Assessing the Overall Sufficiency of Safety Arguments

Anaheed Ayoub, Jian Chang, Oleg Sokolsky and Insup Lee

Computer and Information Science Department, University of Pennsylvania

Philadelphia, PA, USA

Abstract Safety cases offer a means for communicating information about the system safety among the system stakeholders. Recently, the requirement for a safety case has been considered by regulators for safety-critical systems. Adopting safety cases is necessarily dependent on the value added for regulatory authorities. In this work, we outline a structured approach for assessing the level of sufficiency of safety arguments. We use the notion of basic probability assignment to provide a measure of sufficiency and insufficiency for each argument node. We use the concept of belief combination to calculate the overall sufficiency and insufficiency of a safety argument based on the sufficiency and insufficiency of its nodes. The application of the proposed approach is illustrated by examples.

1 Introduction

A safety assurance case presents an argument, supported by a body of evidence, that a system is acceptably safe to be used in the given context (Menon et al. 2009). Recently, the U.S. Food and Drug Administration (FDA) has been highlighting the upcoming call for certain 510(k) medical device submissions to include safety assurance cases (FDA 2010). Adopting safety cases necessarily requires the existence of proper reviewing mechanisms.

Safety cases are, by their nature, often subjective (Kelly 2007). The objective of safety case development, therefore, is to facilitate mutual acceptance of this subjective position. The goal of safety case evaluation, therefore, is to assess if there is a mutual acceptance of the subjective position. We define the safety argument assessment as answering a question about the overall sufficiency of the argument, i.e., are the premises of the argument 'strong enough' to support the conclusions being drawn. The simplest way to assess a safety argument is to ask an expert reviewer to evaluate the overall sufficiency of the argument. Although this is a commonly accepted practice, most probably the final decision is not accurate enough. Research in experimental psychology shows that the human mind does not deal properly with complex inferences based on uncertain sources of

knowledge (Cyra and Gorski 2008a), which is common in safety arguments. Therefore, reviewers should only be required to express their opinions about the basic elements in the safety argument. Then a mechanism should provide a way to aggregate the reviewer opinions to communicate a message about the overall sufficiency of the safety argument. A potential problem with evaluating the safety arguments lies in psychology and the notion of a *mindset*. A mindset is a set of assumptions, methods or notations held by one or more people or groups of people which is so established that it creates a powerful incentive within these people or groups to continue to adopt or accept prior behaviours, choices, or tools (Wikipedia 2012b). An important component of mindset is the concept of *confirmation bias*. Confirmation bias is a tendency for people to favour information that confirms their preconceptions or hypotheses regardless of whether the information is true (Leveson 2011). Confirmation bias is a prime example of a mindset that can produce defective decision making. The problem of the confirmation bias is not easy to eliminate. But it can be reduced by changing the goal. In other words, the reviewer should take the opposite goal: try to show that the provided safety argument is insufficient to support the system safety conclusion. We propose assessing the safety argument insufficiency as well as its sufficiency.

In this paper, we outline a structured method for assessing the level of sufficiency and insufficiency of safety arguments (Section 4). The reviewer assessments and the results of their aggregation are represented in the Dempster-Shafer model (Sentz and Ferson 2002). We use the notion of *basic probability assignment* (also referred to as a *degree of belief* or a *mass*) to provide a measure of the degree of belief in the sufficiency and insufficiency of each argument node to do its role. For example, a mass of the sufficiency and insufficiency of an evidence node Ev, which is directly addressing a conclusion node n, is a measure of the degree of belief in the sufficiency and insufficiency of Ev to support n.

We propose *aggregation rules* (Section 3) to calculate the mass of the overall sufficiency and insufficiency of a safety argument by aggregating the mass of the sufficiency and insufficiency of the safety argument nodes. The selection of the appropriate aggregation rule is discussed in Section 4.1. The assessing of the missing support (if any) and its impact on the degree of beliefs is given in Section 4.2. The application of the proposed method is illustrated by a complete example in Section 5. The related work is discussed in Section 6, and Section 7 concludes the paper.

2 Safety cases

Recently, safety cases are being explored as ways for communicating ideas and information about the safety-critical systems among the system stakeholders. The manufactures submit safety cases (to present a clear, comprehensive and defensible argument supported by evidence) to the regulators to show that their products are acceptably safe to operate in the intended context (Kelly 1999). There are dif-

ferent approaches to structure and present safety cases. The Goal Structuring Notation (GSN) is one of the description techniques that have been proven to be useful for constructing safety cases (Kelly and Weaver 2004). In this work, we use the GSN notation in presenting safety cases. In GSN a top-level goal (i.e., conclusion) is decomposed into sub-goals through implicit or explicit strategy, and eventually supported by evidence. In this paper, we use the term *conclusion* to describe the relation between any goal, sub-goal or strategy and its child nodes. Also we use the term *supporting node* to describe how any sub-goal, strategy or evidence node is related to its parent node. For example, Strategy S1 in Figure 7 is a conclusion for G2 and G3, and a supporting node for G1.

A new approach for creating clear safety cases is introduced in (Hawkins et al. 2011). This new approach basically separates the major components of the safety cases into safety argument and confidence argument. A safety argument is limited to give arguments and evidence that directly target the system safety. A confidence argument is given separately to justify the sufficiency of confidence in this safety argument. The separation between safety and confidence related aspects facilitates the development and reviewing processes for safety cases. In this paper, we introduce a structured mechanism to assess the overall sufficiency and insufficiency of safety arguments. Confidence arguments should be used by the reviewer to make his/her assessments on the sufficiency and insufficiency of the evidence nodes.

3 Aggregation rules

We define the safety argument assessment as answering the question about the overall sufficiency of the argument, i.e., are the premises of the argument 'strong enough' to support the conclusions being drawn. The first step of the proposed assessment procedure is to ask the reviewer to express his/her opinion about the basic elements in the safety argument. Then a systematic mechanism is used to aggregate the reviewer opinions to communicate a message about the overall sufficiency of the safety argument. The process of assessing the argument basic elements is nothing but a *decision-making* process. Decision making can be regarded as the mental processes resulting in the selection of a course of action among several alternative scenarios (Wikipedia 2012a). Every decision making process produces a final choice. In the case of evaluating the safety argument node, the output would be a choice either the node under evaluation is sufficient or not. According to psychologists, a natural phenomenon known as *confirmation bias* contaminates the mental process of the decision-making. Confirmation bias is a tendency for people to favour information that confirms their preconceptions or hypotheses regardless of whether the information is true (Leveson 2011). For example, in the case of evaluating the safety argument nodes, the reviewer may have an existing belief that the submitter of the safety argument is trusted based on his reputation, and probably the system is safe based on past success. In this case, the reviewer will choose to focus on the facts that conform to his/her existing belief. If the reviewer is not careful, his/her mind that is biased toward confirming the safety ar-

gument sufficiency would prevent him/her from seeing any contrary evidence that is actually there. Consequently, the reviewer decision would be that the node is sufficient; however this may not be the truth. This is one of the main problems of the Nimrod safety case (NSC) (Haddon-Cave 2009), the safety case was built and reviewed with the mindset that the system is safe based on past success. Consequently, NSC failed to identify the design flaws that led to the total loss of the Nimrod. The case of NSC shows how the consequences of confirmation bias to decision making are serious. In order to fight the confirmation bias the reviewer should do the opposite; assess the safety argument insufficiency. Although assessing the insufficiency fights the confirmation bias, it increases the vulnerability to negative confirmation bias. The existing belief in case of negative confirmation bias would be that the node is insufficient. To fight confirmation bias and negative confirmation bias, we propose that the reviewer assesses his/her belief in the sufficiency and insufficiency of the argument nodes. This forces the reviewer to evaluate the node from two different perspectives; the positive one (the node is sufficient) and the negative one (it is insufficient). In addition, the reviewer may not be able to precisely determine if the node is sufficient or insufficient, meaning that he/she has a belief that this node could either be sufficient or insufficient. In other words, the reviewer describes his/her belief in the node sufficiency and insufficiency, and then the gap between these two beliefs presents the uncertainty.

3.1 Dempster-Shafer theory

Dempster-Shafer Theory (DST) is a mathematical theory of evidence (Sentz and Ferson 2002). It offers an alternative to traditional probabilistic theory for the mathematical representation for uncertainty, which is required for the safety arguments assessment. DST is a potentially valuable tool when knowledge is obtained from expert's elicitation, which is the case of safety arguments assessment. We use the Dempster-Shafer model to present the reviewer assessments and the results of their aggregation. The most important part of DST is *basic probability assignment (BPA)*, also referred to as a *degree of belief* or a *mass*. Let x be a finite set known as frame of discernment. In the safety argument assessment, x = *{Sufficient, Insufficient}*. The power set $P(x)$ is the set of all subsets of x including itself and null set ϕ. For x = *{Sufficient, Insufficient}* then $P(x) = \{\phi$, *{Sufficient}, {Insufficient}, {Sufficient, Insufficient}}*. The *BPA*, represented by m, defines a mapping from every subset of the power set to interval between 0 and 1. Formally, $m{:}P(x) \rightarrow [0, 1]$. The *BPA* satisfies the following two conditions:

- The *BPA* of the null set is 0. Formally, $m(\phi) = 0$.
- The summation of the *BPA's* of all the subsets of the power set is 1. That is, $\sum_{A \in P(x)} m(A) = 1$, where A is a set in the power set ($A \in P(x)$).

Every set in the power set of the frame of discernment which has mass > 0 is a focal element (i.e., hypotheses). For the safety argument assessment, $P(x)$ has three focal elements: hypothesis S = *{Sufficient}* that the node is sufficient, hy-

pothesis *I* = *{Insufficient}* that it is insufficient, and (universe) hypothesis *U* = *{Sufficient, Inefficient}* that the node is either sufficient or insufficient. For each argument node n, $m_n(S)$ and $m_n(I)$ represent the degree of belief in the sufficiency and insufficiency of n. And $m_n(U)$ represents the uncertainty; n is either sufficient or insufficient, where $m_n(S) + m_n(I) + m_n(U) = 1$.

Example 1. Suppose the reviewer checked the confidence argument/measure provided for node n, and got a belief of 0.5 that n is sufficient ($m_n(S) = 0.5$). However, he/she got a belief in n insufficiency with a degree 0.2 ($m_n(I) = 0.2$). The remaining mass of 0.3, which is the gap between the 0.5 supporting n sufficiency on one hand and the 0.2 for n insufficiency on the other hand is 'indeterminate'. Which means that n could either be sufficiency or insufficient ($m_n(U) = 0.3$).

One of the main attractions of DST is the availability of a rule to combine the data obtained from multiple sources. It allows one to combine evidence from different sources and arrive at a degree of belief. This is the case of addressing a conclusion supported by different nodes, where each supporting node arrives with a degree of belief. DST is based on the assumption that these supporting nodes are independent. The original combination rule of multiple *BPA's* is known as Dempster's rule of combination (Voorbraak 2012).

Definition 1. Given two *BPA's* m_1 and m_2, the combination (called joint *m*) is calculated from the aggregation of m_1 and m_2 as:

$$m(\phi) = 0$$

$$m(C) = m_1(C) \oplus m_2(C) = \frac{\sum_{A \cap B = C} m_1(A) * m_2(B)}{1 - K} \text{ where } C \neq \phi$$

$$K = \sum_{A \cap B = \phi} m_1(A) * m_2(B) .$$

The denominator, $1 - K$, is a normalization factor, which is a measure of the amount of conflict between the two supporting nodes. For the safety argument assessment, *C* is *S*, *I*, or *U*. The Dempster's rule of combination is commutative and associative. For the simple case shown in Figure 1, a conclusion node n1 (i.e., GSN goal, sub-goal, or strategy node) is addressed by two supporting nodes n2 and n3 (i.e., GSN sub-goal, strategy, or evidence nodes). The pairs for *A* and *B* sets for which $A \cap B = S$ are (*S* and *S*), (*S* and *U*), and (*U* and *S*). The pairs for *A* and *B* sets for which $A \cap B = I$ are (*I* and *I*), (*I* and *U*), and (*U* and *I*). The pairs for *A* and *B* sets for which $A \cap B = \phi$ are (*S* and *I*), and (*I* and *S*).

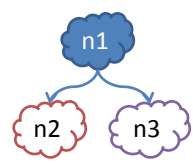


Fig. 1. A simple case

The following definition, which we will frequently use in the rest of the paper, is the application of Definition 1 to the simple case given in Figure 1.

Definition 2.

$$m_{n1}(S) = m_{n2}(S) \oplus m_{n3}(S)$$
$$= \frac{m_{n2}(S) * m_{n3}(S) + m_{n2}(S) * m_{n3}(U) + m_{n2}(U) * m_{n3}(S)}{1 - K}$$

$$m_{n1}(I) = m_{n2}(I) \oplus m_{n3}(I)$$
$$= \frac{m_{n2}(I) * m_{n3}(I) + m_{n2}(I) * m_{n3}(U) + m_{n2}(U) * m_{n3}(I)}{1 - K}$$

$$K = m_{n2}(S) * m_{n3}(I) + m_{n2}(I) * m_{n3}(S).$$

There is a known problem with the normalization of Dempster's rule of combination as explained in (Zadeh 84). We avoid this issue because the elements of our sets (S, *I* and *U*) are not independent.

3.2 Mixing

In addition to using Dempster's rule of combination, we also use a mixing (i.e., weighted averaging) rule.

Definition 3. The mixing rule: $m(C) = \frac{w_1 * m_1(C) + w_2 * m_2(C)}{w_1 + w_2}$, where w_1 and w_2 are the weights assigned to the supporting nodes. These weights represent the coverage of the conclusion by each supporting node. The weighted averaging rule of combination is commutative and associative. The following definition, which we will frequently use in the rest of the paper, is the application of Definition 3 to the simple case given in Figure 1.

Definition 4. $m_{n1}(C) = \frac{w_1 * m_{n2}(C) + w_2 * m_{n3}(C)}{w_1 + w_2}$, where w_1 and w_2 represent the coverage of n1 by n2 and n3 respectively. The use of mixing rules with Dempster-Shafer structures is justified in (Sentz and Ferson 2002).

Definitions 1 and 3 give the general form of the aggregation rules. Definitions 2 and 4 are the case of instantiation to the safety argument assessment settings for two supporting nodes. Discussions for the criteria to select between the aggregation rules (Definition 2 and Definition 4), and how to apply these rules to aggregate the degree of belief in the sufficiency and insufficiency of the argument nodes, are given in Section 4.1.

4 Assessment mechanism

The move toward using safety cases in certification and regulation requires a way to review them. Safety cases are, by their nature, often subjective (Kelly 2007). The goal of safety case evaluation, therefore, is to assess if there is a mutual acceptance of the subjective position.

We propose an assessment method to assess the overall sufficiency and insufficiency of safety arguments. This method consists of the two steps shown in Algorithm 1. The sufficiency and insufficiency of the top goal is the overall sufficiency and insufficiency of the safety argument.

Algorithm 1. Assessment procedure

Step 1. Evidence assessment

- Estimate the 'sufficiency' and 'insufficiency' of each evidence node to address the goal it is used to support. These estimations express the degree of the reviewer belief in the sufficiency and insufficiency of the evidence (e.g., *Ev*) to support its goal. We use the notion of the *BPA* to represent these estimations. Formally, $m_{Ev}(C)$, where $C \in \{S, I, U\}$.

Step 2. Automatic aggregation

- Starting from the leaves of the safety argument, apply the next steps for each conclusion node.
  - *Aggregate* the estimates of the supporting nodes to obtain the degree of belief in the sufficiency and insufficiency of the conclusion (Section 4.1).
  - *Recalculate* the degree of belief in the sufficiency and insufficiency of the conclusion, in case of identifying missing supports (Section 4.2).
- Repeat the process until the top goal has been reached.

Confidence. The reviewer should use the confidence arguments (Ayoub et al. 2012, Hawkins et al. 2011) or any confidence measure (Bloomfield et al. 2007, Denney et al. 2011) in the evidence assessment.

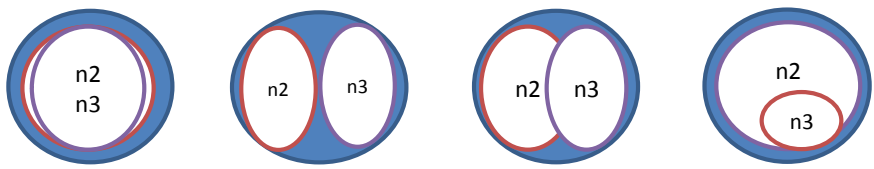
Assumptions. We assume that the safety argument is understandable, free from structural errors (e.g., no circular arguments), fully connected (i.e., no dangling evidence or unsupported goal), sufficiently expressed (e.g., no missing context), and contains no conflicts (e.g., assumptions attached to *all* the argument nodes are consistent). These assumptions can be satisfied by applying the step-by-step reviewing mechanism (Kelly 2007) before running the proposed assessment method. More discussion of the complementary use of these two methods is given in Section 6.

4.1 Argument types

Note that the degree of belief in the support given to a conclusion by its supporting nodes depends on the kind of inference used to derive the conclusion. We therefore begin by characterizing inference types (i.e., argument types). We use the case shown in Figure 1 to define the main argument types. Let n1 be a goal node claiming that {*the system satisfies 10 safety requirements; SRs 1-10*}. Table 1 shows examples of different argument types for the same conclusion. For the case shown in Figure 1, we distinguish four argument types (see Figure 2).

Table 1: Argument types example

	Description	Argument Type
n2 n3	the system is formally verified against SRs 1-10 the system is tested against SRs 1-10	Alternative
n2 n3	the system is formally verified against SRs 1-3 the system is tested against SRs 4-10	Disjoint, $w_1 = 3$ and $w_2 = 7$
n2 n3	the system is formally verified against SRs 1-4 the system is tested against SRs 4-10	Overlap, $w_1 = 3$, $w_{Overlap} = 1$, and $w_2 = 6$
n2 n3	the system is formally verified against SRs 1-4 the system is tested against SRs 1-10	Containment, $w_1 = 4$ and $w_{Containment} = 6$



(a) Alternative (b) Disjoint (c) Overlap (d) Containment

Fig. 2. Argument types

Alternative relates to a situation where more than one independent support of the common conclusion is provided. In other words, each of the supporting nodes supports the whole conclusion (e.g., the first row in Table 1). In this case, the rule given in Definition 2 is used to aggregate the mass of the sufficiency and insufficiency of n2 and n3 to obtain the mass of the sufficiency and insufficiency of n1.

Disjoint relates to a situation where the supporting nodes provide complementary support for the conclusion. This means, each of the supporting nodes covers part of the conclusion (e.g., the second row in Table 1). In such a case, not only the assessments of the supporting nodes but also the weights, representing the coverage, associated with each supporting node are taken into account. The final assessment of the conclusion is a sort of weighed average (i.e., Definition 4) of the contribution of all the supporting nodes.

Overlap relates to a situation where the supporting nodes support overlap parts of the conclusion. So each of the supporting nodes covers 'not disjoint' part of the conclusion (e.g., the third row in Table 1). In such a case, the weights associated with each supporting node and the weight of the overlap are taken into account. For the simple example given in Figure 2c, first the two overlapped nodes are restructured into three disjoint parts as shown in Figure 3a. This restructuring is valid under the assumption that the degree of belief in a node equals the degree of belief in its pieces. E.g., the degree of belief in an evidence node referring to *the testing results for 3 safety requirements SR1, SR2, and SR3* is the same as the degree of belief in the testing results for each single safety requirement *SR1, SR2*, and *SR3*. In other words, this combined evidence can be seen as three smaller evidence nodes each of which points to the testing results for one safety requirement. In this case, the degree of belief in each small piece of evidence equals the degree of belief in the combined evidence as the same testing mechanism and settings are used for the three safety requirements. The middle part in Figure 3a is covered by both evidence nodes, so Definition 2 is applied for hypothesis $C \in \{S, I, U\}$

$$m_{Overlap}(C) = m_{n2}(C) \oplus m_{n3}(C).$$

The result is three disjoint nodes as shown in Figure 3a, so Definition 4 is applied twice. We can compute this by combining any pair of m_{n2}, $m_{Overlap}$, and m_{n3} with the corresponding weights, and then combine the result with the remaining third mass. Let's first combine m_{n2} and $m_{Overlap}$

$$m_{intermediate}(C) = \frac{w_1 * m_{n2}(C) + w_{Overlap} * m_{Overlap}(C)}{w_1 + w_{Overlap}}.$$

Then combine the result with m_{n3}

$$m_{n1}(C) = \frac{(w_1 + w_{Overlap}) * m_{intermediate}(C) + w_2 * m_{n3}(C)}{(w_1 + w_{Overlap}) + w_2}.$$

The result would be

$$m_{n1}(C) = \frac{w_1 * m_{n2}(C) + w_{Overlap} * m_{Overlap}(C) + w_2 * m_{n3}(C)}{w_1 + w_{Overlap} + w_2},$$

where w_1, $w_{Overlap}$ and w_2 represent the coverage of n1 by n2 only, the overlap, and n3 only respectively.

Containment relates to a situation where one supporting node coverage is included in a bigger supporting node coverage. In other words, each of the supporting nodes covers part of the conclusion; this part is covered also by the next larger support (e.g., the last row in Table 1). In such a case, the weights associated with each supporting node are taken into account. For the simple example given in Fig-

ure 2d, first the two nodes are restructured into two disjoint parts as shown in Figure 3b. This restructure is valid under the same assumption given for the overlap case. In Figure 3b, the degree of belief in the part covered by both supporting nodes is calculated by applying the rule given in Definition 2.

$$m_{Containmen\ t}(C) = m_{n2}(C) \oplus m_{n3}(C)$$

The result is two disjoint nodes, so Definition 4 is used:

$$m_{n1}(C) = \frac{w_1 * m_{n2}(C) + w_{Containmen\ t} * m_{Containmen\ t}(C)}{w_1 + w_{Containmen\ t}}$$

where w_1, and $w_{Containment}$ represent the coverage of n1 by n2 only and the containment respectively.

(a) The overlap case (b) The containment case

Figure 3. Argument types restructure

Example 2. Table 2 shows the results of applying the aggregation rules for each argument type with different weights for each supporting node.

Table 2. Example of different argument type aggregation results

	Alternative	Disjoint $\{w_1, w_2\}$	Overlap $\{w_1, w_{Overlap}, w_2\}$			Containment $\{w_1, w_{Containmen\ t}\}$		
weights		{6, 4}	{6,1,3}	{5,3,2}	{4,5,1}	{8, 2}	{5, 5}	{2, 8}
$m_{n1}(S)$	0.8148	0.58	0.592	0.6344	0.7185	0.563	0.657	0.752
$m_{n1}(I)$	0.1111	0.16	0.161	0.1533	0.1343	0.182	0.156	0.129

We can see that for the overlap case, when the overlapping is significant then the results are close to the alternative case, and when the overlapping is insignificant then the results are close to the disjoint case (using the same weights). For the containment case, when one of the supports is very small, then its contribution is negligible, and the results are determined by the significant support. But when the coverage of the smaller support increases, the results become closer to the alternative case. These observations are useful in practice, because it is not always obvious how to characterize the overlapping part. In such cases, the overlap is approximated to either alternative or disjoint case based on the reviewer assessment for the overlapping significance. In the same way, the containment can be ap-

proximated to either alternative case or the significant evidence numbers based on the reviewer assessment for the small evidence coverage.

It is assumed that the reviewer is an expert with sufficient competence to assess the argument nodes and express his/her opinion (as required for Step 1 in Algorithm 1), and to determine the argument type of each decomposition in the safety argument. The expert defines the argument type based on his/her understanding to the conclusion and its basis, and the supporting nodes and their basis. Where the node basis is defined by *all* context, justification, and assumption nodes attached to this node.

Example 3. Figure 4 is an example to show how the node basis is important in identifying the argument type. By checking the supporting nodes G1 and G2, it is clear that the argument type is alternative as both G1 and G2 cover the whole G0 conclusion. However, by checking A1 the expert may believe that testing reveals only 80% of the cases but there is 20% of the cases will not be covered by testing. That means G1 covers only 80% of G0, but G2 covers the whole conclusion. So the argument type is not alternative but containment where G1 is the small support, its weight is 0.8.

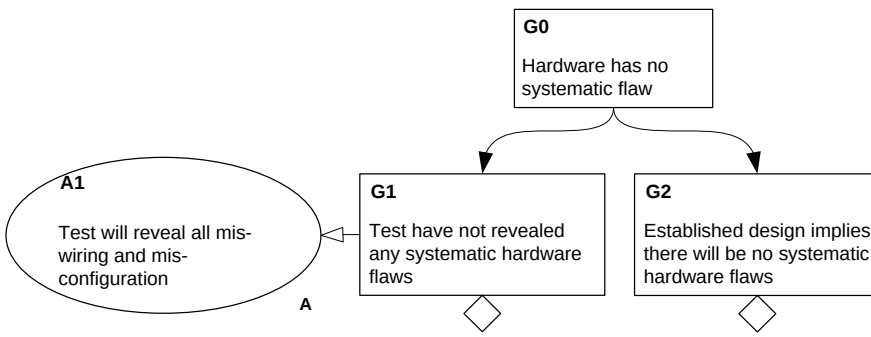


Fig. 4. Identify the argument type

In the general case, i.e., when more than two nodes support the conclusion, the four argument types are still valid. In addition, a general argument type, called *arbitrary*, can be found. The arbitrary argument type corresponds to the situation where there is no common part to *all* supporting nodes, though some supporting nodes may cover a common part. Figure 5 shows an example of such case. The arbitrary argument type can be structured as a combination of alternative and disjoint cases. Although Dempster's rule of combination and the mixing rule are associative and commutative, they do not have the same precedence and so when both rules are applied the order matters. It is the same as the case of the multiplication and addition operators, both are associative and commutative but that does not mean that $a + b + c = (a + b) * c = a + (b * c)$. The precedence of the aggregation rules impacts significantly the calculations for the arbitrary argument type. For this paper, we focus on the basic four argument types (alternative, disjoint, overlap, and containment). Elaborating the impact of the aggregation rule prece-

dence on the degree of belief calculations for the arbitrary case is one of the directions for future work.

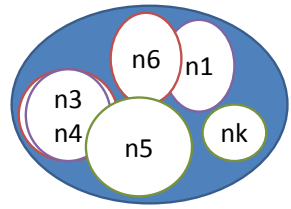


Fig. 5. Arbitrary argument type

4.2 Missing support

The first step in the aggregation process (Step 2 in Algorithm 1) is applying the corresponding aggregation rules (i.e., Definition 2 and/or Definition 4) based on the argument type. Then independently from the argument type, there may be part of the conclusion that is not covered by any of the supporting nodes. This uncovered part is represented by the shading in Figures 2, 3 and 5. For *all* the argument types, the uncovered part of the conclusion can be presented as a missing supporting node; node nm in Figure 6, where ns represents the part of n1 that is covered by the provided supporting nodes. As nm is missing then we know nothing about it. Which means, unless the reviewer has a different opinion, the degree of belief in the sufficiency and insufficiency of nm are set to 0; $m_{nm}(S) = m_{nm}(I) = 0$. In this case, nm is defined with the missing coverage weight w_m and the weight of the part of n1 that is covered by ns is w_s. In this case, the relation between n1 as a conclusion and ns and nm can be seen as the disjoint case. For n1, the mass for hypothesis $D \in \{S, I\}$ is recalculated using Definition 4 as:

$$m_{n1}(D) = \frac{w_s * m_{ns}(D) + w_m * m_{nm}(D)}{w_s + w_m} = \frac{w_s * m_{ns}(D)}{w_s + w_m} \quad (1)$$

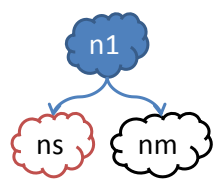


Fig. 5. Representing the missing coverage

If the missing coverage is significant, then this recalculation will significantly decrease the mass of the conclusion sufficiency and insufficiency. On the other hand, the uncertainty about the conclusion increases.

Worth notice is that based on the discussion given in Section 4.1 regarding the different precedence for the aggregation rules, the order of applying the rules matters. And as m_{ns} can be calculated by applying any aggregation rule based on the argument type, then to get consistent results the recalculation because of the missing support would be the last step to be done for any conclusion node (as shown in Algorithm 1).

There are three possible sources of missing supports. We use Figure 7 as an example to show these sources.

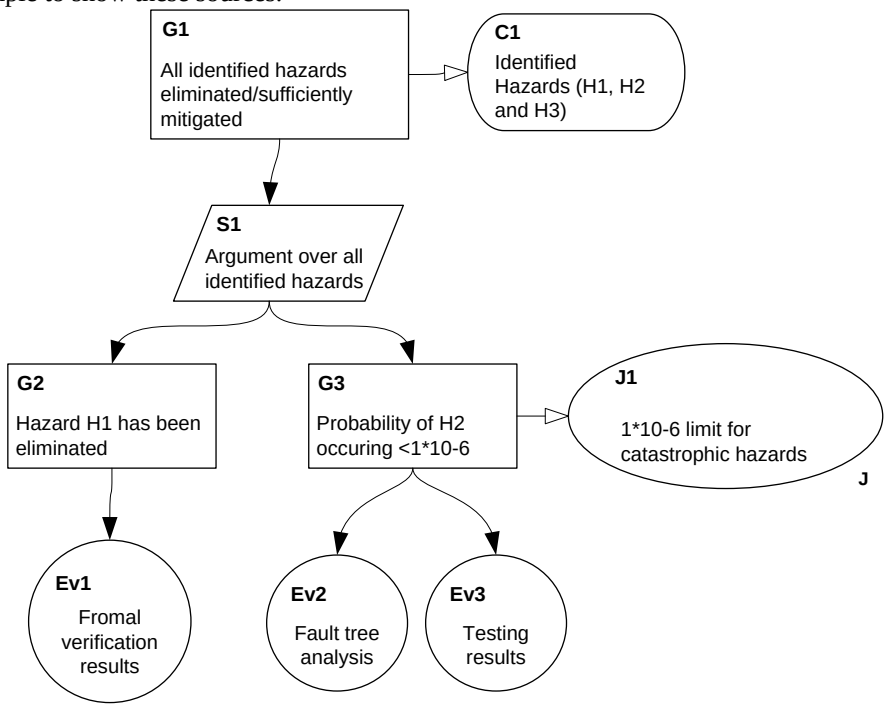


Fig. 6. A simple safety argument example

Case 1. The expert assesses the uncovered part of the conclusion by *checking the conclusion, the supporting nodes, and their basis including the inherited basis* (i.e., the context, justification, and assumption nodes attached to the conclusion, the supporting nodes, and any node higher in the argument tree). For example, S1 inherits context C1 form G1 (Kelly 1999). For S1, C1 identifies three hazards, but only two hazards are covered by G2 and G3, so only 2/3 of S1 is covered. In this case, using Equation 1, $m_{s1}(D) = \frac{2 * m_{s1}(D)}{3}$.

Case 2. In case of explicit or implicit strategy node, the expert assesses the uncovered part by *evaluating the used inference*. For example, for G1, the inference rule given in the explicit strategy S1 states {*argument by hazards mitigation*}. This inference supports G1 by showing that each individual hazard is adequately mitigated, but nothing is given to cover the situation of accumulated hazards. If the probability of getting accumulated hazards is 30%, then there is a missing support to G1 with a weight of 0.3. In this case, $m_{G1}(D) = 0.7 * m_{G1}(D)$.

Case 3. Another source of the missing support is a *defective definition for the conclusion basis*. For example, for G1, the sufficiency of the identified hazards list should be taken into consideration. In other words, if the hazards are not sufficiently identified then a missing support is identified. For example, the expert would believe that only three hazards are identified in C1, while there are three other hazards that were not mentioned. In this case, $m_{G1}(D)$ is recalculated as

$$m_{G1}(D) = \frac{3 * m_{G1}(D)}{6}.$$

As shown, more than one source of missing supports can be found for the same conclusion. E.g., Case 2 and Case 3 define missing supports to G1 in Figure 7. It is clear that the missing mitigation for *H3* is already impacted the calculations for S1. And so for G1, Case 1 does not define missing supports.

It is also clear that it is not important which source of missing supports is assessed first as the multiplication operator is associative and commutative. But the important thing is to check *all* sources (e.g., for the given example, the resultant calculations for G1 would be $m_{G1}(D) = \frac{0.7 * 3 * m_{G1}(D)}{6}$).

Note that, we assume the expert is certain about his/her opinion about the node basis, so that the expert is certain about his/her opinion that C1 in Figure 7 is missing three hazards. This may not always be the case. The expert may be uncertain about his/her opinion in many different ways. For example, the expert may have a belief that some hazards are not defined but cannot certainly tell how many hazards are missing. This uncertainty should be considered in the calculations; this is one of the directions for the future work to extend the proposed mechanism.

5 Illustrated example

We use the complete simple example given in Figure 7, inspired by the example given in (Weaver et al. 2003), to illustrate how to apply the proposed procedure given in Algorithm 1.

Step 1. Evidence assessment

- Assume the expert estimates as: $m_{Ev1}(S) = 0.8$, $m_{Ev1}(I) = 0.1$, $m_{E21}(S) = 0.7$, $m_{Ev2}(I) = 0.1$, $m_{Ev3}(S) = 0.5$ and $m_{Ev3}(I) = 0.2$.

Step 2. Automatic aggregation

- Starting from the leaves
  - For G2
    - Only one evidence Ev1 supports G2, so $m_{G2}(S) = m_{Ev1}(S) = 0.8$ and $m_{G2}(I) = m_{Ev1}(I) = 0.1$.
    - Missing support
      - **Case 1.** Assume the expert opinion is that Ev1 covers the whole G2 conclusion.
      - **Case 2.** Not applicable as there is no implicit or explicit strategy between G2 and Ev1.
      - **Case 3.** The conclusion node G2 has no basis and so no dimensioning is required.
  - For G2
    - Assume the expert opinion is that both Ev2 and Ev3 cover the whole G3 conclusion so the rule given in Definition 2 is applied. $m_{G3}(S) = 0.815$ and $m_{G3}(I) = 0.111$.
    - Missing support
      - **Case 1.** Assume the expert opinion is that conclusion G3 is totally covered by Ev2 and Ev3.
      - **Case 2.** Not applicable as there is no implicit or explicit strategy.
      - **Case 3.** Assume the expert opinion is that the justification J1 covers only 80% of the cases. By applying equation 1, $m_{G3}(S) = 0.652$ and $m_{G3}(I) = 0.089$.

- Repeat the process
  - For S1
    - It is clear that the argument type is disjoint. Assume all hazards have the same importance; the weights of all hazards are equal. Using Definition 4, $m_{S1}(S) = 0.726$ and $m_{S1}(I) = 0.094$.
    - Missing support: this is given as example in Section 4.2. The recalculation result: $m_{S1}(S) = 0.484$ and $m_{S1}(I) = 0.063$.
  - For G1
    - Strategy S1 is the only support to G1, so $m_{G1}(S) = m_{S1}(S) = 0.484$ and $m_{G1}(I) = m_{S1}(I) = 0.063$.

- Missing support: this is the example discussed in Section 4.2. The result is that $m_{G1}(S) = 0.169$ and $m_{G1}(I) = 0.022$.

The calculations are summarized in Figure 8. In this case, the mass of the overall sufficiency and insufficiency of the safety argument given in Figure 7 equal 0.169 and 0.022 respectively, and the uncertainty equals 0.809. Using these numbers the expert can decide if this safety argument should be rejected or not. In the given example, most probably the expert would reject this safety argument as the degree of uncertainty is very high relative to the degree of belief of the argument sufficiency which is quite low. In addition, the expert can provide aclear feedback to the submitter guiding where enhancements are required. For the given argument, it is clear that the first effective weakness of this argument is missing mitigation for *H3*. The inappropriate identification for the hazards has a significant negative impact as well.

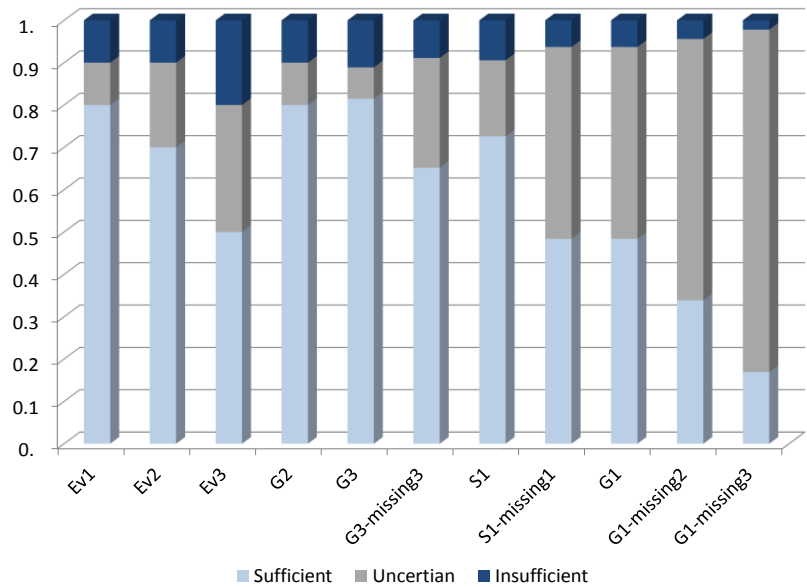


Fig. 7. The example numbers

6 Related work

The proposed assessing mechanism can be used in conjunction with the step-by-step review mechanism proposed in (Kelly 2007) to answer the question given in the last step of this reviewing mechanism, which is of the overall sufficiency of the safety argument. At the same time, applying the step-by-step reviewing mechanism guarantees the assumptions of the proposed mechanism as given in

Section 4. In other words, the step-by-step review mechanism provides a skeleton for a systematic review process; however the proposed assessment mechanism provides a systematic procedure to measure the sufficiency and insufficiency of the safety arguments.

An appraisal mechanism is proposed in (Cyra and Gorski 2008a) to assess the trust cases. Although this mechanism targets trust cases and the proposed mechanism targets safety cases, both mechanisms use the Dempster-Shaffer model. Both mechanisms propose different aggregation rules based on the argument types. However, the argument types are not identical. The proposed argument types cover the case when the falsification of one of the premises decreases, but not nullifies, the support for the conclusion. This case is also covered for the trust cases appraisal mechanism, however only two argument types are defined; alternative and complementary (i.e., disjoint). The additional argument types we defined (i.e., overlap and containment) are treated as an alternative or a complementary type based on the overlap significance. Another case defined for the appraisal mechanism is when the falsification of a single premise leads to the rebuttal of the conclusion or to the rejection of the whole argument because nothing can be inferred about the conclusion. Two argument types are defined for this case; necessary and sufficient condition list (NSC-argument) and sufficient condition list (SC-argument). These two argument types are not considered in the proposed mechanism because we believe that for safety cases, each premise should have a contribution in supporting the conclusion. So no single premise has such a huge impact that it may lead to the rejection of the whole argument. And in case of a single premise that is sufficient to reject the whole argument then it is a simple process to check that premise and decide about the argument rejection without any aggregation.

The linguistic scales given in (Cyra and Gorski 2008b) to express the expert opinions and the aggregation results are appealing. In their work, the linguistic values are mapped into the interval [0, 1] and then quantitative rules are used for the aggregation (Cyra and Gorski 2008a). Although linguistic scales are more appropriate for human decisions than numbers, the mapping has a significant impact on the computed results and there is no evidence that the used mapping is proper.

5 Conclusion

In this paper, we propose an assessment method to assess the overall sufficiency and insufficiency of safety arguments. For the proposed mechanism, there are various parts that require interaction with the reviewer. The reviewer has to assess the evidence hypotheses, the argument types, the existence and the weight of missing supports, and the inference deficits. In other words, the proposed method does not replace the reviewer; instead it provides a framework to lead the reviewer through the evaluation process and to combine the reviewer estimates.

One of the main limitations of the proposed method is that the evidence nodes have to be independent as required by Dempster-Shafer Theory. However in many

safety arguments, evidence nodes are not independent. Extending the proposed mechanism to cover this dependency is one of the directions for future work.

Our preliminary experience of applying the proposed method has revealed that the assessing mechanism yields the expected benefits in guiding the safety argument reviewer and helping him/her to reduce the effect of the confirmation bias mindset.

Acknowledgments This work is supported in part by the NSF CPS grant CNS-1035715 and the NSF/FDA Scholar-in-Residence grant CNS-1042829.

References

Ayoub A, Kim B, Lee I, Sokolsky O (2012) A systematic approach to justifying sufficient confidence in software safety arguments. In: SAFECOMP

Bloomfield R, Littlewood B, Wright D (2007) Confidence: its role in dependability cases for risk assessment. In: DSN

Cyra L, Gorski J (2008a) Supporting expert assessment of argument structures in trust cases. In: PSAM

Cyra L, Gorski J (2008b) Expert assessment of arguments: a method and its experimental evaluation. In: SAFECOMP

Denney E, Pai G, Habli I (2011) Towards measurement of confidence in safety cases. In: ESEM'11

FDA (2010) Guidance for industry and FDA staff – total product life cycle: infusion pump – premarket notification [510(k)] submissions. US Food and Drug Administration Center for Devices and Radiological Health

Haddon-Cave C (2009) The Nimrod review: an independent review into the broader issues surrounding the loss of the RAF Nimrod MR2 aircraft XV230 in Afghanistan in 2006. Technical report, The Stationery Office (TSO)

Hawkins R, Kelly T, Knight J, Graydon P (2011) A new approach to creating clear safety arguments. In: Dale C, Anderson T (eds) Advances in systems safety. Springer

Kelly T (1999) Arguing safety. A systematic approach to managing safety cases. PhD thesis, Department of Computer Science, University of York

Kelly T (2007) Reviewing assurance arguments - a step-by-step approach. In Proceedings of Workshop on Assurance Cases for Security – The Metrics Challenge, DSN '07.

Kelly T, Weaver R (2004) The goal structuring notation – a safety argument notation. In: DSN 2004 Workshop on Assurance Cases

Leveson N (2011) The use of safety cases in certification and assurance. Working paper

Menon C, Hawkins R, McDermid J (2009) Defense standard 00-56 issue 4: towards evidence-based safety standards. In: Dale C, Anderson T (eds) Safety-critical systems: problems, process and practice. Springer

Sentz K, Ferson S (2002) Combination of evidence in Dempster-Shafer theory. Technical report, Sandia National Laboratories, SAND 2002-0835

Voorbraak F (2012) Dempster-Shafer theory. www.blutner.de/uncert/DSTh.pdf. Accessed 30 September 2012

Weaver R, Fenn J, Kelly T (2003) A pragmatic approach to reasoning about the assurance of safety arguments. In the Australian workshop on Safety critical systems and software

Wikipedia (2012a) Decision making. http://en.wikipedia.org/wiki/Decision_making. Accessed 30 September 2012

Wikipedia (2012b) Mindset. http://en.wikipedia.org/wiki/Mindset. Accessed 30 Sept 2012

Zadeh L (1984) Book review: A mathematical theory of evidence. AI Magazine, 5(3):81-83