An Evidential Reasoning Approach for Assessing Confidence in Safety Evidence

Sunil Nair[*], Neil Walkinshaw[+], Tim Kelly[Ψ] and Jose Luis de la Vara[§]
[*]Institute for Energy Technology, Norway
[+]University of Leicester, United Kingdom
[Ψ]University of York, United Kingdom
[§]Carlos III University of Madrid, Spain
sunil.nair@ife.no, n.walkinshaw@mcs.le.ac.uk, tim.kelly@york.ac.uk, jvara@inf.uc3m.es

***Abstract*—Safety cases present the arguments and evidence that can be used to justify the acceptable safety of a system. Many secondary factors such as the tools used, the techniques applied, and the experience of the people who created the evidence, can affect an assessor's confidence in the evidence cited by a safety case. One means of reasoning about this confidence and its inherent uncertainties is to present a 'confidence argument' that explicitly justifies the provenance of the evidence used. In this paper, we propose a novel approach to automatically construct these confidence arguments by enabling assessors to provide individual judgements concerning the trustworthiness and the appropriateness of the evidence. The approach is based on *Evidential Reasoning* and enables the derivation of a quantified aggregate of the overall confidence. The proposed approach is supported by a prototype tool (EviCA) and has been evaluated using the Technology Acceptance Model.**

Keywords—safety case; safety evidence; confidence argument; uncertainity; evidential reasoning; expert judgement

I. INTRODUCTION

Goal-based system safety standards such as DS 00-56 [1] often require the construction and provision of a safety case. A safety case is defined as "*a structured argument, supported by a body of evidence, that provides a compelling, comprehensible and valid case that a system is safe for a given application in a given environment"* [1]. A structured safety argument explains how the available body of evidence supports the overall claim of acceptable safety.

Inevitably, both an argument and its supporting evidence are typically imperfect. 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. A recent survey on the state of the practice of evidence management [2] suggests that expert judgement is the most commonly used basis for assessing safety evidence.

Determining the type or amount of evidence required to satisfy a claim can be difficult. Both the developer and the assessor may be uncertain about certain characteristics of the evidence. As uncertainty in the evidence to support a given claim increases, the confidence in the overall safety case diminishes.

In current practice, the reasons for establishing confidence (or lack thereof) in the evidence and the uncertainties associated to an assessment often remains implicit in the assessment process [3]. In addition, the assessment process can vary substantially from expert to expert, and important assessments are frequently based on subjective evaluations [3]. These subjective beliefs are often based on many factors, including the process of the evidence creation, the techniques used, the people involved, and certain characteristics of the evidence itself (e.g., its role in the argument).

One approach of reasoning about the confidence established in the safety evidence is to build an explicit secondary *confidence argument* [4]. The role of the confidence argument is to explicitly detail the various reasons for having confidence in the evidence. It communicates the reasoning by identifying the various assurance deficits (uncertainties) related to the evidence and explicitly manage them such that the overall confidence in the safety argument is considered acceptable.

In this paper, we propose a novel approach to automatically construct confidence arguments for the evidence cited in a primary safety argument, and show how confidence can be quantified by employing a Multi-Criterion Decision Analysis technique known as Evidential Reasoning (ER) [5]. The paper makes the following specific contributions:

- A confidence argument pattern that explicitly details the various reasons for having confidence in a singular piece of evidence. The pattern is derived from confidence factors [3] and checklists used in industry for evidence assessment.
- A technique, based upon ER [5], by which the low-level confidence information on individual factors can be propagated to a higher-level safety claim. The technique (1) explicitly details the reasons for having confidence in the evidence, (2) captures uncertainties associated with the evidence assessment, and (3) presents the confidence at each level both quantitatively and visually.
- An implementation of the approach as an Eclipse plugin named *EviCA (**Evi**dence **C**onfidence **A**ssessor*).
- A user-evaluation of the proposed approach and its tool support, using the Technology Acceptance Model [6].

The remainder of the paper is organized as follows. Section II presents the background. Section III describes the proposed approach, including the confidence argument and confidence quantification using ER. Section IV presents the *EviCA* tool. Section V presents the evaluation. Section VI presents related work and finally, Section VII presents conclusions and future work.

II. BACKGROUND

This section provides a brief overview of the relevant background information required to understand our proposed approach.

A. Safety Cases and Confidence Arguments

The definition of a safety case presented in the introduction highlights that both arguments and evidence are crucial elements of a safety case that must go hand-in-hand. An argument without supporting evidence is unfounded, and therefore unconvincing. Evidence without argument is unexplained – i.e. it can be unclear whether (or how) safety claims have been substantiated. An explicit argument is required in order to communicate the relationship between the evidence and safety objectives.

The Goal Structuring Notation (GSN) [7] - a graphical argumentation notation – can be used to explicitly represent the individual elements of a safety argument (requirements, claims, evidence, and context) and the relationships that exist between these elements (i.e. how individual requirements are supported by specific claims, how claims are supported by evidence, and the context in which the argument is defined). When the elements of a GSN model are linked together in a network they are described as a '*goal structure*'. The principal purpose of any goal structure is to show how goals (claims about the system) are successively broken down into sub-goals until a point is reached where claims can be supported by direct reference to available evidence (solutions).

Hawkins et al. [4] propose that the arguments within safety cases can be usefully defined in terms of two separate but interrelated arguments:

- A *safety* (or *technical risk*) argument that documents the arguments and evidence used to establish direct claims of system safety.
- An accompanying *confidence* argument that justifies the sufficiency of confidence in the safety argument.

The technical risk argument must decompose the overall claim of acceptable safety into arguments that justify the acceptability of the risks posed by identified system hazards. For each hazard, the argument states what the 'adequately' addressed means are for mitigating that hazard and then identifies the evidence supporting the conclusion. This structure explains the purpose of each piece of evidence.

A confidence argument records the justification for confidence in a safety argument. There will be uncertainties associated with aspects of the safety argument or supporting evidence. The knowledge gaps that prohibit perfect (100%) confidence in a safety argument can be referred as 'assurance deficits'. In order to gain complete confidence in the evidence, these uncertainties must be identified and managed explicitly. The role of the confidence argument is to explicitly address the uncertainties by detailing the factors that establish confidence. Each time evidence is referenced as a solution in the technical risk argument, an assertion is being made that the evidence being put forward is sufficient to support the claim. The assurance of the solution (evidence) depends upon the confidence that the evidence is appropriate to support the claim and that the evidence is trustworthy.

B. Evidential Reasoning

Evidential Reasoning (ER) [5] provides a means by which to assimilate multiple assessments of individual facets of the evidence into a single, coherent macroscopic assessment. ER considers a hierarchy of 'attributes' by which the evidence is to be assessed. For example, as a higher-level attribute in a hierarchy, one might consider the quality of the personnel/team to be a factor, which could in turn be decomposed into sub-attributes such as competence or experience.

Formally, each attribute can be subdivided into a set E of lower-level sub-attributes $\{e_1, \cdots, e_n\}$. Each sub-attribute e_i can also be given a weight w_i representing the relative importance, such that $\sum_{i=1}^{n} w_i = 1$ (by default the weight is evenly distributed across all attributes as $1/n$).

A human expert assesses each of the lowest-level attributes by providing their assessment of the "quality" of the attribute. This is provided in terms of a Likert-scale consisting of g grades $H =< H_1, \cdots, H_g >$ (i.e. if $g = 5$, H_1 and H_5 might correspond to "very poor" and "excellent" respectively).

Instead of providing just a single value on this scale (implying that they are 100% certain of their assessment – e.g. that the experience of the team is 'excellent'), the assessment can be provided as a distribution. This distribution is referred to as *'Belief Function'* – a term that will be used throughout the paper. This belief function enables the assessors to capture any uncertainty that they might have about their assessment. For example, an assessor could indicate that they have a confidence of 50% that the team's level of experience is "excellent" and 50% that the level of experience is "good".

Formally, the expert´s confidence that a particular attribute e_i achieves a grade H_n is denoted $\beta_{n,i}$. For a given attribute, $\sum_{i=1}^{n} \beta_{n,i} \leq 1$. Thus, an expert´s complete assessment of attribute e_i (encompassing all possible grades) can be expressed as the distribution:

$$S(e_i) = \{(H_n, \beta_{n,i}), n = 1, \cdots, g\}$$

A key feature of ER is that, alongside uncertainty, it is also possible to capture complete ignorance on the part of the assessor. The beliefs in $\beta_{n,i}$ do not have to sum up to 1 for a given attribute (as would be expected with conventional Bayesian probabilities). The sum of beliefs $\sum_{i=1}^{n} \beta_{n,i}$ can be interpreted as their overall confidence of the assessment, where a sum of 1 amounts to total confidence, and a sum of 0 amounts to total ignorance (no confidence).

Given a hierarchy of attributes, where the lowest-level attributes are associated with distributions corresponding to the assessments as presented above, ER provides an algorithm by which to assimilate them. Distributions of 'beliefs' are propagated up from lower-level nodes to higher-level nodes, and are combined with the distributions from their sibling nodes to produce a representative macroscopic assessment.

Crucially, this process of propagation from basic attributes to an aggregated result obeys certain desirable axioms that ensure the following [5]:

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: $\sum_{i=0}^{n} \beta_{i,y} < 1$.

III. A Structured Approach For Constructing And Assessing Confidence Arguments

The proposed approach will (a) devise a GSN-based confidence argument structure that details the reason for establishing confidence in the evidence, and (b) use ER to reason about the confidence in the evidence and the assessment of the evidence. Though presented for GSN, the approach could readily be adapted to any technique that explicitly links evidence and safety claims, graphically (e.g., CAE) or textually (e.g., safety accomplishment report).

A. Overview of our Approach

Previous work [3] has demonstrated that there are many secondary factors that can affect the confidence in safety evidence. Given this (often broad) range of factors, which can often be conflicting, it is vital that any inherent uncertainty is identified and acknowledged. In our approach, we make this uncertainty explicit by defining a confidence argument pattern that enables the abstract notion of confidence to be broken down into confidence factors that can in turn be linked to the evidence. This break-down into individual confidence-factors enables us to gauge the assessor's doubt or uncertainty with respect to each atomic factor, via a series of questions based on evidence type-specific checklists.

As a running example, let us consider a primary argument claiming, "*Hazards related to System X have been identified and recorded"*. The claim is supported by citing the evidence *Hazard log* as an asserted solution. When evidence is cited in a safety case, it is typically asserted that the evidence presented is sufficient to support a claim. In reality the truth of this assertion depends on the appropriateness and trustworthiness of the evidence (in this case the hazard log).

Our first contribution is to generate a confidence argument structure to support primary safety arguments. In GSN an assertion in a safety argument is linked to a confidence argument by an *Assurance Claim Point* (ACP). An ACP is indicated in GSN with a named black rectangle on the relevant link, as shown in Fig. 1(a).

We build a confidence argument as a hierarchy by decomposing overall confidence into sub factors. Fig. 1(b) shows an example of how the overall confidence in the hazard log can be decomposed into individual factors (above the dashed line). For example, trustworthiness of the hazard log is broken into the process employed to create the hazard log, the personnel who carried out the activity, and the tools used to create the log. The personnel factor is further broken into competence and guarantee of independence (e.g., the personnel verifying the log being independent of the creators). The relative (example) importance of each factor is annotated on each edge.

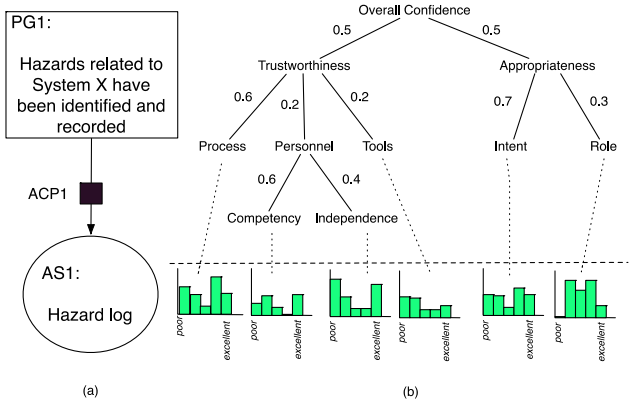

Fig. 1. Running example showing (a) primary argument with ACP, (b) confidence factors associated to the hazard log.

This decomposition is carried out until we can gauge the assessor's subjective confidence in the lowest level factors (below the dashed line) through a series of questions. The rationale is that it is often easier to provide confidence values to individual questions rather than to high-level factors such as trustworthiness. The responses from the assessors are calculated on a scale of (1) agreement to the question using Likert scale, (2) confidence in their own assessment (presented in detail in Section III.B).

The ER algorithm is then applied to combine the lower-level confidence values provided for each question, and propagate them up the hierarchy. This automatic propagation happens until the top most parent factor is reached yielding a general confidence assessment. With the use of GSN patterns, the end result is an argument structure that graphically depicts the various confidence factors associated to the evidence with a quantified confidence value based on lower-level assessments at each claim level.

Section III.B will elaborate on our development of the confidence argument pattern. Section III.C details our use of questionnaires to obtain inputs, and the use of ER to build an accurate overview of an assessor's confidence.

B. Confidence Argument Pattern

The confidence argument pattern is represented using the GSN pattern extensions [7][8][9]. We followed four steps to

build the confidence argument pattern. The first step involved determining the criteria that should be considered for confidence assessment. The results from the systematic analysis of the practice of evidence assessment (via surveys and interviews with safety experts) presented in [3] were used to gather a set of factors (e.g., the process and techniques used, and the personnel involved). As a second step, we specified an initial structure (based on the results of step 1) for the argument pattern. Our structure has two main criteria for assessing confidence in safety evidence: *trustworthiness and appropriateness* (discussed in detail below). These two criteria are widely accepted in the systems engineering domain [4][10][11]. Each criterion was broken down into subsequent specific factors that affect their confidence. All the authors reviewed the confidence argument pattern and discussed its structure and content. This resulted in an initial version of the pattern.

To improve the pattern's coverage of factors influencing assessor confidence, we analysed 16 checklists from the aerospace, avionics, railway, and defence domains (checklists are extensively used in the industry [2][3]). Most of the checklists (13 out of 16) are in the public domain[†]. As a result, we identified three factors (*Bound Qualification, Scope for Document format*, and *Expected Structure of the Evidence Type*) that had not been included in the initial pattern. These were consequently included in the argument pattern and discussed amongst the authors.

Fig. 2 shows our proposed confidence argument pattern expressed in GSN. The pattern is broken into two different parts: *1) trustworthiness* and *2) appropriateness*.

Relating to the running example, the top most goal (G1) of the pattern describes that there is sufficient *overall confidence* in the evidence used as *asserted solution* (hazard log). In our approach, we define confidence using GSN context element (Con1) as a measure of the belief that the evidence cited for a particular claim is trusted for its integrity and is appropriate for its intended purposes.

The trustworthiness and appropriateness factors can be decomposed as follows.

1) Trustworthiness

In the systems engineering domain, trustworthiness of evidence often relates to the data and processes used to generate the evidence [12]. In our approach, we define trustworthiness (Con2) as the property of the evidence to provide trust or belief that the evidence can be assured to be as specified. We decompose *trustworthiness* of the *asserted solution* as shown in Fig. 3 (Appendix B). The sub-factors are as follows:

***Personnel* (G4)** – Arguments regarding the personnel or the team(s) involved in the creation or verification of the asserted solution. When documenting how claims are supported by sub-claims, it is recommended to document the reasoning step. This is done in GSN by documenting the strategy of the argument that links the two goals. Therefore, using the strategy S2, arguments can be made over each personnel factor that influences the overall confidence. There is a set of pre-defined factors in the proposed pattern. In addition, as mentioned above, we allow addition of user-defined personnel factors (Con3).

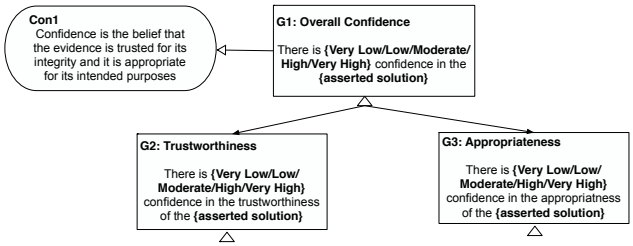

Fig. 2. Overall confidence argument pattern for the asserted solution

We make use of the multiplicity element in GSN to denote different user-defined personnel factors. For each personnel or team, we provide the following decomposition:

- *Past Knowledge* (G12) – the confidence that the personnel or team(s) involved have knowledge about the creation or verification of the asserted solution in a similar context in the past. The similar context (e.g., project information) should be explicitly shown in the argument structure by using the context Con14.
- *Competency* (G13) – the confidence that the personnel or team(s) involved in the creation or verification have the required skills, training and competency to produce the asserted solution. Any further evidence supporting this (e.g., developer resume) can be provided if available.
- *Independence* (G14) – the confidence that the different personnel or teams involved in the creation or verification of the asserted solution are independent of each other. This is to ensure that the person creating the hazard log is not the same as the one verifying the log.
- *Domain Experience* (G15) – the confidence that the personnel or team(s) involved in the creation or verification of the asserted solution have the required domain experience.

***Processes/Techniques (G5)* –** Arguments regarding the different processes or techniques employed to create or verify the asserted solution. Similar to the personnel (G4), arguments can be made over each pre-defined or user-defined process factors (S3). For each process or technique, we provide the following decomposition:

- *Past Use* (G17) – the confidence based on the past use of the same process or technique.
- *Definition* (G18) – the confidence that the process or technique has been clearly defined and the definitions have been followed.
- *Peer Review* (G19) – the confidence that the various outcomes of applying the process or technique have

[†] Public checklists have been collected and shared at https://drive.google.com/a/simula.no/folderview?id=0B42RvDI04vjnbzgtM0RXRFJqZWs&usp=drive_web

been peer reviewed (e.g. the hazard log has been peer reviewed at least once).

***Tool Integrity* (G6)** – Arguments regarding the integrity of the different tools used to create or verify the asserted solution. Arguments can be made over pre-defined or user-defined factors (S4). The following supporting tool integrity arguments are defined:

- *Bound Qualification* (G21) – the confidence that the specific usage of the different tools used for creation or verification of the evidence is within the constraints of its qualification, e.g. the tool used to record the hazard log was used in the context of the appropriate process or that the tool was configured appropriately.
- *Standard Qualification* (G22) – the confidence that the different tools were qualified in accordance to the safety standard used, e.g. DO-178C [13] can require tools to be qualified when used as part of the software assurance process. The safety standard in question should be made explicit in context Con15.

***Content Compliance* (S1)** – Arguments regarding the structural integrity and structural compliance of the asserted solution. The strategy is decomposed into two sub-goals:

- *Scope* (G24) – the confidence that the asserted solution has been scoped and defined according to the required document format to demonstrate compliance with a specific safety standard, e.g. all the terms and acronyms are defined in the hazard log. Scope arguments detail the structural integrity of the asserted solution. The safety standard (or reference document) in question must be made explicit (Con16).
- *Expected structure* (G25) – the confidence that the asserted solution conforms to the expected structure for the particular evidence type, e.g., hazard log structure as mandated by hazard specification. This detail the structural compliance of the asserted solution. The type of the evidence must be made explicit (Con 17).

***Evidence of Past Use* (G7)** – The confidence that the evidence was approved to be trustworthy in an earlier instance of use in a similar context. The similar context (e.g. the associated techniques and tools) should be made explicit here via the context Con7.

2) Appropriateness

The appropriateness of the evidence (hazard log) relates to the satisfaction of the claim (e.g., all hazards were identified and recorded) [10].

In our approach, we define appropriateness of the evidence as the property of the evidence to sufficiently corroborate the claim it was cited for. The type of evidence that is most appropriate can only be determined based on the nature of the claim and the argument that the evidence is intended to support. Hence, we decompose the *appropriateness* of the *asserted solution* (Fig. 4) as:

***Argument role* (G9)** – the confidence that the evidence *type* of the asserted solution is capable of providing the asserted role in the argument [10]. For example, the role of the hazard log is to identify hazardous functional failures that may occur in system X. The type of the evidence and the asserted role has to be made explicit using contexts Con8 and Con9 respectively.

***Intent* (G10)** – the confidence that the *specific* evidence satisfies the asserted role. For example, the intent of the hazard log to demonstrate that the system X does not contain errors that could manifest as hazards. The intent strengths the role of the asserted solution. The asserted role of the evidence has to be made explicit (Con10).

***User-Defined Appropriateness Factor* (G8, G11, G16, G20, G23 and G26)** – User-defined factors can be created at any level of the pattern. The required claim description, strategies related to the goal, and context description must be added to the pattern if required. Further decomposition can also be carried out.

Each of the lower level goals can be further decomposed if required. However, in the current approach we provide question prompts for each of the leaf claims. The process of collecting the individual belief functions for each of the lower-level claims and the confidence quantification is discussed in the following subsection.

C. Confidence Quantification and Combination with ER

The confidence argument pattern discussed above presents us with a hierarchy of factors, in terms of which we can assess confidence in the evidence. In this step we show how an assessor can feed their data into this hierarchy, and how ER can combine this data to provide a macroscopic picture of the assessment.

1) Collecting Assessment Data

The hierarchical nature of the confidence pattern means that each factor at a branch-node is composed of its lower-level nodes. Accordingly, the assessment of a parent-node (e.g. Appropriateness) is composed of the assessments of its child-nodes (Personnel, processes/techniques, etc.). It follows that if we can provide assessments for all of the *leaf*-nodes in the hierarchy, it is possible to propagate these assessments up the hierarchy, and so to yield an overall assessment.

As mentioned previously, we solicit input from the assessor by means of questionnaires. Each leaf-node is associated with a questionnaire that asks for two inputs: (1) An agreement to the question regarding aspects of the factor using Likert-scale (by default from 1 to 5, but this can be customised), (2) level of *confidence* in the capacity to answer the question (on a continuous scale from 0-100).

2) Combining Assessments with the ER Algorithm

Having answered the questions, we are left with a range of assessments for the `leaf' nodes.

The challenge now is to combine these, to provide aggregate assessments for higher-level factors in the confidence argument. This combination can be achieved by employing the ER algorithm [5].

To employ ER, each of the assessments provided for the leaf-node is encoded as a belief function (see Section II.B).

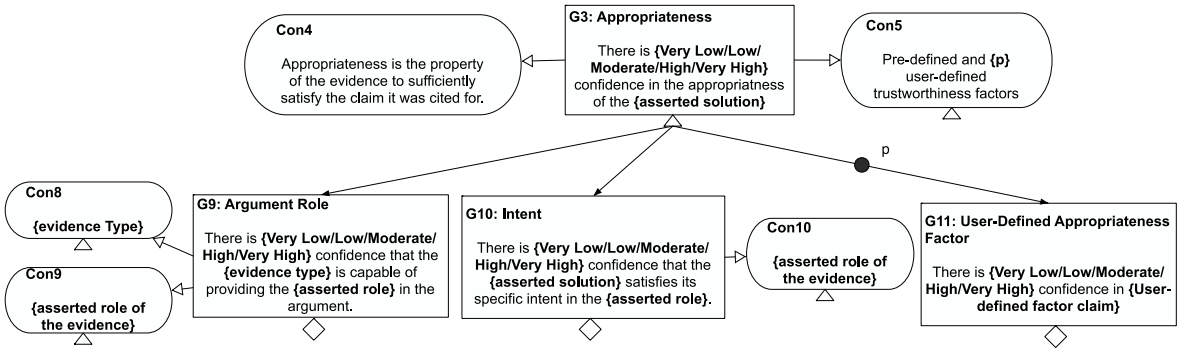

Fig. 4. Appropriateness argument pattern for the asserted solution

The algorithm then uses depth-first traversal to combine the belief-functions from the bottom-up, to populate the entire hierarchy. The following description will illustrate how an example set of lower-level attribute assessments is combined using ER. Details regarding the algorithm and the processing are provided in Section II.B, and can be found in [5][14].

Let us consider the example of the hazard log. In order to gain confidence in the evidence source, and more specifically in the *trustworthiness* of the *hazard log*, let us consider two factors related to the *Personnel/Teams* from Section III.B.1): *Independence* (I) among the personnel and *Competency* (C) of the personnel. Each of these two factors is further broken into two leaf-node attributes by means of two questions (Q) respectively. These lowest-level attributes form the basis for obtaining the assessor's atomic assessments that are to be aggregated to a high-level assessment of the hazard log. For this example, we consider some questions identified from the Railway Safety Commission checklist [15] that are used in practice to assess hazard logs.

- **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?

For each question $q \in Q$, the assessment consists of two parts:

1. The assessor's agreement a on the Likert scale $1 \geq a \geq 5$ where (for these questions) 1 represents *Definitely not*, *2* represents *No*, *3* represents *Maybe*, *4* represents *Yes*, and 5 represents *Absolutely*.
2. The assessor's confidence on his/her assessment $0 \leq c \leq 1$, where 0 represents no confidence at all (total ignorance), and 1 represents total confidence.

These answers are then used to construct a distribution $S(q) = \{(H_{n=a}, c), (H_{n\neq a}, 0), n = 1, \cdots, 5\}$. For example, the assessor might answer the questions (in the order listed above) as follows: (4, 0.8), (5, 1), (3, 0.5), (4, 0.8). In other words, he/she *agrees* with a certainty of 80% that the person doing the verification is different from the one responsible for developing the documents (q1), he/she *strongly agrees* with a certainty of 100% that the manager to whom the team reports is identified (q2), etc. The resulting distributions then look as follows: $[0,0,0,0.8,0], [0,0,0,0,1], [0,0,0.5,0,0], [0,0,0,0.8,0]$.

To begin with, the algorithm combines the two leaf-node attributes for *Independence* ($[0,0,0,0.8,0]$, $[0,0,0,0,1]$), setting the weights for each attribute to 0.5 (1/[number of attributes]). This results in the aggregate distribution for *Independence* [0,0,0,0.364,0.545]. The *doubt* is quantified explicitly as $1 - (0.364 + 0.545) = 0.091$. Similarly, combining the answers for *Competence* ($[0,0,0.5,0,0]$ and $[0,0,0,0.8,0]$) yields $[0,0,0.231,0.462,0]$, with an explicit doubt of 0.308. Now, to compute the final assessment of the *trustworthiness* of the hazard log, ER combines the two scales computed for *Independence* and *Competence* again ($[0,0,0,0.364,0.545]$ and $[0,0,0.231,0.462,0]$), to produce a final assessment of $[0,0,0.099,0.452,0.281]$, with an explicit uncertainty of 0.168 (16.8%).

This final distribution reflects the following two confidence values:

(1) The assessor's confidence in the factor – 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*. From the above final distribution, we obtain 9.9% of the mass is attributed to *Medium* confidence, 45.2% can be attributed to *High* confidence, and 28.1% can be attributed to *Very High* confidence. 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.

(2) The assessor's confidence in their assessment – 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*. In the above distribution the assessor has 16.8% uncertainty related to the assessment of the hazard log. This translates to 83.2% confidence in the assessment of the

trustworthiness of the hazard log, which represents *Very High* confidence.

To summarise, the final distribution indicates that (1) the assessor has *Medium* confidence in the *trustworthiness* of the hazard log, and (2) the assessor has *Very High* confidence in his/her assessment of the *trustworthiness* of the hazard log.

IV. Tool Support

In this section we briefly describe the prototype tool named EviCA (**Evi**dence **C**onfidence **A**ssessor), developed to support the proposed approach. Specifically, EviCA allows users to: (1) create and edit safety arguments using GSN, (2) question the various reasons for having confidence in the evidence used in a primary argument, (3) automatically structure confidence arguments based on a predefined GSN pattern that is customisable, and (4) calculate the confidence at each level of the pattern automatically using ER. EviCA is written in Java programming language as a plug-in to the Eclipse IDE. It uses some utilities of the underlying Eclipse framework, notably the Graphical Editing Framework (GEF).

A. Creating and Editing Safety Arguments

Fig. 5 shows a screenshot of a safety argument fragment built with EviCA. The pallet to the right of the screen provides users with the various GSN elements (goals, solutions, strategies, context, etc.) that they need to model safety cases. The properties of a selected item can be accessed at the bottom of the screen. The element description can be edited either in the properties window or directly on the canvas. The pane in the left of the window is a project explorer that displays the different projects and their associated safety case diagrams. To our knowledge, the GSN editor developed as part of EviCA is the first of its type that allows users to create and manipulate confidence arguments. Users can click and drag *Assertion Claim Points* (ACP) between goals and solutions.

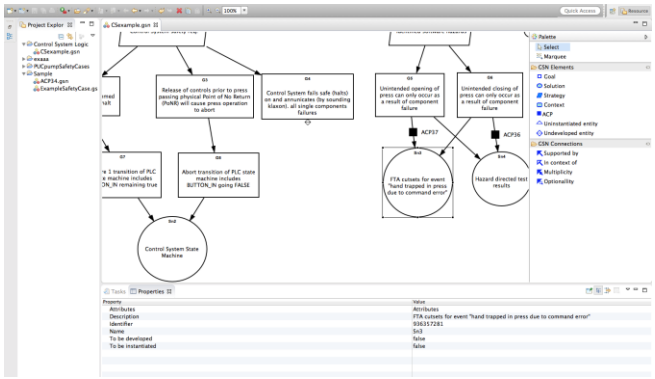

Fig. 5. EviCA's GSN editor screenshot with sample safety case fragment

B. Confidence Argument Generator

Double clicking on any ACP between a goal and an asserted solution opens the *Confidence Argument Generator* wizard (Fig. 6). This wizard guides the creation of the confidence argument based on the various confidence factors. It initially has a pre-defined tree structure with a set of pre-defined confidence factors (as presented in Section III.B). Right clicking on any parent factor allows users to edit the tree structure. Each element in the tree has a weight function. The user can define the weights for any factor in the tree. The sum of weights of any factor must not exceed 1. By default, the weights are equally split among all the children depending on the number of children.

C. Collecting Belief Functions

EviCA allows users to add individual questions to each confidence factor to obtain the belief function. There are two ways to add questions. Users can manually add one question at a time by right clicking the lowest child. This will bring up the *Add Question* window (Fig. 6). The user can define the weight of each question the same way as the parent.

The second way to add questions is with the help of the *Import Questions* button. EviCA allows users to import a set of checklist questions from Microsoft Excel. In an Excel-based checklist, the rows represent the different checklist items and the columns represent the confidence factors. A checklist item can be mapped to a confidence factor by marking an 'X' in the Excel file..EviCA then automatically imports the marked questions into their corresponding factors in the *Confidence Argument Generator* dialog. This way EviCA allows users to import large sets of questions at once with ease.

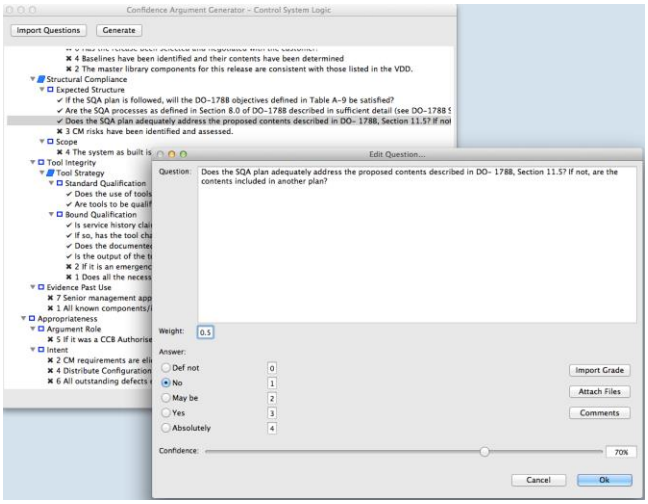

Fig. 6. Screenshot of confidence argument generator window and add question dialog

To obtain the belief function for each question, the user has to input two values: (1) an agreement grade and (2) a confidence in the assessment. By default, we use a five point Likert scale for the grade values for all the questions - *Definitely not (1), No (2), Maybe (3), Yes (4)* and *Absolutely (5)*. However, the user can change the default scale by selecting from a list of pre-defined scales (e.g., *Strongly disagree to Strongly agree*) using the *Import Grade* button.

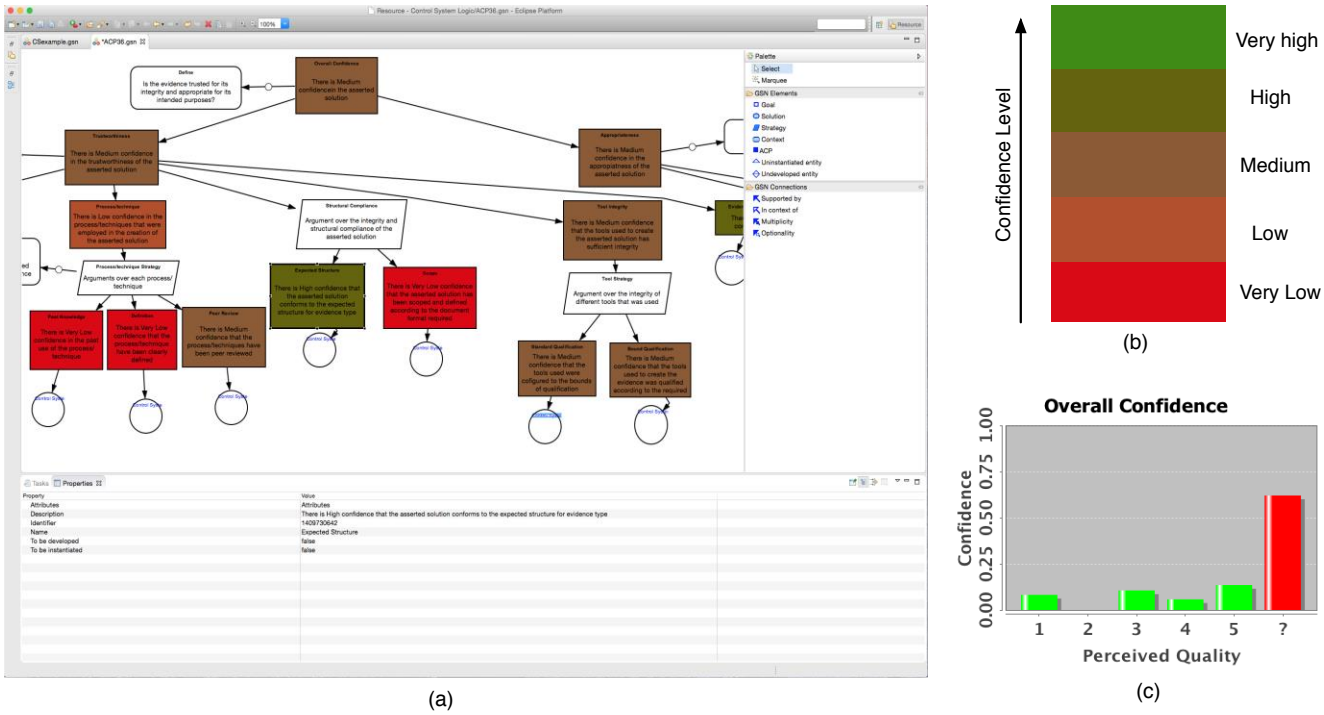

Fig. 7. Final output of EviCA (a) screenshot of final confidence argument fragment generated; (b) colour code used to represent confidence in the assessment; (c) a bar chart presenting the individual belief function for overall confidence with explicit uncertainty (in red)

For example, when answering a question regarding the competency of a developer, additional evidence such as the resume of the developer can be provided.

In addition, the users can justify their rationale behind a particular answer by providing further comments. Unanswered questions in the *Confidence Argument Generator* window are denoted by a ✕ on the left side, while answered questions are denoted by a ✓ (Fig. 6).

Once all the questions are answered, the user can click the *Generate* button to automatically build the confidence argument for the asserted solution with the confidence in the evidence and the assessment visually presented. Fig. 7(a) shows the fragment of the confidence argument structure built by EviCA. The structure is editable by the user.

To present the assessor's confidence on the evidence, EviCA uses the placeholders **"{0}"** in the goal description to be automatically replaced with the calculated confidence value tags ranging from *Very Low* to *Very High*. The answers to all the questions and the corresponding comments are computed and attached as solutions to the corresponding parent goal. Any additional evidence provided in the question is also added as a solution to the corresponding parent goal.

To present the assessor's confidence in the assessment, EviCA uses a colour scheme for representing different levels of confidence (shown in Fig. 7(b)). Additionally, EviCA provides individual bar charts of *belief function*s for each goal and solution node in the structure. Fig. 7(c) shows the distribution of values as individual bar chart for the overall confidence in the evidence. The values 1-5 correspond to the assessor's confidence in the evidence, 1 being *Very Low* and 5 being *Very High* (see Section III.B). The height of a bar represents the assessor's confidence in his/her assessment. An additional bar in red denotes the uncertainty or doubt inherent in the *overall confidence*. This is computed from those belief functions where the total sum of beliefs did not add up to 1.

V. Evaluation

In order to evaluate the user acceptance of our proposed approach and tool support, we adopted the Technology Acceptance Model (TAM) [6]. TAM focuses on three main facets of user acceptance:

- *Perceived Ease of Use:* degree to which a person believes that using a particular method would require less effort.
- *Perceived Usefulness:* subjective probability that using a particular system would enhance job performance.
- *Intention to Use:* extent to which a person intends to use a particular system.

An evaluation based on the above three facets is best suited for this scenario because the proposed technology is fundamentally improving an existing activity by making it more systematic, repeatable, and explicit. Since safety is a critical aspect of the system development, unless this technology is perceived as being feasible (e.g., the effort required) and useful (e.g., leading to better assessments) by the practitioners, it is unlikely that it will be explored further

with other validations (e.g., a case study). Using TAM allows us to evaluate whether the approach and the tool are appealing to practitioners (1) without intruding on the practitioners work settings and (2) without imposing any artificial experimental constraints that could skew their judgment.

The study targeted safety experts, who are directly involved in safety case development and safety evidence assessment for critical systems. Since the tool support is a direct implementation of the approach itself, we developed a short presentation of the proposed approach and a tool demo video[‡] showing the main features of EviCA. This allowed us to mitigate threats related to possible bias in the way the technology was presented to each participant. At the end of the presentation, the participants were asked to fill a short questionnaire with 12 questions.

The questionnaire was divided into four parts. The first section contained a short description of the aim of the survey and five background questions. The second section consisted of three questions regarding the participants perceived usefulness of the approach. The third section consisted of three questions regarding the perceived ease of use of the approach and the tool support. And finally, the last section consisted of one question regarding the intention of use of the approach and any further comments regarding the proposed approach. We used a five-point Likert Scale ranging from *Strongly Agree* to *Strongly Disagree* to collect answers for all questions. We also allowed participants to make additional comments for each question. Other than the first section, the remaining sections of the questionnaire were randomised. The entire questionnaire can be found in [10].

The presentation and the survey questionnaire were sent via personal email invitations and subsequent reminders to practitioners that we knew. We also asked them to let other colleagues know about the survey. Additionally, the presentation and the survey were posted on a social networking websites for people in professional occupations (http://www.linkedin.com).

A total of 21 participants provided their feedback on the approach. All had more than two years of experience in safety certification, assurance, or assessment. All also indicated that they assess safety evidence information as part of their work, or develop or assess safety cases.

Relating to the perceived usefulness of the approach and the tool support (Appendix A, Fig. 8), 86% of the participants agreed or strongly agreed that the ability to express ignorance or doubt is useful for the assessment of safety evidence, 85% agreed or strongly agreed that the range of safety evidence assessment factors covered in the approach is adequate, and 62% mentioned that they agree or strongly agree that the use of the approach will lead to more accurate safety evidence assessments. We acknowledge that the completeness of the confidence argument pattern for all types of evidence cannot be guaranteed. We built the pattern based on a small set of identified factors and checklists that are used in practice. These are restricted to a relatively narrow range of domains and standards. Nonetheless, our evaluation suggests that the presented structure covers all major areas of concern relating to evidence confidence assessment. To accommodate situations where the set of criteria are incomplete, our approach and tool support allows users to add new factors and criteria to the existing proposed pattern.

Regarding the perceived ease of use of the approach and the tool support (Appendix A, Fig. 9), 76% of the participants agreed or strongly agreed that it is easy to express any doubt or ignorance about a particular safety evidence assessment with the approach. Similarly, 71% of the participants strongly agreed that it is easy to customise a safety evidence assessment in the tool. Additionally, 76% of the participants agreed or strongly agreed that it is easy to interpret the results produced by the approach.

When asked about the intention of use of the approach and tool support (Appendix A, Fig. 10), 57% of the participants agreed or strongly agreed that they would use the approach for safety evidence assessment tasks if it were made available to them within their organization. Only 10% of the respondents disagreed.

In summary, the results show that the approach was generally viewed as being easy to understand and easy to use. Most of the participants also indicated that it would be advantageous to use the approach within their organizational context. Although the number of responses is limited, the evaluation indicates that practitioners consider that the proposed approach can be valuable in practice for safety evidence assessment. This gives us sufficient confidence to continue with a further, more extensive evaluation of the approach in the future (e.g. a case study).

VI. Related Work

Safety case development and assessment has been of much interest in research over the past years. One method to construct safety arguments using GSN is the *Six Step* method [16]. However, this does not explicitly consider the confidence of the safety argument and the confidence in the safety evidence [17]. Some works have provided various criteria and factors that should be considered to determine the confidence in safety evidence and arguments [18]. These have been complementary to our proposed confidence argument pattern.

A new approach for creating clear safety cases was introduced by Hawkins *et al*. [4]. Their approach suggests building a secondary confidence argument that explicitly states the reason for having confidence in the evidence cited. The paper presents an indicative argument pattern for the confidence argument that is based upon the identification and management of assurance deficits (uncertainties). However, this pattern is not exhaustive and does not identify specific confidence factors. Similar pattern for process-based arguments have also been proposed using model-driven

[‡] *https://www.youtube.com/watch?v=Rz_P0YlMPBU*

safety certification approaches [19]. By combining this approach of a secondary confidence argument with the various factors that influence the assessor's confidence, we propose a more exhaustive pattern that covers the specific reasons for having confidence in the evidence.

One approach similar to ours in terms of building a confidence model was proposed by Ayoub *et al*. [11]. It uses common concerns associated with an argument and systematically builds confidence arguments based on them. The limitation of their approach, as acknowledged by the authors, is it only covers only some of the *Trustworthiness* factors. Details as to how these factors were identified and the sources of their origin are not detailed in the paper. Moreover, other concerns related to *Appropriateness* are yet to be categorised.

Other attempts have been made to quantitatively measure confidence in safety cases. Bloomfield *et al*. [20] suggest that confidence depends upon uncertainty about the underpinnings of the dependability case, e.g. truth of assumptions, correctness of reasoning, and strength of evidence. Their paper discusses some challenges related to quantitative assessment of confidence and shows that confidence in a claim can sometimes be surprisingly low. Although this may be true, the paper does not provide detailed guidance on how to tackle this issue. Our approach builds on this work to provide a systematic guided process that considers an exhaustive list of confidence factors to quantify evidence and associated uncertainty.

Past studies have detailed the notion of uncertainty in safety cases [21][22] and provided ways to handle them e.g. using Bayesian Belief Networks (BBN) [23][24]. BBN rely heavily on their probability tables, which in turn rely on the availability of prior probability information. This is often difficult to obtain given the scarcity of priors, making it difficult to provide a thorough assessment on confidence when the assessor is doubtful of any of the factors. Our approach does not depend on prior probabilities and allows assessors to accommodate doubt or uncertainty in their assessment.

ER is an example of Dempster-Schäfer (DS) theory [25][26]. DS has been applied to a multitude of diverse `Multi-Criteria Decision Analysis problems' such as environmental impact assessments [27], assessment of weapons systems capabilities [28], and safety analysis [29]. An approach based on trust cases to support and improve expert assessment of arguments is proposed in [30]. Similar to our proposed approach, the trust case approach is based on DS theory and it provides a way to issue assessments and their aggregation depending on the types of inference used in arguments. What differentiates our work is the use of an explicit secondary argument structure for demonstrating the reason for having confidence in the evidence. DS theory and aggregated assessment alone may not suffice to demonstrate sufficient confidence in the assessment of the evidence or the argument. The approach proposed in this paper provides a systematic confidence argument pattern that allows explicit presentation of the various factors that provide confidence in the evidence. Our approach also aggregates the individual beliefs into a final assessment, and the final assessment can be broken down into lowest level reasons for having confidence and making uncertainties explicit.

VII. Conclusions

This paper has proposed a novel approach to automatically construct confidence arguments and quantify confidence using Evidential Reasoning. The approach enables safety experts to assess evidence by explicitly reasoning (1) the trustworthiness of the evidence creation and verification process and (2) the appropriateness of the evidence in a particular claim-argument context. The approach also allows experts to indicate their uncertainty or doubt related to the assessment. As part of the approach, we proposed a confidence argument pattern that details the various reasons for establishing confidence in the evidence. The confidence argument pattern decomposes the notion of overall confidence in the evidence into lower-level factors associated to the evidence. The approach then enables experts to provide individual belief functions for the lowest-level factors. With the help of the ER algorithm, the lowest-level belief functions are propagated to higher-level claims, until it provides an aggregate belief value for an overall confidence claim. The final result of the approach is an explicit confidence argument structure that visually and quantitatively presents the confidence in the evidence assessed and in the assessment of the evidence.

As a proof of concept, the proposed approach has been implemented as a prototype tool named EviCA (Evidence Confidence Assessor). The proposed approach and the tool support were also evaluated using the Technology Acceptance Model, by conducting a survey with safety experts who are directly involved in safety case development and evidence assessment. A total of 21 experts responded to questions relating to the perceived use of the approach, perceived usefulness of the approach and future intention to use the approach. Most of the participants perceived the approach to be useful and easy to use, and indicated that they would use the approach and tool support if it were available in their organizational context. Therefore, we believe that the proposed approach can be valuable in practice for the assessment of safety evidence.

As an immediate future work, we will further validate the proposed approach in an industrial case study. We will further validate the completeness of the confidence argument pattern with more checklists used in practice. We will also develop a classification of questions for each confidence factor in the pattern for different evidence types. Finally, our work will seek to improve the tool support, eliminate fallacies (if any introduced) and mitigate fatigue when answering the questionnaire for the various factors.

Acknowledgments

The research leading to this paper has received funding from the FP7 programme under the grant agreement n° 289011 (OPENCOSS) and from the Research Council of Norway under the project Certus-SFI.

REFERENCES

[1] Interim Defence Standard 00-56 Part 1 - Issue 5, in, UK MOD (2014)

[2] S. Nair, J. de La vara, M. Sabetzadeh, and D. Falessi.: Evidence Management for Compliance of Critical Systems with Safety Standards: A Survey on the State of Practice, Technical report, Simula Research Laboratory (2014)

[3] S. Nair, T. Kelly, and M Jørgensen.: Understanding the practice of Safety Evidence Assessment: A Qualitative Semi-Structured Interview Study, Technical report, Simula Research Laboratory (2014)

[4] R. Hawkins, T. Kelly, J. Knight, and P. Graydon.: A new approach to creating clear safety arguments. In Advances in Systems Safety (pp. 3-23) (2011)

[5] J.B. Yang, and D.L. Xu.: On the evidential reasoning algorithm for multiple attribute decision analysis under uncertainty, IEEE Transactions on Systems, Man, and Cybernetics, Part A, vol. 32,no. 3 (2002)

[6] F.D. Davis.: A technology acceptance model for empirically testing new end-user information systems: Theory and results. Diss. Massachusetts Institute of Technology (1985).

[7] GSN Committee.: Draft GSN Standard. Version 1.0 (2010).

[8] T. Kelly, and J. McDermid.: Safety case construction and reuse using patterns. *Safe Comp 97*. Springer London, 1997. 55-69.

[9] T. Kelly.: Arguing safety – a systematic approach to managing safety cases. PhD thesis. Department of Computer Science, University of York (1998)

[10] R. Hawkins, and J. McDermid.: Software Systems Engineering Initiative, SSEI-TR-0000041, *Software Safety Evidence Selection and Assurance*, Issue 1, University of York, October 2009.

[11] A. Ayoub, BG. Kim, I. Lee, and O. Sokolsky.: A systematic approach to justifying sufficient confidence in software safety arguments. *Computer Safety, Reliability, and Security*. Springer Berlin Heidelberg, 2012. 305-316.

[12] I. Habli, and T. Kelly.: Achiveing intergrated process and product safety arguments, Proceddings of 15th Safety Critical Systems Symposium (2007).

[13] DO-178C/ED-12C, Software Considerations in Airborne Systems and Equipment Certification. (2012).

[14] N. Walkinshaw.: Using evidential reasoning to make qualified predictions of software quality. *Proceedings of the 9th International Conference on Predictive Models in Software Engineering*. ACM, 2013.

[15] RSC Guidelines, Railways Safety Commission, http://www.rsc.ie/publications/rscguidelines.html (accessed September 2014).

[16] T. Kelly.: A six-step Method for Developing Arguments in the Goal Structuring Notation (GSN). Technical report. York Software Engineering, UK (1998)

[17] R. Hawkins, and T. Kelly.: Software Safety Assurance – What Is Sufficient?. In: 4th IET International Conference of System Safety (2009

[18] C. Menon, R. Hawkins, and J. McDermid.: Defence standard 00-56 issue 4: Towards evidence-based safety standards. In: Safety-Critical Systems: Problems, Process and Practice, pp. 223–243. Springer, London (2009)

[19] B. Gallina.: A Model-Driven Safety Certification Method for Process Compliance, Software Reliability Engineering Workshops (ISSREW), pp. 204,209. IEEE. (2014)

[20] R. Bloomfield, B. Littlewood, and D. Wright.: Confidence: Its Role in Dependability Cases for Risk Assessment. In: 37th Annual IEEE/IFIP International Conference on Dependable Systems and Networks, DSN 2007, pp. 338–346 (2007)

[21] E. Denney, and G. Pai.: A lightweight methodology for safety case assembly. In Computer Safety, Reliability, and Security (pp. 1-12). Springer Berlin Heidelberg (2012).

[22] R. Weaver, T. Kelly, and P. Mayo.: Gaining confidence in goal-based safety cases. In Developments in Risk-based Approaches to Safety (pp. 277-290) (2006)

[23] E. Denney, G. Pai, and I. Habli.: Towards measurement of confidence in safety cases. In IEEE ESEM, pp. 380-383 (2011)

[24] W. Weihang, and T. Kelly.: Combining bayesian belief networks and the goal structuring notation to support architectural reasoning about safety. *SAFECOMP* Springer Berlin Heidelberg, 2007. 172-186.

[25] A.P. Dempster.: A generalization of Bayesian inference, Journal of the Royal Statistical Society, Series B, vol. 30, pp. 205–247 (1968)

[26] G. Shafer.: A Mathematical Theory of Evidence. Princeton University Press, (1976)

[27] Y. Wang, J.B. Yang, and D.L. Xu.: Environmental impact assessment using the evidential reasoning approach. *European Journal of Operational Research* 174.3 (2006): 1885-1913.

[28] J. Jiang, X. Li, Z. Zhou, D. Xu, and Y. Chen.: Weapon system capability assessment under uncertainty based on the evidential reasoning approach. *Expert Systems with Applications* 38.11 (2011): 13773-13784.

[29] J. Wang, J. B. Yang, and P. Sen.: Safety analysis and synthesis using fuzzy sets and evidential reasoning. *Reliability Engineering & System Safety* 47.2 (1995): 103-118.

[30] L. Cyra, and J. Górski.: Expert assessment of arguments: A method and its experimental evaluation. *Computer Safety, Reliability, and Security*. Springer Berlin Heidelberg, 2008. 291-304.

APPENDIX A: Evaluation Results

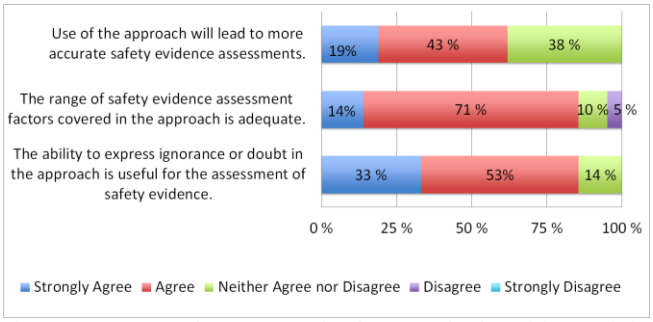

Fig. 9. Percentage of responses related to perceived usefulness of the approach and the tool support

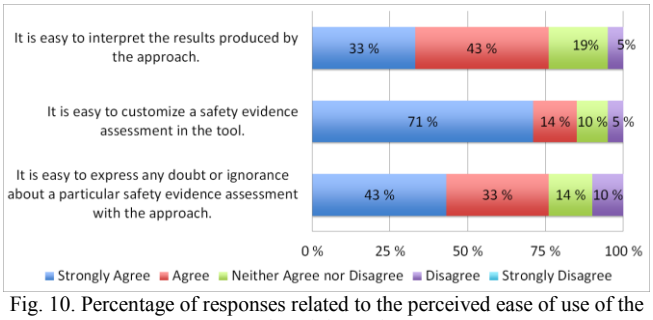

Fig. 10. Percentage of responses related to the perceived ease of use of the approach and the tool support

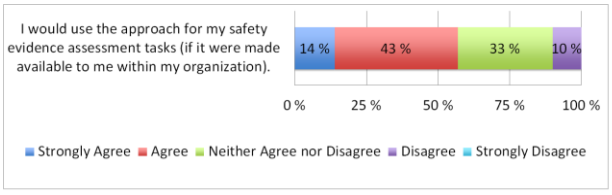

Fig. 11. Percentage of responses related to the intention of use of the approach and tool support

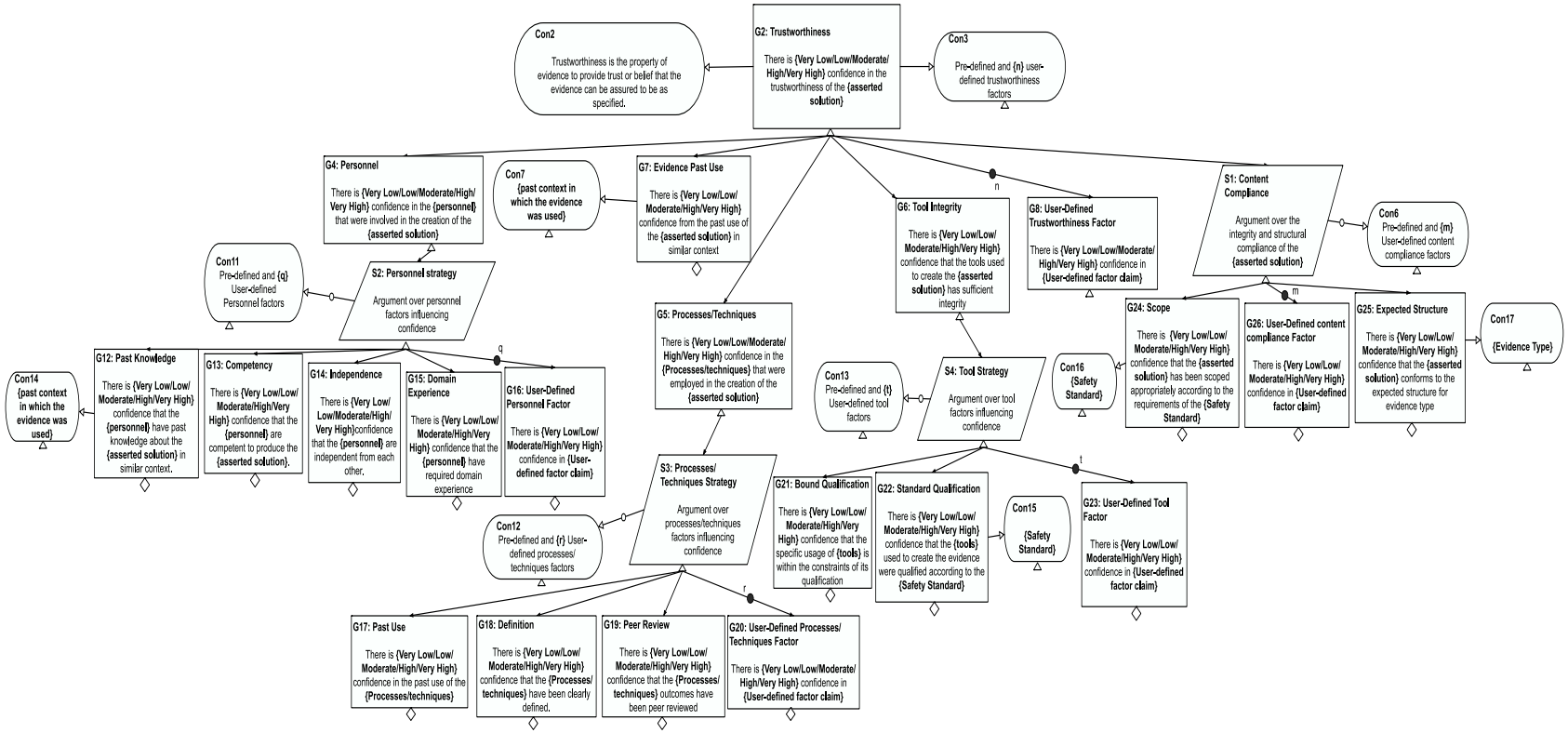

Figure 3. Trustworthiness argument pattern for the asserted solution