rui Wang, Jérémie Guiochet, Gilles Motet

LAAS-CNRS, Université de Toulouse, CNRS, INSA,UPS, Toulouse, France
{firstname.surname}@laas.fr

Abstract. Software applications dependability is frequently assessed through degrees of constraints imposed on development activities. The statement of achieving these constraints are documented in safety arguments, often known as safety cases. However, such approach raises several questions. How ensuring that these objectives are actually effective and meet dependability expectations? How these objectives can be adapted or extended to a given development context preserving the expected safety level? In this paper, we investigate these issues and propose a quantitative approach to assess the confidence in assurance case. The features of this work are: 1) fully consistent with the Dempster Shafer theory; 2) considering different types of arguments when aggregating confidence; 3) a complete set of parameters with intuitive interpretations. This paper highlights the contribution of this approach by an experiment application on an extract of the avionics DO-178C standard.

Keywords: Dependability; Confidence assessment; Assurance case; Goal Structuring Notation; Belief function theory; DO-178C

1 Introduction

Common practices to assess the software system dependability can be classified in three categories [12]: quantitative assessment, prescriptive standards, and rigorous arguments. *Quantitative assessment* of software system dependability (probabilistic approach) has always been controversial due to the difficulty of probability calculation and interpretation [13]. *Prescriptive standard* is a regulation for software systems required by many government institutions. Nevertheless, in these standards, little explanations are given regarding to the justification and rationale of the prescriptive requirements or techniques. Meanwhile, the prescriptive standards limit to great extent the flexibility of system development process and the freedom for adopting alternative approaches to provide safety evidence. *Rigorous argument* might be another approach to deal with the drawbacks of quantitative assessment and prescriptive standard. It is typically presented in an assurance case [12]. This kind of argumentation is often well structured and provides the rationale how a body of evidence supports that a system is acceptably safe in a given operating environment [2]. It consists of

the safety evidence, objectives to be achieved and safety argument. A graphical argumentation notation, named as Goal Structuring Notation (GSN), has been developed [10] to represent the different elements of an assurance case and their relationships with individual notations. Figure 1 provides an example that will be studied later on. Such graphical assurance case representation can definitely facilitates the reviewing process. However, it is a consensus that safety argument is subjective [11] and uncertainties may exist in safety argument or supporting evidence [9]. Therefore, the actual contribution of safety argument has to be evaluated.

A common solution for assessing the safety argument is to ask an expert to judge whether the argument is strong enough [1]. However, some researchers emphasize the necessity to qualitatively assess the confidence in these arguments and propose to develop a confidence argument in parallel with the safety argument [9]. Besides, various quantitative assessments of confidence in arguments are provided in several works (using the Bayesian Networks [5], the belief function theory [3], or both [8]). In the report [7], authors study 12 approaches for quantitative assessments of confidence in assurance case. They study the flaws and counterarguments for each approaches, and conclude that whereas quantitative approaches for confidence are of high interest, no method is fully applicable. Moreover, these quantitative approaches lack of tractability between assurance case and confidence assessment, or do not provide clear interpretation of confidence calculation parameters.

The preliminary work presented in this paper is a quantitative approach to assess the confidence in a safety argument. Compared to other works, we take into account different types of inference among arguments and integrate them in the calculation. **We also provide calculation parameters with intuitive interpretation in terms of confidence in argument, weights or dependencies among arguments.** Firstly, we use GSN to model the arguments; then, the confidence of this argumentation is assessed using the belief function theory, also called the Dempster-Shafer theory (D-S theory) [4, 15]. Among the uncertainty theories (including probabilistic approaches), we choose the belief function theory, as it is particularly well-adapted to explicitly express uncertainty and calculate human's belief. This paper highlights the contribution of assessing the confidence in safety argument and the interpretation of each measurement, by studying an extract of the DO-178C standard as a fragment of an assurance case.

2 DO-178C Modeling

DO-178C [6] is a guidance for the development of software for airborne systems and equipment. For each Development Assurance Level (from DAL A, the highest, to DAL D, the lowest), it specifies objectives and activities. An extract of objectives and activities demanded by the DO-178C are listed in Table 1. There are 9 objectives. The applicability of each objective depends on the DAL. In Table 1, a black dot means that "the objective should be satisfied with indepen-

Table 1. Objectives for "verification of verification process" results, extracted from the DO-178C standard [6]

		Objective		Activity	Applicability by Software Level				Output		Control Category by Software Level			
		Description	Ref	Ref	A	B	C	D	Data Item	Ref	A	B	C	D
G2	1	Test procedures are correct.	6.4.5.b	6.4.5	●	○	○		Software Verification Results	11.14	②	②	②	
G2	2	Test results are correct and discrepancies explained.	6.4.5.c	6.4.5	●	○	○		Software Verification Results	11.14	②	②	②	
G3	3	Test coverage of high-level requirements is achieved.	6.4.4.a	6.4.4.1	●	○	○	○	Software Verification Results	11.14	②	②	②	②
G3	4	Test coverage of low-level requirements is achieved.	6.4.4.b	6.4.4.1	●	○	○		Software Verification Results	11.14	②	②	②	
G4	5	Test coverage of software structure (modified condition/decision coverage) is achieved.	6.4.4.c	6.4.4.2.a 6.4.4.2.b 6.4.4.2.d 6.4.4.3	●				Software Verification Results	11.14	②			
G4	6	Test coverage of software structure (decision coverage) is achieved.	6.4.4.c	6.4.4.2.a 6.4.4.2.b 6.4.4.2.d 6.4.4.3	●	●			Software Verification Results	11.14	②	②		
G4	7	Test coverage of software structure (statement coverage) is achieved.	6.4.4.c	6.4.4.2.a 6.4.4.2.b 6.4.4.2.d 6.4.4.3	●	●	○		Software Verification Results	11.14	②	②	②	
G4	8	Test coverage of software structure (data coupling and control coupling) is achieved.	6.4.4.d	6.4.4.2.c 6.4.4.2.d 6.4.4.3	●	●	○		Software Verification Results	11.14	②	②	②	
G4	9	Verification of additional code, that cannot be traced to Source Code, is achieved.	6.4.4.c	6.4.4.2.b	●				Software Verification Results	11.14	②			

dence", i.e. by an independent team. White dots represent that "the objective should be satisfied" (it may be achieved by the development team) and blank ones mean that "the satisfaction of objectives is at applicant's discretion".

This table will serve as a running example for all the paper. The first step is to transfer this table into a GSN assurance case. In order to simplify, we will consider that this table is the only one in the DO-178C to demonstrate the top goal : "Correctness of software is justified". We thus obtain the GSN presented in Figure 1. S1 represents the strategy to assure the achievement of the goal. With this strategy, G1 can be broken down into sub-claims. Table 1 contains 9 lines relative to 9 objectives. They are automatically translated into 9 *solutions* (Sn1 to Sn9). These objectives can be achieved by three groups of activities: reviews and analyses of test cases, procedures and results (Objectives 1 and 2), requirements-based test coverage analysis (Objectives 3 and 4), and structure coverage analysis (Objectives 5 to 9). Each activity has one main objective, annotated by G2, G3 and G4 in Table 1, which can be broken down into sub-objectives. In Figure 1, G2, G3 and G4 are the sub goals to achieve G1; meanwhile, they are directly supported by evidence Sn1 to Sn9. As this paper focuses on the confidence assessment approach, the other elements in GSN (such as *context*, *assumption*, etc.) are not studied here, which should be also considered for a complete study.

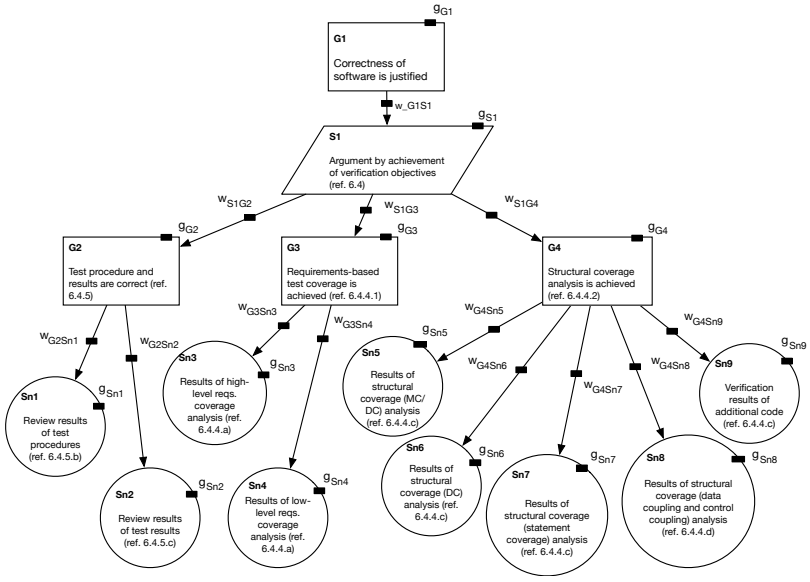

Fig. 1. GSN model of a subset of the DO-178C objectives

3 Confidence Assessment with D-S Theory

3.1 Confidence Definition

We consider two types of confidence parameters in an assurance case, which are similar to those presented in [9] named "appropriateness" and "trustworthiness", or "confidence in inference" and "confidence in argument" in [8]. In both cases, a quantitative value of confidence will lead to manage complexity of assurance cases. Among uncertainty theories (such as probabilistic approaches, possibility theory, fuzzy set, etc.), we avoid to use Bayesian Networks to express this value, as it requires a large number of parameters, or suffers from a difficult interpretation of parameters when using combination rules such as Noisy OR/Noisy AND. We propose to use the D-S theory as it is able to explicitly express uncertainty, imprecision or ignorance, i.e., "we know that we don't know". Besides, it is particularly convenient for intuitive parameter interpretation.

Consider the confidence g_{Snx} in a Solution Snx. Experts might have some doubts about its trustworthiness. For instance, the solution Sn2 "review results of test results" might not be completely trusted due to uncertainties in the quality of the expertise, or the tools used to perform the tests. Let X be a variable taking values in a finite set Ω representing a *frame of discernment*. Ω is composed of all the possible situations of interest. In this paper, the binary frame of discernment is $\Omega_X = \{\bar{X}, X\}$. An opinion about a statement X is assessed with 3 measures coming from DS-Theory: *belief* $(bel(X))$, *disbelief* $(bel(\bar{X}))$, and the *uncertainty*. Compared to probability theory where $P(X) + P(\bar{X}) = 1$, in the D-S theory a third value represents the uncertainty. This leads to $m(X) + m(\bar{X}) + m(\Omega) = 1$ (*belief + disbelief + uncertainty = 1*). In this theory, a mass $m(X)$ reflects the

degree of belief committed to the hypothesis that the truth lies in X. Based on D-S theory, we propose the following definitions:

$$\begin{cases} bel(\bar{X}) = m(\bar{X}) = f_X \text{ represents the disbelief} \\ bel(X) = m(X) = g_X \text{ represents the belief} \\ m(\Omega) = 1 - m(X) - m(\bar{X}) = 1 - g_X - f_X \text{ represents the uncertainty} \end{cases} \tag{1}$$

where $g_X,\ f_X \in [0, 1]$.

3.2 Confidence Aggregation

As introduced in Equation 1, the mass g_X is assigned for the belief in the statement X. When X is a premise of Y, interpreted as "Y is supported by X" (represented with a black arrow in Figure 1, from a statement X towards a statement Y), we assigned another mass to this inference which is (note that we use $m(X)$ for $m(X = true)$):

$$m((\bar{X}, \bar{Y}), (X, Y)) = w_{YX} \tag{2}$$

This mass actually represents the "appropriateness" i.e. the belief in the inference *"Y is supported by X"* (i.e. the mass of having Y false when X is false, and Y true when X true). Using the the Dempster combination rule [15], we combine the two masses from Equations 1 and 2 to obtain the belief (result is quite obvious but detailed calculation is given in report [16]):

$$bel(Y) = m(Y) = g_X \cdot w_{YX}$$

Nevertheless, in situations with 2 or more premises supporting a goal (e.g. G3 is supported by Sn3 and Sn4), we have to consider the contribution of the combination of the premises. Additionally to the belief in the arguments as introduced in Equation 1 ($m_1(X) = g_X$ and $m_2(W) = g_W$ where m_1 and m_2 are two independent sources of information), we have to consider a third source of information, m_3 to express that each premise contributes alone to the overall belief of Y, or in combination with the other premises. Let us consider that X and W support the goal Y, and use the notation (W, X, Y) for the vector where the three statements are true, and $(*, X, Y)$ when W might have any value (we do not know its value). We then define the weights:

$$\begin{cases} m_3((\bar{W}, *, \bar{Y}), (W, *, Y)) = w_{YW} \\ m_3(*, \bar{X}, \bar{Y}), (*, X, Y)) = w_{YX} \\ m_3((\bar{W}, \bar{X}, \bar{Y}), (\bar{W}, X, \bar{Y}), (W, \bar{X}, \bar{Y}), (W, X, Y)) = 1 - w_{YW} - w_{YX} = d_Y \end{cases} \tag{3}$$

where $w_{YW}, w_{YX} \in [0, 1]$, and $w_{YW} + w_{YX} \leq 1$.

The variable d_Y actually represents the contribution of the combination (similar to an AND gate) of W and X to the belief in Y. We propose to use this value as the assessment of the dependency between W and X to contribute to belief in Y, that is, the common contribution of W and X on demand to achieve Y. In this paper we will use three values for dependency, $d_Y = 0$ for independent premises,

$d_Y = 0.5$ for partial dependency, and $d_Y = 1$ for full dependency. At this step of our study, we did not find a way to extract from expert judgments a continuous value of d. Examples of interpretation of these values are given in next section. We then combine m_1, m_2 and m_3 using the DS rule (complete calculation and cases for other argument types are presented in report [16]):

$$bel(Y) = m(Y) = g_Y = d_Y \cdot g_X \cdot g_W + w_{YX} \cdot g_W + w_{YW} \cdot g_X \tag{4}$$

Where $g_W, g_X, w_{YX}, w_{YW} \in [0,1]$, $d_Y = 1 - w_{YX} - w_{YW} \in [0,1]$.

When applied to G2, we obtain:

$$g_{G2} = d_{G2} \cdot g_{Sn1} \cdot g_{Sn2} + w_{Sn1} \cdot g_{Sn1} + w_{Sn2} \cdot g_{Sn2} \tag{5}$$

Furthermore, a general equation (6) is obtained for goal Gx supported by n solutions Sni. The deduction process is consistent with D-S Theory and its extension work [14]:

$$g_{Gx} = d_{Gx} \cdot \prod_{i=1}^{n} g_{Sni} + \sum_{i=1}^{n} g_{Sni} \cdot w_{GxSni} \tag{6}$$

Where $n > 1$, $g_{Sni}, w_{Sni} \in [0,1]$, and $d_{Gx} = 1 - \sum_{i=1}^{n} w_{Sni} \in [0,1]$.

4 DO-178C Confidence Assessment

In the GSN in Figure 1, black rectangles represent belief in elements (g_{Sni}) and weights on the inferences (w_{GiSni}). The top goal is "Correctness of software is justified" and our objective is to estimate the belief in this statement. The value of dependency between argument (d_{Gi}) are not presented in this figure for readability. In order to perform a first experiment of our approach, we propose to consider the belief in correctness of DAL A software as a reference value 1.We attempt to extract from Table 1, the expert judgment of their belief in an objective to contribute to obtain a certain DAL. Table 1 is then used to calculate the weight (w_{GiSni}), belief in elements (g_{Sni}) and dependency (d_{Gi}).

4.1 Contributing Weight (w_{GiSni})

We propose to specify the contributing weights (w_{YX}), based on an assessment of the effectiveness of a premise X (e_X) to support Y. When several premises support one goal, their dependency (d_Y) is also used together to estimate the contributing weights. Regarding G2, Sn1 and Sn2 are full dependent arguments, as confidence in test results rely on trustworthy test procedures , i.e., $d_{G2} = 1$. d_{G3} for Sn3 and Sn4 is estimated over a first phase to 0.5. For structural coverage analysis (G4), the decision coverage analysis and the MC/DC analysis are extensions to the statement coverage analysis. Their contribution to the correctness of software is cumulative, i.e., $d_{G4}= 0$. Similarly, in order to achieve the top objective (G1), the goals G2, G3 and G4 are independent, i.e., $d_{G1}= 0$.

For each DAL, objectives were defined by safety experts depending on their implicit belief in technique effectiveness. For each objective, a recommended applicability is given by each level (dot or not dot in Table 1), as well as the external implementation by an independent team (black or white dot). Ideally, all possible assurance techniques should be used to obtain a high confidence in the correctness of any avionics software application. However, practically, a cost-benefit consideration should be regarded when recommending activities in a standard. Table 1 brings this consideration out showing that experts considered the effectiveness of a technique, but also its efficiency.

Only one dot is listed in the column of level D: "Test coverage of high-level requirements is achieved". This objective is recommended for all DALs. We infer that, for the given amount of resource consumed, this activity is regarded as the most effective one. Thus, for a given objective, the greater the number of dots is, the higher is the belief of experts. Hence, we propose to measure the effectiveness (e_X) in the following way: each dot is regarded as 1 unit effectiveness; and the effectiveness of an objective is measured by the number of dots listed in the Table 1. Of course, we focus on the dots to conduct an experimental application of our approach, but a next step is to replace them by expert judgment.

Based on rules in the D-S Theory, the sum of dependency and contributing weights is 1. Under this constraint, we deduced the contributing weights of each objective from its normalized effectiveness and the degree of dependency (see Table 2).

4.2 Confidence in Argument (g_i)

Coming back to Table 1, the black dot, which means the implementation of the activity needs to be deemed by another team, implies higher confidence in achieving the corresponding objective. The activities marked with the white dot are conducted by the same developing team, which give relatively lower confidence in achieving the goal. In order to calculate a reference value of 1 for the DAL A, we specify that we have a full confidence when the activity is implemented by an independent team ($g_{Sni} = 1$), an arbitrary value of 80% confidence when the activity is done by the same team ($g_{Sni} = 0.8$), and no confidence when the activity is not carried out ($g_{Sni} = 0$, see the g_{Sni} example for DAL B in Table 2).

4.3 Overall Confidence

Following the confidence aggregation formula given in Section 3.2, the confidence in claim G1 ("Correctness of software is justified") on DAL B is figured out as g_{G1} in Table 2. Objective 5 and 9 are not required for DAL B. Thus, we remove Sn5 and Sn9, which decrease the confidence in G4.

We perform the assessment for the four DAL levels. The contributing weights and dependency (w_{GiSni}, w_{G1Gi} and d_{Gi}) remain unchanged. The confidence in each solution depend on the verification work done by internal or external team. The different combinations of activities implemented within the development

Table 2. Confidence assessment for DAL B

	G1								
	G2		G3		G4				
	Sn1	Sn2	Sn3	Sn4	Sn5	Sn6	Sn7	Sn8	Sn9
g_{Sni}	0.8	0.8	0.8	0.8	0	1	1	1	0
e_{Sni}	3	3	4	3	1	2	3	3	1
d_{Gi}	1		0.5		0				
w_{GiSni}	0	0	2/7	1.5/7	1/10	1/10	2/10	3/10	1/10
e_{Gi}	6		7		10				
d_{G1}	0								
w_{G1Gi}	6/23		7/23		10/23				
g_{G1}	0.7339								

Table 3. Overall belief in system correctness

DAL	A	B	C	D
g_{DAL_x}	1	0.7339	0.5948	0.1391

team or by an external team provide different degrees of confidence in software correctness. Table 3 gives the assessment of the confidence deduced from the DO-178C, with a reference value of 1 for DAL A.

Our first important result is that compared to failure rates, such a calculation provides a level of confidence in the correctness of the software. For instance, the significant difference between confidence in C and D, compared to the others differences, clearly makes explicit what is already considered by experts in aeronautics: level A, B and C are obtained through costly verification methods, whereas D may be obtained with lower efforts. Review of test procedures and results (Objectives 1,2), components testing (Objective 4) and code structural verification (statement coverage, data and control coupling) (Objectives 7,8) should be applied additionally to achieve the DAL C. The confidence in correctness of software increases from 0.1391 to 0.5948. From DAL C to DAL B, decision coverage (Objective 6) is added to code structural verification and all structural analysis are required to be implemented by an independent team.

5 Conclusion

In this paper, we provide a contribution to the confidence assessment of a safety argument, and as a first experiment we apply it to the DO-178C objectives. Our first results show that this approach is efficient to make explicit confidence assessment. However, several limitations and open issues need to be studied. The estimation of the belief in an objective (g_X), its contribution to a goal (w_{YX}) and the dependency between arguments (d_Y) based on experts opinions is an important issue, and needs to be clearly defined and validated through several experiments. We choose here to reflect what is in the standard considering the black and white dots, but it is surely a debating choice, as experts are

required to effectively estimate the confidence in arguments or inferences. This is out of the scope of this paper. The dependency among arguments is also an important concern to make explicit expert judgment on confidence. As a long-term objective, this would provide a technique to facilitate standards adaptation or extensions.

References

1. Anaheed Ayoub, Jian Chang, Oleg Sokolsky, and Insup Lee. Assessing the overall sufficiency of safety arguments. In *21st Safety-Critical Systems Symposium (SSS'13)*, pages 127–144, 2013.
2. Peter Bishop and Robin Bloomfield. A methodology for safety case development. In *Industrial Perspectives of Safety-Critical Systems*, pages 194–203. Springer, 1998.
3. Lukasz Cyra and Janusz Gorski. Support for argument structures review and assessment. *Reliability Engineering & System Safety*, 96(1):26–37, 2011.
4. Arthur P Dempster. New methods for reasoning towards posterior distributions based on sample data. *The Annals of Mathematical Statistics*, pages 355–374, 1966.
5. Ewen Denney, Ganesh Pai, and Ibrahim Habli. Towards measurement of confidence in safety cases. In *Empirical Software Engineering and Measurement (ESEM), 2011 International Symposium on*, pages 380–383. IEEE, 2011.
6. DO-178C/ED-12C. Software considerations in airborne systems and equipment certification, 2011. RTCA/EUROCAE.
7. Patrick J Graydon and C Michael Holloway. An investigation of proposed techniques for quantifying confidence in assurance arguments. 2016.
8. Jérémie Guiochet, Quynh Anh Do Hoang, and Mohamed Kaaniche. A model for safety case confidence assessment. In *Computer Safety, Reliability, and Security (SAFECOMP)*, pages 313–327. Springer, 2015.
9. Richard Hawkins, Tim Kelly, John Knight, and Patrick Graydon. A new approach to creating clear safety arguments. In *Advances in systems safety*, pages 3–23. Springer, 2011.
10. Tim Kelly. *Arguing Safety - A Systematic Approach to Safety Case Management*. PhD thesis, Department of Computer Science, University of York, 1998.
11. Tim Kelly and Rob Weaver. The goal structuring notation–a safety argument notation. In *Proceedings of the Dependable Systems and Networks (DSN) workshop on assurance cases*, 2004.
12. John Knight. *Fundamentals of Dependable Computing for Software Engineers*. CRC Press, 2012.
13. Emmanuel Ledinot, Jean-Paul Blanquart, Jean Gassino, Bertrand Ricque, Philippe Baufreton, Jean-Louis Boulanger, Jean-Louis Camus, Cyrille Comar, Hervé Delseny, and Philippe Quéré. Perspectives on probabilistic assessment of systems and software. In *8th European Congress on Embedded Real Time Software and Systems (ERTS 2016)*, 2016.
14. David Mercier, Benjamin Quost, and Thierry Denœux. Contextual discounting of belief functions. In *Symbolic and Quantitative Approaches to Reasoning with Uncertainty*, pages 552–562. Springer, 2005.
15. Glenn Shafer. *A mathematical theory of evidence*, volume 1. Princeton university press Princeton, 1976.
16. Rui Wang, Jérémie Guiochet, Gilles Motet, and Walter Schön. D-S theory for argument confidence assessment. In *The 4th International Conference on Belief Functions, Prague, CZ*. Springer, 2016.