Defence Standard 00-56 Issue 4: Towards Evidence-Based Safety Standards

Catherine Menon, Richard Hawkins, John McDermid

Software Systems Engineering Initiative, Department of Computer Science, University of York

Heslington, N. Yorks., UK

Abstract Defence Standard 00-56 Issue 4 is goal-based, and requires system developers to demonstrate how they have achieved safety. To this end, evidence is used to support claims relating to software safety. One of the most subtle questions when constructing a safety argument is the determination of whether the evidence presented is sufficient to assure the safety of the system to the level required. This paper presents a framework for assessing the assurance of evidence and claims. We also present a vocabulary for discussing factors which influence assurance. This framework and vocabulary together enable us to construct and discuss safety arguments for software. Using this framework and vocabulary, we present some sample discussions which demonstrate how the factors influencing assurance can interact.

1 Introduction

DS 00-56 Issue 4 (Ministry of Defence 2007) is goal-based – it sets out requirements relating to safety management, but does not prescribe how those requirements will be met. In general, before a system can be deployed, it is necessary to produce a safety case setting out an argument and supporting evidence that the system is acceptably safe.

The UK Ministry of Defence has adopted a principle that standards should be 'as civil as possible, and only as military as necessary'. DS 00-56 Issue 4 deliberately moved away from prescription to allow, for example, the software elements of a system to be developed to appropriate civil standards, e.g. DO178B (RTCA and EUROCAE 1992) for aircraft software. However, it would also be equally valid to use a standard such as EN 50128 (British Standards 2001) for software in a railway signalling system, or a bespoke approach for military-unique systems. In all these cases there is an issue of what amounts to an acceptable argument and

what constitutes sufficient evidence – which we can perhaps best summarise as sufficient assurance of evidence and argument.

The MOD has funded the Software Systems Engineering Initiative (SSEI) as a centre of excellence for defence software systems. One of the initial tasks to be undertaken by the SSEI is to provide guidance on software safety, in the context of DS 00-56 Issue 4. One of the key issues to be addressed by this work is to provide a sound basis for assessing, or measuring, assurance in evidence. This paper focuses on the core technical issues in assessing assurance, and the conclusions outline how the model set out here will fit into the guidance to be produced by the SSEI.

One of the most subtle aspects of the evidential approach is assessing whether the safety requirements have been satisfied by the arguments and evidence presented. The assurance of a system is the measure of how much confidence we have in the safety argument and supporting evidence. That is, the assurance of a system is the extent to which we are confident that the safety requirements have been met. The assurance requirements on a system vary with the risk of the system hazards, and any failure to meet these requirements must be justified. We recommend that the justification take the form of an argument based on the ALARP principle, and will refer to this as ACARP (As Confident As Reasonably Practical).

We present here a framework for assessing the assurance of a safety argument. This framework identifies the major factors which influence the assurance of a claim, and therefore influence the confidence we have in the safety of a system. The framework also provides a means of calculating assurance from these factors. One of the primary advantages of such a framework is that it provides a generalised foundation for calculating assurance in any system, and furthermore for verifying the accuracy of any assurance claimed to have been achieved by a safety argument and the associated evidence. In addition, this framework establishes a vocabulary to discuss safety arguments. Thus, it is possible to communicate to the people responsible for providing evidence – such as testing evidence or formal analysis of the system – precisely what evidence would be needed to achieve the required assurance.

Section 2 establishes a framework for addressing assurance in arguments, and section 3 extends this to evidence presented in support of an argument. Section 3 also identifies questions which should be asked when determining the quality of the evidence presented. Assurance is a multi-faceted concept, and there is a risk of ‘double-accounting’ when assessing assurance; section 4 discusses this issue and considers how to combine assurance. Section 4 also considers the issue of propagating assurance through argument structures, and section 5 draws conclusions.

2 Claims and Arguments

Throughout this paper we will be referring to the constituent parts of a safety argument. Our terminology for discussing arguments is based on GSN (Kelly 1999), although there are other acceptable methods of presenting safety arguments (Adelard 1998). We define the key concepts we will use below.

Definition 1. Within a system safety case, a *claim* is a statement made about the system which may or may not be true.

For example 'All omission failures are detected and handled acceptably' is a claim about a system.

Definition 2. An *argument* is a connected sequence of statements intended to establish a claim.

For example, 'All omission failures are detected and handled acceptably, because Components A and B are present in the system, and tests show that they detect and handle all possible omission failures' is an argument.

Definition 3. A *higher-level claim* is a claim which is supported by other claims.

For example, the claim 'Software safety requirements mitigate all system hazards' is supported by the three claims 'Software requirements are adequate and address all hazards', 'Software requirements are met' and 'Software requirements are traceable'.

Definition 4. A *leaf claim* is a claim which is supported directly by evidence.

For example, a leaf claim might be 'Function X has no side effects', being supported by static analysis.

In the course of refining an argument, it is possible for a claim which was originally presented as a leaf claim to become a higher-level claim. It is also possible for claims to be simultaneously higher-level claims and leaf claims, depending on whether they are supported exclusively by other claims, by evidence, or by a mixture of both.

The assurance of a higher-level claim is dependent upon the assurance of its supporting claims. However, the assurance of the higher-level claim may not be dependent upon every supporting claim to an equal extent. In the following section, we identify some factors which influence the degree to which a higher-level claim may be dependent upon a specific supporting claim. We also identify some more general influences which can increase or decrease the assurance of a claim. In this way we will present a means of analysing an argument to identify where assurance deficits (discrepancies between the assurance required and the assurance achieved) may have been introduced.

2.1 Assurance Factors

Throughout this discussion our model argument will consist of a higher-level claim *HC* supported by two supporting claims SC_1 and SC_2. That is, the purpose of the safety argument will be to justify the inference $SC_1 * SC_2 \rightarrow HC$ (where the symbol * is to be interpreted as conjunction). The assurance of *HC* is a combination of the assurance of SC_1 and SC_2 and the strength of the inference $SC_1 * SC_2 \rightarrow HC$. The strength of the inference is subject to the following factors.

2.1.1 Scope of Supporting Claims

Definition 5. *Scope* is defined as the degree to which the supporting claims entail the entirety of the higher-level claim.

Scope is most easily understood where it refers to the extent of the claim *HC* which is addressed by either SC_1 or SC_2. For example, consider the argument $SC_1 * SC_2 \rightarrow HC$, where the claims are instantiated as follows:

HC: 'Function X is fault-free in all 10 operational modes'
SC_1: 'Function X is fault-free in operational modes 1, 2, 3, 4, 5'
SC_2: 'Function X is fault-free in operational modes 6, 7'

In this case, we deduce that the *scope* of $SC_1 * SC_2$ is most, but not the entirety, of *HC* (in practice, we would also be interested in other variables such as the time spent in each operational mode during typical use). We can also deduce that the scope of SC_1 is larger than the scope of SC_2. Consequently, the assurance of *HC* is dependent upon the assurance of SC_1 to a greater extent than it is upon the assurance of SC_2. A more complex example of scope can be observed in the following example, where *HC* is supported by claims SC_1, SC_2 and SC_3.

HC: 'Software safety requirements mitigate all system hazards'
SC_1: 'Software requirements are adequate and address all hazards'
SC_2: 'Software requirements are met'
SC_3: 'Software requirements are traceable'

Here, several supporting claims each address a different facet of the higher level claim. In the absence of any argument to justify why one supporting claim is more important than another, we assume each of these to have equal scope.

To avoid confusion, we will use the expression *equal scope* to describe the situation where multiple supporting claims influence the higher-level claim to the same extent. That is, equal scope implies that each supporting claim addresses different aspects from the others, but all these aspects are equally important to the higher-level claim. We will use the expression *identical scope* to describe the situation where multiple claims address the same aspect of the higher-level claim. This has also been referred to as convergent support (Govier 1988), and is described more fully in Section 2.1.4.

The advantage of identifying scope as a factor in assurance is twofold. Firstly, it formalises the idea that a higher level claim can be dependent upon one supporting claim to a greater extent than the others. That is, in the absence of all other factors (see below) we can state that the assurance of the higher level claim is most strongly influenced by the assurance of the supporting claim with the greater scope. Secondly, scope helps us to understand why assurance deficits may have been allowed within a safety argument. Where the combined scope of supporting claims does not address the entirety of the higher level claim, this is an indication that the argument structure is flawed or limited. In other words, this indicates that an essential premise of the safety argument is missing, meaning that some aspects of the higher level claim are not supported in any way. The visible consequence of this is that the assurance of that aspect of the higher level claim is zero, and the assurance of the entire higher level claim is therefore diminished. In general, such assurance deficits would need to be justified along ACARP principles if they remained in a final safety case.

2.1.2 Independence of Supporting Claims

Definition 6. *Independence* is defined as the diversity between the sets of evidence used to support the claims SC_1, ..., SC_n in the inference $SC_1 * ... * SC_n \rightarrow HC$.

More specifically, independence is the measure of how qualitatively 'different' the evidence supporting SC_1 is from the evidence supporting SC_2 (note that if SC_1 and SC_2 are not leaf claims, then evidence can only support them via other supporting claims). If the evidence that supports SC_1 shares some significant characteristics with the evidence which supports SC_2, then SC_1 and SC_2 are said to demonstrate low independence from each other, and the assurance of the higher-level claim *HC* will consequently be diminished.

Independence may be *conceptual* or *mechanistic*. Items of conceptually different independence are based on different underlying theories, while items of mechanistically independent evidence are obtained by implementing the same underlying theory in different ways (Weaver et al. 2003).For example, formal methods and testing will produce conceptually independent evidence, while conducting testing alone using a variety of techniques will produce mechanistically independent evidence.

One illustration of the consequences of a lack of independence is the effect that common-cause failures can have on the assurance of a claim. If the evidence for SC_1 is generated using a particular tool, then any undetected failure in that tool will result in flaws in the evidence and consequently an incorrect (often unjustifiably high) confidence in the truth of SC_1. Because the assurance of higher level claims depends upon the assurance of the supporting claims, the assurance of *HC* will also be affected by this tool failure. If the same tool is then used to generate evidence for SC_2, the effect of the tool error on the assurance of *HC* will be com-

pounded. In other words, a high degree of conceptual and mechanistic independence between the evidence supporting SC_1 and SC_2 will reduce the impact of any common-cause failure when generating these groups of evidence. Section 3.7 contains further discussion on how to estimate the independence of items of evidence.

2.1.3 User-defined Importance

Definition 7. *User-defined importance* is the additional weighting placed upon one or more supporting claims due to legislative or other precedents.

Legislation and standards often identify certain safety principles as carrying more weight than others in an argument. This can lead to a supporting claim SC_1 being considered more 'important' than another, say SC_2, even though SC_1 and SC_2 may have equal scope. Consequently, the assurance of *HC* is then affected by the assurance of SC_1 more than the assurance of SC_2.

The safety of children or the general public as compared to the safety of adults or defence force personnel is a common example of this (Health and Safety Executive 2001). For example, consider the argument $SC_1 * SC_2 \rightarrow HC$, where the claims are instantiated as follows:

> *HC*: 'Personnel in all 4 rooms of the nuclear plant are acceptably protected against failures'
> *SC1*: 'Safety systems are in place for rooms 1 and 2 (approx. 10 people)'
> *SC2*: 'Safety systems are in place for rooms 3 and 4 (approx. 10 people)'

The scope of SC_1 is equal to the scope of SC_2, because they both address equal extents of the claim *HC*. However, if rooms 3 and 4 are the nuclear plant childcare centre (assuming there is one), then guidance for decision-making issued by the Health and Safety Executive (Health and Safety Executive 2001) would consider the claim SC_2 to be more important.

User-defined importance therefore allows us to consider many of the unstated or 'intuitive' considerations which can affect the quality of a safety argument. While theoretically it is certainly the case that the safety of the power plant in the example above is dependent equally upon the assurance with which we can state both sets of rooms are safe, a doubt as to the safety of the childcare centre is far less palatable than a doubt as to the safety of the other rooms. That is, this example demonstrates a situation where a doubt about one supporting claim SC_1 is more easily justified than a doubt about another supporting claim SC_2, even though both these doubts would logically have the same impact on the assurance of the higher-level claim *HC*. In other words, due to user-defined importance SC_2 has a potentially much greater impact on the assurance of *HC* than SC_1, so the consequences of failing to meet assurance requirements on SC_2 will be correspondingly greater.

Previous attempts to define assurance (Weaver 2004) have not formally codified this factor. In general, however, arguments are always written with an in-

tended reader in mind, and are written to be compelling from the point of view of that reader. Consequently, all arguments consider user-defined importance to a certain extent. By formalising this, we make it possible to assess, for example, the impact of public feeling upon the required assurance for safety-critical systems.

User-defined importance is usually expressed in general terms, meaning that certain principles (such as the safety of children) are defined to be of greater importance than others (such as the safety of defence force personnel) when constructing a safety case. In some cases this has been codified in legislation, standards or guidance (Health and Safety Executive 2001). An example of the latter case is the prioritisation of certain types of evidence expressed in DS 00-56 Issue 4 (Ministry of Defence 2007). Section 3.6 discusses this in further detail.

2.1.4 Reinforcement

Definition 8. *Reinforcement* is defined as the extent to which multiple supporting claims address the same aspects of a higher-level claim.

Arguments where reinforcement is relevant are those where two (or more) supporting claims SC_1 and SC_2 have *identical* scope. That is, where SC_1 and SC_2 address the same aspects of the higher-level claim *HC*. A high degree of reinforcement within the supporting claims means that the assurance of *HC* will be increased.

When assessing the effects of reinforcing a supporting claim SC_1, it is important to consider the assurance of the other claims which will be used to reinforce SC_1. While it is certainly true that the assurance of the higher level claim *HC* can be increased by introducing a claim SC_3 which reinforces SC_1, the extent of this increase can vary. If both SC_3 and SC_1 are strongly assured themselves, the reinforcement will have a significant positive effect on the assurance of *HC*. Equally, reinforcing a supporting claim SC_1 which has low assurance with a claim SC_3 with high assurance will greatly increase the assurance of *HC*. However, where two supporting claims are only weakly assured themselves, a high degree of reinforcement between these claims will only marginally increase the assurance of *HC*. The interaction of independence and other factors will influence the effect of reinforcement, as we discuss in Section 4.1.4. Nevertheless, the value of reinforcement within each individual safety argument must be assessed on its own merits.

Reinforcement can also be used to express the effect of *counter-evidence*, a quantity recommended for consideration by DS 00-56 Issue 4. This is discussed in more detail in Section 3.8.

2.2 Applying Assurance Factors

We can use the factors introduced in Section 2.1 to provide a means of calculating the assurance of a higher level claim. The assurance of any claim *HC* is dependent upon:

- The assurance of each supporting claim, allowing for:
  - The scope of this supporting claim relative to *HC*
  - The user-defined importance of this supporting claim
- The independence of all supporting claims
- The degree to which any supporting claims are reinforced

That is, the assurance of the claim *HC* is dependent solely upon the assurance of the supporting claims SC_1 and SC_2, the independence of these supporting claims, and the extent of reinforcement between these claims. The degree to which each the assurance of an individual supporting claim SC_1 affects the assurance of *HC* is determined by the scope of SC_1 and any additional importance placed upon principles which affect SC_1.

The two primary advantages of decomposing assurance as described above are an ease in communication, and a more standardised approach to safety arguments. This framework makes explicit a number of different ways to improve the assurance of a claim, as well as providing a means to assess the impact of each individual claim on the assurance of the entire argument. Because safety arguments depend ultimately on the assurance of leaf claims, we devote the next section to a discussion of how to ensure maximum assurance at this level.

3 Leaf Claims

As described above, assurance within a safety argument 'cascades upwards', with the assurance of higher-level claims being determined from the assurance of supporting claims. This means that the assurance of the leaf claims is of vital importance, supporting as they do the entire safety argument. Unfortunately, when generating evidence it is common for there to be limited visibility of the proposed safety argument structure. Consequently, it is often difficult to determine the value of an item of evidence to the safety argument. Furthermore, the people generating the evidence may not be system safety experts and may not have the resources to interpret abstract principles for increasing assurance, and apply them to evidence. To negate this problem, we have provided sample checklists of questions which will help determine the quality of an item of evidence. These checklists facilitate discussion between safety experts and system developers, by providing an accessible language to discuss those properties which are required for the evidence to support a compelling safety argument.

3.1 Assessing the Assurance of a Leaf Claim

The assurance of a leaf claim depends upon the quality, or rigour, of the evidence provided. We have identified seven factors which must be considered in determining the assurance of a leaf claim from evidence. Four of these, *scope*, *user-defined importance*, *independence* and *reinforcement* have been discussed previously in Section 2.1. There are also three factors which apply solely when assessing the assurance for a leaf claim: *replicability*, *trustworthiness* and *coverage*. These factors are as follows:

- Replicability: the ease with which the evidence could be replicated.
- Trustworthiness: the likelihood that the evidence is free from errors.
- Scope: the extent of the claim which this type of evidence could be reasonably expected to address.
- Coverage: the extent of the claim which is actually addressed by the evidence, relative to the scope.
- User-defined importance: the additional weighting placed upon certain types of evidence by legislation, standards or client preference.
- Independence: the diversity of the evidence, as well as the different tools and methods used to obtain evidence.
- Reinforcement: the extent to which multiple items of evidence support the same aspects of a leaf claim.

The following sections discuss how each factor is to be interpreted with respect to evidence. We also present sample checklists of questions which can be used to assess the quality of evidence. These checklists require neither visibility of the overall safety argument, nor a background in software safety management. Section 3.9 then describes how to estimate the assurance of a leaf claim from these factors.

3.2 Replicability

The replicability of evidence is the extent to which this evidence can be reproduced. Evidence may not be replicable for two reasons:

- The circumstances under which the evidence was obtained no longer hold.
- The evidence is by nature subjective.

The first situation commonly arises for evidence which is the result of discussion or analysis during an early part of the development. HAZOP analysis is a good example of this, in that it cannot be reproduced at a later date. Firstly, the system may have changed so that hazards which were present have now been removed entirely, or new hazards introduced. Secondly, even if this is not the case, the discussion and thought-processes of the participants will not be exactly the same as before. Consequently, it is impossible to reproduce the analysis and gain the same

results. Another common example of evidence which lacks replicability is in-service or historical evidence.

The second reason for a lack of replicability is a lack of objectivity in the evidence, a situation which commonly arises with review evidence. Although the competence of multiple reviewers may be judged to be equal, there is no guarantee that they will produce identical reviews. Similarly, any evidence which relies on interpretation is said to lack replicability.

Supporting a leaf claim with replicable evidence will increase the assurance of that claim to a greater degree than supporting it with evidence which is not replicable. This is a commonly accepted principle, to the extent where replicable evidence is often officially preferred within standards and contract conditions. For example, DS 00-56 Issue 4 (Ministry of Defence 2007) expresses a preference for analytical evidence, whereas in DS 00-55 Issue 2 (Ministry of Defence 1997) the emphasis is on formal techniques.

3.3 Trustworthiness

The trustworthiness of evidence is the faith which we place in the integrity of the evidence. Untrustworthy evidence is often characterised as evidence which is 'buggy' (Weaver et al. 2005) and can greatly reduce the assurance of claims it supports. Questions to consider when determining the trustworthiness of evidence include:

- Was the evidence gathered in accordance with any documented standard (for example, a COTS product developed to DO-178B will have some evidence in its safety case, gathered according to the principles of this standard)?
- Are the evidence-gathering personnel competent? Are they certified to an appropriate standard? Have they performed the tests before?
- How valid are the assumptions and simplifications that were made?
- Is there a culture of safety in the environment where the evidence was gathered?
- For COTS products especially, has the evidence been obtained from disinterested sources?
- Is there any counter-evidence (see Section 3.8.1)?

Supporting a leaf claim with trustworthy evidence will increase the assurance of that claim to a greater degree than supporting it with evidence which is not trustworthy. It is important to note that evidence which is deemed trustworthy may still contain flaws. For example, while there may be every indication that a tool is trustworthy, it is still possible that the results produced by applying this tool contain false negatives. Consequently, independence (Section 3.7) is still an important factor in guarding against common cause failures no matter how apparently trustworthy the evidence. Section 4.1.6 explains more about how trustworthiness can interact with the other factors which influence assurance.

3.4 Scope

Scope has been introduced in Section 2.1.1 and, for evidence, is to be interpreted as the extent to which a particular type of evidence can imply the truth of a leaf claim. Questions to consider to help determine the scope of a particular item of evidence include:

- Does this type of evidence typically produce results which would support all parts of the leaf claim? For example, evidence produced from formal analysis is unlikely to support a claim about a lack of timing failures in the system.
- If the evidence is to be obtained from testing, how much of the relevant functionality referred to in the leaf claim will be tested?
- If the leaf claim refers to multiple components, do you envisage testing all these components and their interactions?
- Will all applicable operational modes be examined when generating this evidence?

In keeping with the earlier notation, if *E* represents an item of evidence, and *LC* a leaf claim, the scope of *E* helps determine the strength of the inference $E \rightarrow LC$. The higher the combined scope of all evidence supporting *LC*, the higher the assurance of *LC* will be. If two items of evidence are provided to support a claim, the evidence with the greater scope will have the greater effect on the assurance of the claim.

3.4.1 Scope and Evidence

If the evidence does not cover the full scope of the leaf claim, this will result in diminished assurance for the leaf claim. While this does not necessarily imply that the assurance of the system will not meet the requirements, it does signal the need for an ACARP argument to justify this decrease in assurance. Such an argument would provide reasons as to why it is not necessary to generate further evidence to address the missing scope. The scope of evidence is determined mainly by the type of the evidence, and the extent to which it is possible for such evidence to fully address the leaf claim. This could also be referred to as the intent of the evidence.

The scope of evidence can be determined before the evidence has been generated. Scope is affected by considerations such as the fidelity of any models used in formal analysis, the planned coverage of tests, the number of operational modes for which historical data can be sought and so on. There is a closely related concept to scope, which determines how well the evidence gathering processes have been implemented, or the intent has been achieved: *coverage*.

3.5 Coverage

The coverage of an item of evidence supporting a claim is the extent of the claim addressed by this evidence, relative to that which could reasonably have been expected from evidence of this type (that is, relative to the scope of this evidence). Questions to consider when determining coverage are:

- How much of the relevant software functionality was examined, compared with what could have been examined?
- To what degree was consistency of configuration maintained?
- In how many different valid operational modes or environments did the evidence-gathering take place? Was there a reason why not all operational modes were examined?
- For historical evidence, were all major sources considered?
- To what extent did the evidence gathering processes match the usage profile of the system?
- How thorough was any review evidence? Were all relevant documents made available to the reviewers, and was the system adequately completed at the time of review?

Supporting a leaf claim with evidence demonstrating high coverage will increase the assurance of that claim to a greater degree than supporting it with evidence demonstrating low coverage. An evidence-gathering process which is implemented and executed exactly as planned will theoretically generate evidence with maximum coverage. If a leaf claim is supported solely by evidence which does not provide maximum coverage, an argument using ACARP principles will be required to justify why this evidence is thought to provide sufficient assurance.

3.5.1 Coverage and Scope

Scope and coverage illustrate different reasons why the assurance of a leaf claim (and consequently of any higher-level claim it supports) may be lower than desired. If low scope is the cause of the low assurance, this indicates that the evidence-gathering processes were not appropriate to the task. For example, modelling a system using formal methods is not an appropriate technique to gather evidence about a lack of timing failures. Similarly, normal-range testing will address only a small part of a claim that a system is robust to erroneous input. In both these situations, therefore, the proposed item of evidence will have low scope and will not strongly support the claim.

By contrast, if low coverage is the cause of the low assurance, this indicates that the evidence-gathering processes, while appropriate, were not implemented as well as expected. For example, review evidence may be used to support a claim that the system was developed in accordance with good practice. The scope of this evidence may be quite high. However, if the review is undertaken midway

through the development lifecycle the coverage will be low because many lifecycle activities would not yet have been completed.

3.5.2 Rigour: Coverage, Trustworthiness and Replicability

Coverage, trustworthiness and replicability together make up what is termed the *rigour*, or quality, of the evidence. Presenting evidence which is highly rigorous will increase the assurance of a leaf claim more than presenting evidence which is less so. That is, to achieve a given assurance, the quantity of evidence required will vary inversely with its rigour. Similarly, *counter-evidence* (Section 3.8.1) of greater rigour will decrease the assurance of a claim more than counter-evidence of little rigour.

3.6 User-defined Importance

The user-defined importance of a type of evidence is the additional weighting which is to be placed on that evidence by historical or legislative precedence, or by client preference. For example, DS 00-56 Issue 4 presents five evidence categories in order of importance. The assurance of a higher level claim is to an extent dependent upon the presence of evidence which is defined as important in this way. That is, providing a type of evidence which is defined to be more 'important' will increase the assurance of a leaf claim to a greater extent than providing a type of evidence which is not deemed so. Questions to consider when determining if there is any explicit user-defined importance on certain evidence types are:

- Has conformity to a particular standard been requested? If so, does that standard place a weighting on evidence types?
- Does the contract for the work state that particular types of evidence are preferred?
- Has the client specifically stated a preference for certain evidence-gathering processes (perhaps based on the track record of the supplier)?

An item of evidence which is deemed to be of greater importance than the others provided will have a greater effect on the assurance of the leaf claim.

It is important to remember that there may be no *explicit* description of the types of evidence which are deemed to be most compelling. In this case, user-defined importance is taken to be neutral. That is, any *implicit* user-defined importance cannot be considered as binding when determining assurance of claims.

3.7 Independence

Providing multiple items of independent evidence will increase confidence in the claim they support. Conceptual independence is preferred to mechanistic independence, and providing both will maximise the increase in confidence. Questions to consider when determining the degree of independence which has been obtained are:

- Have multiple items of conceptually diverse evidence (e.g. testing, formal analysis, review evidence) been presented?
- Has evidence been gathered in a number of mechanistically different ways? For example, if testing is performed using an automated tool, have a number of different tools been used?
- Have reviews been endorsed or approved by a number of different people?
- For COTS products, does evidence originate from a number of different sources?

Increasing the independence of the evidence provided to support a claim will increase the assurance of that claim. By contrast, if only one type of evidence is provided to support a claim, an argument using ACARP principles will be required to justify why this evidence is thought to provide sufficient assurance to the claim.

3.8 Reinforcement

Reinforcement is the extent to which multiple items of evidence support identical scope of a leaf claim. The assurance of a claim will be increased if the evidence supporting it is adequately reinforced. Note, however, that if evidence does not address the full scope of a leaf claim, then reinforcing this evidence will not increase the scope. That is, the assurance of the claim will still be negatively affected due to inadequate evidence scope.

One of the important aspects when considering how to reinforce evidence is to ensure that the items of evidence in question are independent. We discuss this in more detail in Section 4.1.4. When assessing whether one item of evidence is reinforced by another, the following questions should be considered:

- Do the two types of evidence have identical scope?
- If the two types of evidence do not have identical scope, can the results from one be extrapolated to provide reinforcement for the other?
- Is there another item of evidence which ought to be reinforcing, but in fact contradicts the claim? If so, this is *counter-evidence* and will greatly diminish the assurance of the leaf claim.

In general, the assurance of a leaf claim will be increased if multiple items of evidence can be presented which reinforce each other. However, when determining the extent of this effect, it is necessary to consider the rigour (Section 3.5.2) of the evidence in question. That is, two items of evidence which reinforce each other and both have a high degree of rigour will have a greater positive impact on the assurance of the leaf claim than two reinforcing items of evidence which do not demonstrate this high rigour.

3.8.1 Counter-evidence

Counter-evidence refers to the provision of an item of evidence which has the potential to undermine a claim. Some standards, such as DS 00-56 Issue 4, mandate the search for counter-evidence. If found, counter-evidence will greatly reduce the assurance of a claim. The degree to which the assurance of the claim is reduced will be directly dependent upon the rigour (Section 3.5.2) of the counter-evidence. However, even counter-evidence with a relatively low degree of rigour will have a negative impact upon the assurance of a leaf claim – and consequently on the assurance of any claim supported by this leaf claim.

Furthermore, the provision of counter-evidence may have an effect on the assurance of other claims which are not themselves refuted by the counter-evidence. This is because the presence of counter-evidence which refutes a claim may lower the trustworthiness of any evidence E_2 which supports this claim. If E_2 is also used to support a second claim, then the assurance of this second claim may be lowered due to the lowered trustworthiness of E_2. Section 4.1.5 discusses this in more detail.

3.9 Leaf Claim Assurance

When determining the assurance of a leaf claim, it is necessary to take into account all factors discussed above, for all supporting items of evidence presented. With that in mind, the assurance of a leaf claim is dependent upon:

- The rigour (replicability, trustworthiness and coverage) of each item of evidence, allowing for:
  - The scope of this evidence
  - The user-defined importance of this evidence
- The independence of all evidence supporting this claim
- The degree to which all items of evidence are reinforced

By noting the parallels between this definition and that given in Section 2.2 we can deduce that, for evidence, assurance is interpreted as being rigour. This obser-

vation, combined with the decomposition of rigour as described in Section 3.5.2, provides us with a vocabulary for assessing whether evidence is 'good enough', or 'sufficient' to support a particular argument.

4 Separating and Combining Assurance Factors

In the previous section we discussed how the assurance of a leaf claim is dependent upon the rigour of the evidence presented, and we described factors which influence this dependence. Unfortunately, in the process of determining the assurance of a claim it is inevitable that information is lost. To see this, note that an item of evidence may be judged to lack rigour for three reasons: a lack of trustworthiness, replicability or coverage. All these situations will have the same end result – a decrease in rigour and therefore a decrease in the assurance of the leaf claim supported by this evidence. Consequently, given the situation $SC_1 * SC_2 \rightarrow HC$, if *HC* has a lower assurance than is required, it is difficult to immediately see how this could be rectified without 'propagating up' specific knowledge about how the assurance of SC_1 and SC_2 were determined.

To obviate this problem, we recommend propagating the elements of rigour (replicability, trustworthiness and coverage) up to higher-level claims. This allows us to retain as much information as possible, and to structure the argument in a way which best compensates for any deficiency. Consequently, we will speak of the trustworthiness factor of a leaf claim as being obtained from the trustworthiness of all supporting items of evidence. The trustworthiness factor of a higher-level claim *HC* is a combination of the trustworthiness factors of SC_1 and SC_2, in a degree which is proportional to the scope and user-defined importance of each. Similarly, the replicability and coverage factors of *HC* are a combination of the replicability and coverage factors of supporting claims SC_1 and SC_2 in a degree proportional to the scope and user-defined importance of each. Finally, we will also refer to the independence factor of *HC*; this is the degree of independence between the evidence gathered for supporting claims SC_1 and SC_2 (see Section 2.1.2 for more details).

4.1 Interaction of Assurance factors

While it is impossible to prescribe a single optimal strategy for constructing every argument, there are some general observations which can be made about the interaction of those factors which influence assurance. In the sections below, we discuss how to extrapolate from particular combinations of assurance factors to conclusions about how a safety argument should be structured. Similarly, we provide some examples of where particular combinations of assurance factors can appear to have a more pronounced effect on the assurance of a higher-level claim than

would actually be the case. These are anticipated to form the basis of anti-patterns (Weaver 2004), which are used to analyse common fallacious arguments. It should be emphasised that this section is not intended to be an exhaustive list of all possible interactions, but rather to demonstrate some of the 'flavour' of what is required in terms of propagating assurance upwards.

4.1.1 Trustworthiness and Independence

If a claim *HC* has a low trustworthiness factor (that is, the evidence which eventually supports this claim is not particularly trustworthy), then the assurance of *HC* will be decreased significantly if *HC* also has a low independence factor. That is, a combination of low trustworthiness and low independence factors will result in a significantly lowered assurance, perhaps to a greater degree than another combination of low assurance factors. The reason for this is that a low trustworthiness factor signals that the evidence has been gathered in a manner which would be likely to introduce errors. Furthermore, a low independence factor signals that all the evidence shares some common characteristics. That is, errors which affect one item of evidence are highly likely to affect the others. Consequently, combined low trustworthiness and independence factors signal that errors are likely to be present, and they are likely to affect all of the evidence presented. As a result, the assurance of *HC* will be lowered to a greater degree than would generally be the case.

4.1.2 Coverage and Trustworthiness

A low coverage factor for a claim means that there are parts of this claim that have not been addressed by any evidence, even though they are theoretically within the scope of the types of evidence which have been generated. This indicates that we might expect to see a correspondingly low trustworthiness factor for this claim. The reason for this conclusion is that the implementation of the evidence-gathering process was obviously not as thorough as expected (hence low coverage). This fact indicates that the evidence may have been generated in a careless manner (hence low trustworthiness). In other words, a low coverage factor combined with a high trustworthiness factor signals a possible deficiency in the assessment of the evidence.

4.1.3 User-defined Importance and Replicability

One of the most common examples of user-defined importance is the statement of a preference for a particular type of evidence (Ministry of Defence 2007). In many cases this preference can reasonably be judged to be due to this evidence being highly replicable (for example, formal methods and static analysis have a high de-

gree of replicability). In this situation, when considering the extent to which the assurance of a leaf claim depends upon the rigour of an item of evidence, the user-defined importance should not also be taken into consideration. If it were, then evidence with a high replicability factor would have a disproportionate effect on the assurance of a leaf claim. To understand this, note that firstly high replicability would – in the absence of all other factors – cause this evidence to be judged highly rigorous, thereby increasing the assurance of any leaf claim it supports. Furthermore the leaf claim would be dependent upon this evidence to a greater degree, due to its high user-defined importance. Correspondingly, the assurance of the leaf claim will be "increased twice" solely because the evidence is highly replicable. This type of 'double-accounting' can result in imbalanced arguments which disproportionately favour some claims or types of evidence.

4.1.4 Reinforcement and Independence

The assurance of a claim *HC* is dependent upon the degree to which items of evidence reinforce each other (otherwise known as convergent support). However, reinforcement by independent items of evidence will increase the assurance of a claim to a greater degree than reinforcement by evidence which lacks independence. To see this, consider an item of evidence supporting some claim and resulting from execution of a test suite. It is possible to run this test suite again to obtain a second item of evidence with which to reinforce the first – however, most people would judge that this would not noticeably increase their confidence in the claim! However, reinforcing these test results with evidence obtained from formal analysis (i.e. evidence which is conceptually independent) is likely to increase confidence in the claim. Thus, a high degree of reinforcement coupled with a high degree of independence will increase the assurance of a claim to a greater degree than would generally be the case.

4.1.5 Counter-evidence and Trustworthiness

If counter-evidence is found which has the potential to undermine a claim, then this finding may cast doubt on the trustworthiness of evidence which supports that claim, requiring a reappraisal of the trustworthiness of this evidence. This situation arises when the counter-evidence and the supporting evidence have identical scope – that is, they address the same aspects of the claim and would otherwise be assessed as reinforcing items of evidence.

For example, let E_1 be the results from a test suite, showing that no omission failures in the system have been detected from the tests. If E_2 is an item of counter-evidence showing that there are omission failures in the system that should have been detected by these tests (that is, E_2 and E_1 have identical scope) then this finding will lower the trustworthiness of E_1. Furthermore, the trustworthiness of any evidence sharing certain characteristics with E_1 will be lowered by this find-

ing. Using the example above, the trustworthiness of that test suite is called into question by the existence of counter-evidence. Consequently, all evidence resulting from iterations of that test suite will now be judged to lack trustworthiness. There are multiple types of counter-evidence, all of which undermine a claim in different ways (Toulmin et al. 1979). Thus, the effect of counter-evidence on each individual safety argument must be explicitly assessed.

4.1.6 Trustworthiness

Trustworthy evidence is evidence which is judged to be free of 'bugs' and which has been gathered in a manner which is unlikely to introduce errors. By contrast, the provision of untrustworthy evidence implies that evidence-gathering processes and assessments were not carried out with the necessary care. If one item of evidence is judged to have a low trustworthiness factor, it is reasonable to suppose that this should be the case for all other items of evidence with which it shares certain characteristics, such as a common origin. If this is not the case, then this signals a potential discrepancy in the assessment of evidence. In this way, low trustworthiness functions in a similar manner to counter-evidence (Section 4.1.5).

5 Conclusions

In this paper we have presented a framework for assessing and communicating the assurance of a safety argument. We have identified several factors which determine the confidence we have in the truth of a claim. We have also provided a vocabulary with which to discuss these factors. By making use of the concepts of scope, independence, user-defined importance and reinforcement, we can determine the extent to which any claim depends on those which support it. These concepts also aid us in constructing a safety argument, as they can be used to determine exactly where assurance deficits have been introduced.

Furthermore, we have provided guidance for assessing the quality of any evidence supporting a safety argument. This guidance is written in a way that clearly expresses why one item of evidence may be judged to provide less assurance than another. We have also defined rigour, and shown how the assurance of a claim is dependent upon the rigour of supporting evidence. By using the concepts of coverage, replicability and trustworthiness we have established criteria by which evidence of different types maybe compared. Finally, we have provided some examples of how assurance factors can interact, and discussed the effect these interactions have on the assurance of claims. The principles underlying these discussions may then be used to construct a justifiable and compelling safety argument.

5.1 Context and Further Work

The guidance on software in the context of DS 00-56 Issue 4, to be produced by the SSEI, has two primary audiences: MOD Integrated Project Teams (IPTs) and the industry supply chain, including Independent Safety Auditors (ISAs). The intention is to produce guidance for the supply side, which enables prime contractors to specify evidence requirements, to assess the evidence which is supplied, e.g. coverage versus scope, and to identify any assurance deficit. The guidance will also address arguments about the acceptability of any assurance deficit, probably as a form of ACARP argument and will build on the notion of assurance set out here, and will include questions to help elicit measures of assurance. The guidance will be supported with case studies illustrating the approach and, if practicable, argument patterns showing how the principles can be applied in practice. These patterns will be 'grounded' in evidence types, e.g. review results, test evidence.

The IPT guidance will dovetail with the supply side guidance, but will be presented in a way which does not require expert knowledge. The aim is to identify means of articulating assurance requirements, to help understand the assurance achieved at each milestone in the project, and to provide the ability to challenge what is being done to address any assurance deficit. This should enable IPT staff to engage effectively in assurance decisions, with the relevant experts.

There are many benefits of moving to goal-based, or evidence-based, standards not the least of which are avoiding situations where 'perfectly good' systems have to be re-engineered at high cost, but with minimal added value, because they do not meet some prescriptive standards. The downside is that it is difficult to articulate 'how much is enough' when it comes to evidence – and arguments, for that matter. This paper has outlined an approach to assurance which we plan to use to underpin guidance for software in the context of DS 00-56 Issue 4. We hope that this will help the MoD to realise the benefit of the standard, whilst reducing the uncertainty that can arise in using goal-based, or evidence-based, standards.

Acknowledgements The authors would like to thank the UK Ministry of Defence for their support and funding.

References

Adelard (1998) ASCAD - The Adelard Safety Case Development Manual. ISBN 0 9533771 0 5

British Standards (2001) EN 50128:2001, Railway Applications – Communications, Signalling and Processing Systems – Software for Railway Control and Protection Systems

Govier T (1988) A Practical Study of Argument. Wadsworth

Health and Safety Executive (2001) Reducing Risks, Protecting People. http://www.hse.gov.uk/risk/theory/r2p2.pdf. Accessed 15 September 2008

Kelly T (1999) Arguing Safety – A Systematic Approach to Safety Case Management. DPhil Thesis. Department of Computer Science Green Report YCST99/05

Ministry of Defence (1997): Defence Standard 00-55 Issue 2: Requirements for Safety Related Software in Defence Equipment

Ministry of Defence (2007) Defence Standard 00-56 Issue 4: Safety Management Requirements for Defence Systems

RTCA, EUROCAE (1992) Software Considerations in Airborne Systems and Equipment Certification. Radio Technical Commission for Aeronautics RTCA DO178B/EUROCAE ED-12B

Toulmin S, Rieke R, Janik A (1979) An Introduction to Reasoning. Macmillan Publishing Co., New York

Weaver R (2004) The Safety of Software – Constructing And Assuring Arguments. PhD Thesis. University of York

Weaver R, Fenn J, Kelly T (2003) A Pragmatic Approach to Reasoning about the Assurance of Safety Arguments. In: Proceedings of the 8th Australian Workshop on Safety Critical Systems and Software. Australian Computer Society, Darlinghurst

Weaver R, Despotou G, Kelly T et al (2005) Combining Software Evidence – Arguments and Assurance. In: Proceedings of ICSE-2005: Workshop on Realising Evidence Based Software Engineering