EC6

Methodological Guideline for Assurance Cases Evaluation

Document reference number for ANR

contact@confiance-ai.fr | www.confiance.ai

Document reference: 613C

Contributors

	Name	Organisation	Role
Responsible for the deliverable	Eric Jenn	IRT Saint Exupéry	EC6 Contributor
Scientific responsible	Eric Jenn	IRT Saint Exupéry	EC6 Contributor
Co-authors	Anthony Fernandes Pires	ONERA / IRT SystemX	Author
	Vincent Mussot	IRT Saint Exupéry / IRT SystemX	Reviewer
	Yassir ID MESSAOUD	IRT SystemX	Reviewer

Document control

Revision	Date	Commentary	Author
1.0a	22/12/2023	Version for review	See co-authors
1.0	05/01/2024	Version for delivery	See co-authors

Table of contents

A. Introduction and abstract 4

A.1 General introduction to trustworthy AI challenges 4
A.2 Context 4
A.3 Target audience and disclaimer 5
A.4 Summary of the limitations on this evaluation methodology and perspectives for evolution 5
A.5 Document organization 5
A.6 How to use the document 6

B. Description of the method 7

B.1 Context of the Evaluation 7
B.2 Prerequisites to the Evaluation 7
B.3 Evaluation Procedure 7

C. Conclusion 14

C.1 Checklist for the evaluation 14
C.2 Limitations on this evaluation methodology and perspectives for evolution 15

D. Bibliography 17

A. Introduction and abstract

A.1 General introduction to trustworthy AI challenges

Trustworthiness in AI within critical systems (systems that can directly or indirectly affect human life and moral entities) is essential for its widespread adoption (by the industry, the decision makers, the general public, etc.) and poses the following significant challenges.

- First, how to design AI models, so that, by construction, they satisfy trustworthy properties (accuracy, robustness…).
- Secondly, how to characterize these AI models, for example to understand and explain their behavior and their adequacy to the operational domain.
- Then, how to implement and embed those AI models on hardware, by making them fit for the target without losing their trustworthy properties.
- Another question is, what methods of data engineering to apply in order to, among other topics, manage important volumes of data and adapt to the evolution of the operational domain.
- At system level, what verification and certification processes to consider specifically for AI-based systems.
- Finally, a federation of all these matters is necessary to build an end-to-end methodological approach, supported by a consistent engineering environment compatible with industrial practices.

These are the challenges, among others, that the Confiance.ai program addresses.

A.2 Context

An assurance case (AC) is a structured argument used to justify a desired claim (e.g., The system is safe, reliable, robust, …), based on evidence(s) concerning both the system and the environment in which it operates. As stated by [Rushby, 2015] an argument might be subject of imperfection. Indeed, the evidence used to defend a claim can be questioned by either questioning its truthfulness or relevance to the case. Moreover, the goal of having a perfectly sound argument, for example about a property of a system or a function, outside a highly monitored and controlled environment, is unrealistic. Hence the idea of minimizing uncertainty. To judge the quality of an insurance case, it is important to assess the validity of the evidence used in the argument and the reasoning behind it, through the inferences made between the different stages of the argument. Evaluation is a critical activity in its life cycle. This activity often refers to the review of the assurance case by the certification authority to decide if it is acceptable or not. However, it is important to gain confidence in the Assurance Case and detect problems early on in its life cycle [Kelly, 2007], before reaching the certification authorities. Recent works such as [Id Messaoud, 2022] are dealing with the assessment of uncertainty of the assurance cases to increase confidence in the argument. Those works provide two approaches for propagating epistemic and aleatoric uncertainty in AC. The first approach is based on a continuous uncertainty value in the interval [0, 1] while the second approach uses a scale of linguistic qualifiers for uncertainty. For both approaches, uncertainty propagation consists in applying formulae involving uncertainty related to elements of the structure, such as support relations and goals directly linked to solution, to compute uncertainty of the desired claim. Methods are provided for eliciting uncertainty related to elements of the structure for both approaches.

In our context, we aim at evaluating the uncertainty and validating the reasoning of ACs that have already been internally pre-reviewed. The method proposed by [Id Messaoud, 2022] will be adapted to the specificities of our context. First, the form of the targeted ACs are closer to “patterns” to be instantiated by the applicant in the sense that the applicant may have to make choices between different argumentation paths during the instantiation. Still, we wish to obtain a first uncertainty evaluation for the pattern to inform the applicant on the confidence of the argument and to guide its instantiation. Second, we consider that the ACs has been pre-reviewed internally and their structure and their reasoning are “pre-evaluated” correct. It allows the evaluators to have a sound basis to analysis. Nevertheless, the results of the uncertainty evaluation may reveal issues in the argument, e.g. a very low confidence in an argumentation step may indicate a missing/wrong premise or an issue in the reasoning.

In this document, we propose guidelines for conducting such evaluation with industrial and academic experts, which we differentiate as follows:

- Experts with an industrial background are able to consider the assurance cases as part of their own systems, consider the availability of the engineering items and estimate the feasibility of the future instantiation of the AC.
- Experts with an academic background can have more critical review of the soundness of the science and the deep mechanics behind the argument for specific properties of interest (e.g. *Robustness* or *Explainability of neural networks*).

Establishing evaluation sessions with experts requires a thorough preparation, especially since the availability of experts is limited and the ACs structure can be challenging in terms of size. However, we will not discuss the details of the adaptation of the method of [Id Messaoud, 2022] for our context.

A.3 Target audience and disclaimer

The target audience of this guideline are the persons in charge of the evaluation of the Assurance cases. In our case, it corresponds to the Assurance Case Developers. The only pre-required knowledge to understand this guideline is the vocabulary and definitions associated with the Assurance cases.

A.4 Summary of the limitations on this evaluation methodology and perspectives for evolution

The limitations of this evaluation methodology concern the lack of two stages. First, the methodology could include a stage for the critical review of the AC, i.e. a stage where a group of experts tries to defeat the structured argument. This stage should happen before the evaluation process in this document. Second, the evaluation process should integrate a stage including an evaluation or a discussion with representatives of certification authorities in order to collect their opinion on the AC from the point of view of certification.

In terms of perspectives for evolution, the addition of these two stages to the process given by the methodology and their descriptions are a priority for this guideline.

A.5 Document organization

In the next chapter, we introduce how we organized the sessions based on related works coming from the ACs field but also coming from other fields that have the habit to deal with evaluation sessions involving people, such as in social sciences [Nachmias, 2008] or experts such as in enterprise modelling [Janis Stirna, 2018]. We conclude the document with a summary of the plan to use for the evaluation.

A.6 How to use the document

This document intends to provide a list of activities to fulfil in order to evaluate an Assurance Case. For each activity, a rationale of the content of the activity is given along with a checklist of tasks to accomplish. The reader needs to follow the checklist for each activity in order to conduct the evaluation.

B. Description of the method

B.1 Context of the Evaluation

This methodology guideline does not intervene in the End-to-End Approach but allows evaluating ACs that would be part of the End-to-End approach. It is specially designed to deal with ACs developed following the Assurance Cases guidelines document released by EC6.13. However, the methodology can be customized to deal with other types of Assurance Cases.

This evaluation methodology is based on concepts coming from the ACs literature but also from the recognized literature of other domains (e.g. usability testing, enterprise modelling, etc). At the time of the writing, this evaluation methodology has been applied and experimented on one Assurance Case with a single academic expert. Nevertheless, the methodology will be applied with other Assurance Cases and several experts in the scope of the batch 4 of the Confiance.ai program.

B.2 Prerequisites to the Evaluation

B.2.1 Choice of participants

To perform the evaluation, evaluators must be selected. These evaluators must fulfil some requirements.

First, they need to be expert of the field. Otherwise, their opinion on the validity of the ACs cannot be trusted. Second, they cannot be part of the ACs development team. As Rushby and al. explain in [Rushby, 2015], there is a concern that if the developers were to do the assessment, they will most likely try to justify the ACs rather than challenging it. This is a confirmation bias. On the other hand, it is in the human nature to challenge and to criticise the argument of someone else.

Third, the chosen expert(s) must represent either the industrial domain or the academic domain. It is a known fact that industrial and academic often offer different points of view to the same research problem. It is therefore highly valuable to be able to capture both points of view. Ideally, the evaluation could be done in a consensus fashion (i.e. a group of experts trying to reach a common answer). In that case, it is preferable to make sure that there is no personal animosity between the people of the group in order to avoid moderation during the evaluation or bias in the result.

B.3 Evaluation Procedure

B.3.1 Overview

As shown in Figure 1, the suggested evaluation procedure is structured in five stages for each AC:

- Presentation stage
- Evaluation stage
- Answers Analysis stage
- Debriefing stage
- Closing stage

Note that depending on the result of the debriefing stage, the procedure can be iterated with additional experts before converging to the closing stage. Indeed, in case the corrections made to the AC during the debriefing stage are significant, the expert becomes involved in the AC development. Thus, the evaluation of the corrected AC shall be performed by other experts. Note also, that among the criteria for expert selection the most important criterion is the expertise of the field, then the diversity among academics and industrial, and finally the absence of animosity. This last criterion is quite difficult to check with certainty, however detected presence of animosity could lead to evaluation issues. Finally, even if the procedure involves several experts, during the evaluation stage each questionnaire is filled with no access to the questionnaires filled by other experts.

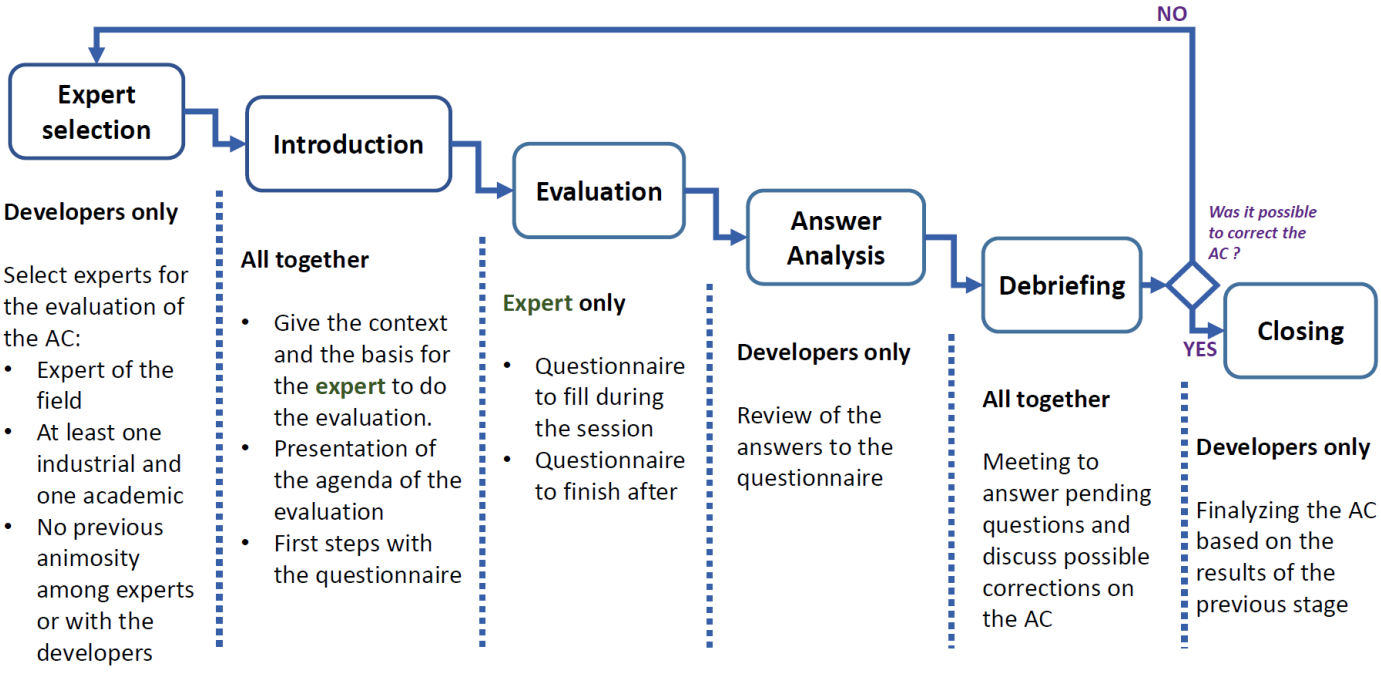

Figure 1: Assurance Case Evaluation Method

B.3.2 Presentation Activity

This stage is here to introduce the evaluation process to participants and make them comfortable in doing the evaluation.

First, the purpose of the evaluation is presented to each participant. This helps them understand the importance of the evaluation. For example, in social sciences, an introduction statement always supports the use of survey methods to collect data [Nachmias, 2008]. This introduction statement explains the purpose of the survey and its importance in order to guarantee a high response rate. A similar introduction is also done during interview for preparing participatory enterprise modelling session and make the participant understand the purpose of the modelling and agree on its importance [Janis Stirna, 2018]. Even if they already agreed to participate in our evaluation, giving them visibility on the purpose of the AC and its evaluation can help consolidate their involvement and motivation in the task.

Second, once the purpose of the evaluation is clear, the evaluation process itself is described. It is important that the participant is aware of what is going to happen during the evaluation [Janis Stirna, 2018] and how the results are going to be used. It might be necessary to ensure the anonymity of each participant with respect to the project to avoid bias in the answer (i.e. experts might be reluctant to give their real opinion if they know they will be under scrutiny in the future).

Third, the AC to evaluate is introduced to the participant. As stated in by T. Kelly in his Staged Argument Review Process [Kelly, 2007], the first step of an argument review is to understand the presented argument. In the same vein, we have to make sure that the participant has the tools to understand the AC under evaluation. A reminder of the basic concepts of Goal Structuring Notation (GSN) may be necessary if the participant is not familiar with it, followed by a presentation of the AC itself and the way to read it. Note that it is decided that this presentation will be done top (i.e., from the high property argued) to bottom (i.e., to the provided evidence) on the AC. The later, makes the reading of an assurance case and the understanding of its reasoning easier.

Finally, the next step of the evaluation process is presented in more detailed to each participant and the evaluation rules are introduced. These rules will be presented in the next section.

To summarise, the Presentation stage consists in:

- **Introducing the session**
  - Present the agenda of the presentation session
  - Introduce each other
- **Introducing the evaluation process**
  - Introduce the goal of the work done so far and the importance of the evaluation
  - Present the agenda of the overall evaluation and its rules
  - Give the calendar of the different stages
- **Introducing the AC to evaluate**
  - Introduce/remind the basic concepts to understand the graphical notation
  - Introduce the AC and how to read it (walk-through)
- **Introducing the next stage**
  - Remind the next step and the rules for the evaluation
  - Present the questionnaire to use for evaluation

B.3.3 Evaluation Activity

This stage starts at the end of the Presentation stage. For the evaluator, it consists in filling a questionnaire adapted from the work of [Id Messaoud, 2022] in order to evaluate the uncertainty of the AC.

Each AC can be significantly large and complex (dozens of goals). It may be cognitively challenging for an expert to have to evaluate it all at once during a face-to-face session. Moreover, it can also be constraining for the expert with limited availability to book long sessions to carry out the evaluation. In this context, the expert is allowed proceeding to the evaluation freely, on its own, outside of face-to-face session. The only

constraint applying will be a deadline to send back the results. This choice is a compromise between quality and efficiency: the evaluator will have less support from the development team in comparison with a face-to-face session, but he or she will be more flexible on managing the time to evaluate the AC. One could also argue that having more time to evaluate an argument can allow the user to deepen his or her thinking. In any case, the lack of support in comparison with a face-to-face session can be compensated by email exchanges and a face-to-face debriefing session coming afterwards. This session is described in the next sections.

In other words, the evaluation stage will proceed as follows. Each participant starts the evaluation after the Presentation stage with a global picture of the AC to evaluate, the questionnaire to fill and rules on how to fill it. Experts unable finishing during the face-to-face session following the Presentation stage, can proceed outside of the session and an adequate amount of time is given to them to fill the questionnaire and send it back to the team.

The questionnaire is adapted from [Id Messaoud, 2022]. In this questionnaire, a set of questions is asked to the evaluator for each goal, from top to bottom. These questions evaluate the uncertainty for the reasoning between a goal and its sub-goals. They do not evaluate the relationship between a solution and a goal. In addition, for each goal yes/no questions associated to open questions let the evaluator freely express its opinion on the argumentation step:

- Can you assess this argument? If not, could you give the reason(s) why?
- Do you think that the argument is incomplete or needs some improvement? If yes, could you give some details?

For filling the questionnaire, different modes are considered:

- Mode 1: each participant answers the questionnaire individually
- Mode 2: an industrial expert is paired with an academic expert to jointly answer the questionnaire. It is up to them to organise the joint sessions to carry out the evaluation

But in practice, mode 1 is more realistic because experts are usually overbooked and will be the default mode.

To summarise, the Evaluation stage consists in:

- **Providing a global picture of the AC, the questionnaire and the rules of evaluation**
  - The questionnaire allows evaluating the AC from top to bottom
  - The developers can only answer questions about where to find information
  - The developers cannot answer questions about its own opinions or the rationale of the AC

B.3.4 Answer Analysis Activity

This stage starts once all the filled questionnaires are retrieved from the evaluator(s). First the answers to the open questions are analysed. Arguments that are not assessed are detected and default uncertainty values are provided depending on the reason of absence of assessment. Each answer is then analysed

and compared with other expert's answers when more than one is selected. Special attention is paid to low confidence goal evaluation and inconsistency in the evaluation. Typical inconsistencies to be checked are:

- Strong acceptation or rejection associated to low confidence.
- Decision concerning a conjunction of supports weaker than decisions concerning each support.

After correcting data from incompleteness and inconsistencies, the uncertainty of each goal and of the whole AC are computed.

Following this analysis, a discussion plan is established to guide the debriefing sessions. This discussion plan is important as it allows the team to moderate the future discussion and ask for clarification on the results of the questionnaire.

To summarise, the Answer analysis stage consists in:

- **Retrieving the questionnaire results**
- **Analysing and comparing the answers**
  - Evaluate the uncertainty for the AC and each goal
  - Identify low confidence points
  - Identify inconsistency in the evaluation
  - Analyse the answers to the open questions
- **Establishing a follow-up questionnaire/discussion plan for the next stage**
  - Allow to guide the debriefing stage
  - The debriefing session needs to be booked at this stage or even directly at the Presentation stage

B.3.5 Debriefing Activity

In this stage, a debriefing session is organised with each participant to go through his or her answers and to clarify ambiguous ones, and to discuss remarks reported by the expert during the analysis stage. This stage makes the link between the understanding of the evaluator and the understanding of the development team. It is even more crucial here, as the participant may have conducted the evaluation stage on its own.

This kind of debriefing session often happens in the field of Usability testing, where the participant has to evaluate a product through a series of tests and questionnaires while being observed. As stated in [Rubin, 2008], the debriefing session is the final opportunity to understand the rationale behind the participant actions and answers during the test. In that sense, the authors of [Rubin, 2008] offer some guidelines on how to behave during a debriefing session. Firstly, they suggest to never making a participant feel defensive about its opinions. The session should feel like a discussion among peers, not an interrogation. Secondly, they suggest to not reacting to the participant's answer in one way or another while questioning. It can lead the participant to answer future questions in a specific direction rather than its true opinion. The authors recognised that this aspect could be very difficult if the questioner is closed to the evaluated product.

Note that if problems are detected in the reasoning of the assurance case, a possible correction has to be discussed, validated and assessed in terms of uncertainty during this session. This is why it is crucial to have analysed and compared the answers of all the experts before starting this debriefing stage.

As it is an interactive session such as the Presentation stage, it is necessary to start the session by presenting the agenda and introduce its purpose. This session is better done physically but can also be done remotely if needed. The question of doing a common session with all participants or individual sessions can be raised. Doing the session individually allow getting deeper in details for each expert answer but if update on the AC are needed, it may require additional validation by the other participants. Doing it as a group may make the participant uncomfortable to see his or her answers exposed and discussed with the others. There also a risk that we cannot go in details in the points to discuss or that the discussion diverged due to debate between participants. In this case, a strong moderation may be necessary. But if updates on the AC are needed, we can validate and re-evaluate these updates more efficiently.

In practice, the debriefing session will be done with all the participants present to make sure that a validated AC is obtained at the end of the session. The session could also be divided in numerous sessions to deal with all the points identified during the analysis session. The duration of the sessions depends on the availability of the experts and the number of points to be discussed.

In the case where the AC cannot be corrected and needs a deep modification of its structure, the AC will need to be revised by the developers before going to a new round of evaluation process, with possibly a selection of new experts to avoid any bias. This case corresponds to the feedback loop in Figure 1. Otherwise, the closing stage can start.

To summarise, the Debriefing stage consists in:

- **Introducing the session**
  - Remind the goal of the session, make the participant comfortable
  - Present the agenda
- **Answering participants pending questions**
- **Discussing issues or uncleared points discovered during answers analysis**
  - It is not an interrogation/trial
  - Do not make the participants feel defensive or influence their answer by your reaction
- **Discussing possible corrections to apply on the AC**
  - Decide corrections/update on the AC
  - Evaluate the confidence on the corrected nodes
  - Make sure there is no side-effects from the correction

B.3.6 Closing Activity

This stage marks the end of the evaluation process for an Assurance Case. It will give an evaluation of the uncertainty of the Assurance cases based on the combined answers of each expert and an internal review to ensure that all the updates discussed in the previous stage have been considered.

Note again that in the case where critical issues have been found in the assurance case, a second round of review might be necessary after the AC rework, and the validation of the assurance case can be delayed.

To summarise, the Closing stage consists in:

- **Computing the final evaluation of the uncertainty of the AC**
- **Validating the AC reasoning**

C. Conclusion

C.1 Checklist for the evaluation

The final plan for ACs evaluation is the following:

Choice of evaluators

- Recognised domain experts
- External to the development team
- Industrials and academics
- If group evaluation, no existing conflict between participants

Presentation stage

- **Introducing the session**
  - Present the agenda of the presentation session
  - Introduce each other
- **Introducing the evaluation process**
  - Introduce the goal of the work done so far and the importance of the evaluation
  - Present the agenda of the overall evaluation and its rules
  - Give the calendar of the different stages
- **Introducing the AC to evaluate**
  - Introduce/remind the basic concepts to understand the graphical notation
  - Introduce the AC and how to read it (walk-through)
- **Introducing the next stage**
  - Remind the next step and the rules for the evaluation
  - Present the questionnaire to use for evaluation
  - Book the Debriefing session if possible

Evaluation stage

- **Providing a global picture of the AC, the questionnaire and the rules of evaluation**
  - The questionnaire allows evaluating the AC from top to bottom
  - The developers can only answer questions about where to find information
  - The developers cannot answer questions about its own opinions or the rationale of the AC

Answers Analysis stage

- **Retrieving the questionnaire results**
- **Analysing and comparing the answers**
  - Evaluate the uncertainty for the AC and each goal

  - Identify low confidence points
  - Identify inconsistency in the evaluation
  - Analyse the answers to the open questions
- **Establishing a follow-up questionnaire/discussion plan for the next stage**
  - Allow to guide the debriefing stage
  - The debriefing session needs to be booked at this stage or even directly at the Presentation stage

Debriefing stage

- **Introducing the session**
  - Remind the goal of the session, make the participant comfortable
  - Present the agenda
- **Answering participants pending questions**
- **Discussing issues or uncleared points discovered during answers analysis**
  - It is not an interrogation/trial
  - Do not make the participants feel defensive or influence his or her answer by your reaction
- **Discussing possible corrections to apply on the AC**
  - Decide corrections/update on the AC
  - Evaluate the confidence on the corrected nodes
  - Make sure there is no side-effects from the correction

Closing stage

- **Computing the final evaluation of the uncertainty of the AC**
- **Validating the AC reasoning**

C.2 Limitations on this evaluation methodology and perspectives for evolution

The limitations of this evaluation methodology are mainly the lack of two stages. First, it would be better to include a stage for the critical evaluation of the AC, i.e. a stage where a group of experts tries to defeat the structured argument. This is an important stage as it is a guarantee of the strength of the argument. In our case, we considered that this stage already happened before the evaluation process presented in this document but a formal inclusion of this stage in the methodology along with its characterization could be an improvement of this guideline.

Second, the evaluation process should integrate a stage including an evaluation or a discussion with representatives of certification authorities in order to collect their opinion on the AC from the point of view of certification. Indeed, as the final purpose of an AC is to be used and instantiated during AI projects, it would be preferable to ensure the AC makes sense for certification authorities.

In terms of perspectives for evolution, the addition of these two stages to the process and their description is a priority for this guideline. In particular, both stages should be positioned in the process and the different

tasks to perform for each of this stage along with a description of the organization of the different sessions to organized should be added. In addition, this methodology will need to be experimented with further on multiple Assurance Cases and refined accordingly.

D. Bibliography

[Id Messaoud, 2022] Id Messaoud, Y. (2022). Uncertainty assessment in safety argument structure : an approach based on Dempster-Shafer theory. LAAS UPS Toulouse.

[Id Messaoud, 2022] Id Messaoud, Y. J. (2022). Questionnaire for estimating uncertainties in assurance cases. Toulouse: LAAS/CNRS.

[Janis Stirna, 2018] Janis Stirna, A. P. (2018). Enterprise Modeling: Facilitating the Process and the People. Springer Cham.

[Kelly, 2007] Kelly, T. (2007). Reviewing assurance arguments-a step-by-step approach. Workshop on assurance cases for security-the metrics challenge, dependable systems and networks (DSN).

[Nachmias, 2008] Nachmias, D. a.-N. (2008). research methods in the social sciences, seventh edition.

[Rubin, 2008] Rubin, J. a. (2008). Handbook of usability testing: how to plan, design and conduct effective tests. John Wiley & Sons.

[Rushby, 2015] Rushby, J. a. (2015). Understanding and evaluating assurance cases. NASA Langley Research Center: NASA Contractor Report NASA/CR-2015-218802.

Title : Methodological Guideline for Assurance Cases Evaluation

Keywords : Assurance case, evaluation, guidelines

Evaluation is a critical activity in the life cycle of an Assurance Case. This activity often referred to the review of the assurance case by the certification authority to decide if it is acceptable or not. However, it is important to gain confidence in the Assurance Case and detect problems early on in its life cycle.
In our context, we aim at evaluating the uncertainty and validating the reasoning of Assurance Cases that have already been internally pre-reviewed. In this document, we propose guidelines for conducting such evaluation with industrial and academic experts.

Our partners