EC6

Scientific Contribution to Uncertainty Assessment in Assurance Cases

Document reference number for ANR

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

Document reference: 613D

Contributors

	Name	Organisation	Role
Responsible for the deliverable	Eric Jenn	IRT Saint Exupéry	EC6 contributor
Scientific responsible	Jean-Loup Farges	IRT System-X, ONERA	EC6 contributor
Co-authors	Yassir Id Messaoud	IRT System-X	EC6 contributor
	Vincent Mussot	IRT System-X, IRT Saint Exupéry	EC6 contributor
	Anthony Fernandes Pires	IRT System-X, ONERA	EC6 contributor
	Florent Chenevier	IRT System-X, Thales AVS	EC6 contributor
	Ramon Conejo Laguna	IRT System-X, IRT Saint Exupéry	EC6 contributor

Document control

Revision	Date	Commentary	Author
1.0	30/11/2023	First release for review	Jean-Loup Farges
1.1	20/12/2023	External review taken into account	Yassir Id Messaoud

Table of contents

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

Of course, all trustworthy AI challenges are applicable to Machine Learning (ML). The development of functions based on ML requires a huge effort to prove their dependability. Convincing people that a function complies with its requirements is a matter of building and exposing a rigorous argumentation. Among many dependability techniques available today, Assurance Cases (AC) is one way to formalize and structure this argumentation. These cases could be based on Goal Structuring Notation (GSN). Basic elements such as goals, solutions and strategies are combined in a tree supporting a root goal claiming that the function complies with one of its requirements. Thus, AC is a relevant technique to argue that an ML model complies with a requirement derived from the desired satisfaction of a trustworthy property. To streamline and normalize the design of AC, AC templates are proposed. The difference between an AC and an AC template lies in the fact that (1) AC templates have argumentation choices that are not frozen while ACs present no choice and (2) AC templates present some not yet existing solutions while ACs solutions are concrete objects.

However, AC only provides a qualitative analysis and no estimation of the uncertainty related to the satisfaction of the trustworthy property is available. This is a serious drawback.

A.3 Introduction

The objective of the research presented here is to add uncertainty or confidence to AC templates.

Indeed, the final objective of uncertainty assessment in instances of AC is to provide certification authorities with an AC presenting a full belief assessment grounded on concrete solutions. Of course, this belief assessment would need to indicate that the trustworthy property of the AC is highly likely.

Two main possibilities exist for generating a specific AC from an AC template. The first possibility consists in considering the AC template only once at the very early beginning of the ML model development process. In that case, the ML model developer must make assumptions about the likeliness of goals associated to not yet existing solutions and based on those assumptions must freeze all choices of the AC template. Freezing those choices configurates the workflow of the ML model development. The assumptions made become likeliness requirements for the goals associated to the concrete solutions to be produced by the workflow. However, workflows including activities of type “select” or “or” are proposed for ML model development. Moreover, the choices made by those activities may impact the choices to be frozen in the AC template. Thus, a second possibility for generating a specific AC may include intermediate steps with intermediate uncertainty assessment objectives at design, implementation, verification or validation stages of the ML model development process. In that case, at the beginning of the ML model development process, the developer only relies, for the verification and validation aspects, on an AC template that provides choices in a pre-defined tree structure. Some of those choices are linked to some workflow activities. Then, as the development of the ML model progresses, choices are made and frozen while executing those activities. As shown in Figure 1, in that context the AC is used inside the activity to assess the consequences of the different options of the choice and at the end of the activity the AC is pruned according to the result of the choice. Other activities instantiate the AC by providing concrete definitions and solutions. In conclusion, for that second possibility, the AC helps in driving the workflow while the workflow refines the AC.

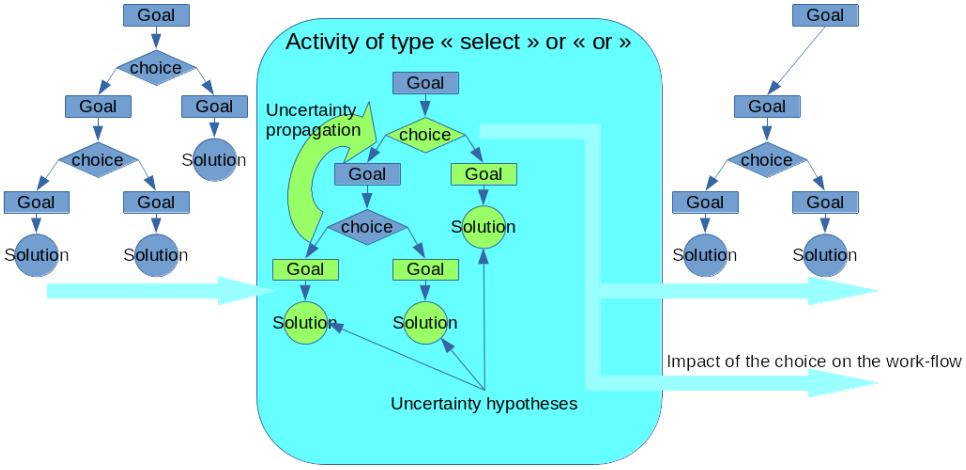


Figure 1. pruning of the AC by an activity of type “select” or “or” of a workflow

The difficulty for making decisions is high when the subject of the AC is a new technology where goals may be reached with many approaches with different levels of readiness, as it is the case for robust ML. In those cases, an uncertainty assessment can be useful for making a judgment about the pertinence of using a specific approach. Moreover, uncertainty assessment of each strategy in the tree structure may be performed at no cost and could be directly provided with the AC template.

At the opposite, the evidence at some leaves of the tree may be subject to a change in its uncertainty assessment: The evidence will be provided at no additional cost by the chosen design process but the uncertainty before producing it may be different from the uncertainty after producing it and according to the

choice made in the AC template, the evidence must be provided independently from the design process by the verification and validation process. This last kind of evidence may be costly.

The objective of the research presented in this report raises several issues: Choice of an uncertainty representation, elicitation of uncertainty associated to atomic elements such as relations and solutions, and propagation of the uncertainty of atomic elements through the AC. Working with AC templates that will become instantiated as actual AC is also quite challenging.

The approach followed here is based on recently published results [Id Messaoud, 2022a; Id Messaoud, 2022b] and brings the following contributions:

(i) An uncertainty assessment based simultaneously on qualitative and quantitative uncertainty modeling

(ii) an elicitation method allowing simultaneous capture of qualitative and quantitative uncertainty

(iii) an analysis of uncertainty modeling and propagation on AC templates and

(iv) demonstration of the approach on a case study related to robustness of ML models.

The next chapter presents the state of the art and provides the basis for understanding this report. Chapter C is related to the method used here, that is the choice of an approach for uncertainty modelling, the data capture method, the data analysis and uncertainty propagation method and finally the demonstration of the method on a case study. Then the data collected from an ML expert on the case study is presented in Chapter D. Chapter E provides the results obtained on the case study using the data analysis and uncertainty propagation method. Finally analysis of results and conclusions are given in chapters F and G.

B. State of the art

The mathematical grounding of the approach presented here are the theory of Dempster-Shafer and the theory of capacities. Those are presented in the first section. Then a section deals with the sources of uncertainty in AC and the type of arguments that can be addressed. Finally, the last section presents the main characteristics of the work over which the methodology is grounded.

B.1 Dempster-Shafer and qualitative capacities theories

This section presents theories on which the uncertainty/confidence propagation and elicitation models are based.

B.1.1 Dempster-Shafer theory

Dempster-Shafer Theory (DST) [Shafer, 1976] (or theory of evidence) is a generalization of probability theory which offers tools to model both aleatoric (due to random events) and epistemic (due to incomplete information) uncertainty. This theory is more adapted to argumentation than probability theory because the uncertainty identified in an argument is often due to the lack of information. In probability theory, epistemic uncertainty is modeled with a uniform probability distribution over all elements of the frame of discernment Ω (i.e. universe of all possibilities). For example, if we do not know if it will rain tomorrow, we will assume that: $P(rain) = P(not\, rain) = 1/2$. However, one cannot distinguish between the situation where no information is available and the case where we have as much information to support the claim that it will rain tomorrow or deny it. When no information is available, it is likely that the probability of rain is biased.

A mass function, or Basic Belief Assignment (BBA), is a probability distribution over the power set of Ω $m: 2^{\Omega} \longrightarrow [0,1]$ is such that:

$$\sum_{E \subseteq \Omega} m(E) = 1, m(\emptyset) = 0$$

Any subset E of Ω such as $m(E) > 0$ is called a *focal set* of m. $m(E)$ quantifies the probability that we know that the truth lies in E, in particular $m(\Omega)$ quantifies the amount of ignorance. Hence this approach handles incomplete information in an unbiased way.

Example:

Let us consider the case of a light bulb. The frame of discernment $\Omega = \{On, Off\}$ includes the two possible states of the latter “lights on” or “lights off”.

- $m(\{On\})$: Quantifies the probability that the light bulb is **“On”**.
- $m(\{Off\})$: Quantifies the probability that the light bulb is **“Off”**.
- $m(\Omega)$: Quantifies ignorance on the state of the light bulb **“On” or “Off”**.
- $m(\emptyset)$: Quantifies contradiction. I.e., **“On” and “Off” at the same time**. DST supposes that this mass is initially null. However, it may have a non-null value when masses from different sources are combined.

A mass assignment induces a so-called belief function Bel: $2^{\Omega} \longrightarrow [0,1]$, defined by:

$$Bel(A) = \sum_{E \subseteq A, E \neq \emptyset} m(E)$$

that represents the sum of all the masses (evidence) supporting a statement A.

Belief in the negation $\neg A$ of a statement A, or *disbelief*, is denoted by $Disb(A) = Bel(\neg A)$.

The value $Uncer(A) = 1 - Bel(A) - Disb(A)$ quantifies the lack of information about A.

Mass functions also induce a so-called plausibility function $Pl: 2^{\Omega} \longrightarrow [0,1]$, defined by:

$$Pl(A) = \sum_{E \cap A \neq \emptyset} m(E)$$

Belief and plausibility functions are related: $Pl(A) = Bel(A) + Uncer(A) = 1 - Disb(A)$. They respectively represent the lower and upper limits of probability: $Bel(A) \leq P(A) \leq Pl(A)$. Figure 2 represents the relation between these metrics.

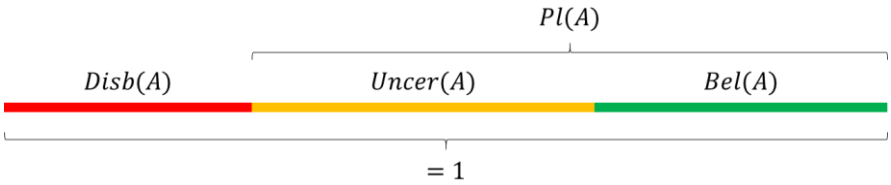


Figure 2. Representation of the 3-tuple (belief, disbelief, uncertainty)

To obtain the uncertainty in proposition A, two pieces of information are needed:

- $(Bel(A), Pl(A))$ which gives direct information about the uncertainty range, or
- $(Bel(A), Disb(A))$ which gives the truth is most likely to be: $A, if\ Bel(A) > Disb(A)$ or $\neg A, if\ Disb(A) > Bel(A)$.

To merge pieces of evidence (masses) from different sources a conjunctive combination rule can be used such as:

$$(m_1 \otimes m_2)(A) = m_{12}(A) = \sum_{E_1 \cap E_2 = A \neq \emptyset} m_1(E_1) \cdot m_2(E_2)$$

There exist merging conflicts when one source has mass on a set and the other source has mass on sets that are disjoint from the first set. Dempster rule of combination normalize the expression above by $1 - \sum_{E_1 \cap E_2 = \emptyset} m_1(E_1) \cdot m_2(E_2)$ to eliminate the effect of conflicts. However, when conflict is important this normalization has no sense.

B.1.2 Qualitative capacities theory

In contrast to a basic probability assignment (BPA or mass) in DST, a Basic possibility (Π) Assignment (BΠA) [Dubois, 2019] is a possibility distribution ρ: $2^{\Omega} \longrightarrow L$ over the power set of the frame of discernment Ω, such $max_{B \subseteq \Omega} \rho(B) = 1$ and $\rho(\emptyset) = 0$. L is a finite totally ordered set representing certainty levels. The value $\rho(B)$ is the strength of a piece of evidence B.

A qualitative capacity (q-capacity, for short) is a function: $\Upsilon: 2^{\Omega} \longrightarrow L$, such that: $\Upsilon(\emptyset) = 0, \Upsilon(\Omega) = 1$ and $A \subseteq B \Rightarrow \Upsilon(A) \leq \Upsilon(B)$. Any q-capacity can be put in the form:

$$\Upsilon(A) = max_{\emptyset \neq B \subseteq A}\, \rho(B), \forall A \subseteq \Omega$$

The value $\Upsilon(A)$ (resp. $\Upsilon(\neg A)$) quantifies the support in favor of (resp. against) A, i.e., belief (resp. disbelief) in A using an element in the qualitative scale L. The pair $(\Upsilon(A), \Upsilon(\neg A))$ thus describes our epistemic stance with respect to A in terms of belief and disbelief ranging from no information (i.e., $(0,0)$), to full conflicting information (i.e., $(1,1)$) and from full belief (i.e., $(1,0)$), to full disbelief (i.e., $(0,1)$). This is more general than possibility theory where the case $(1,1)$ is not allowed. In the quantitative setting (i.e., DST), full conflict situation is represented with the pair $(0.5,0.5)$.

To merge pieces of evidence (BΠA), the following alternative to the quantitative combination rule can be used:

$$(\rho \odot \rho_2)(A) = \rho_{12}(A) = \max_{E_1 \cap E_2 = A} min\,[\rho_1(E_1), \rho_2(E_2)]$$

B.2 Assurance cases and sources of uncertainty

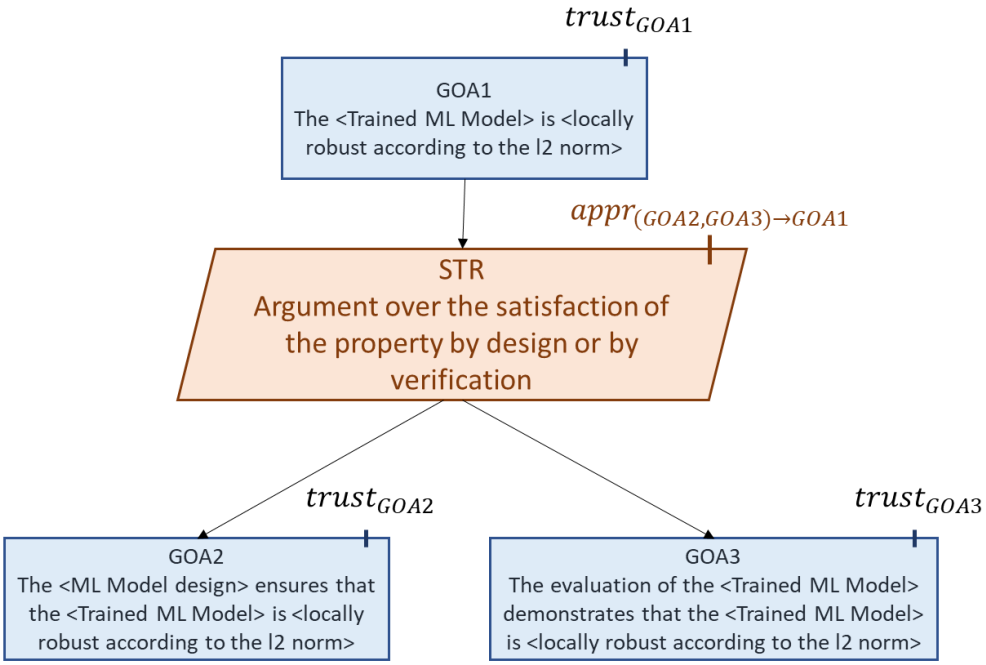
In this section, we present the source of uncertainty in an argument. Then, we specify the different argument types that one may encounter in an assurance case and formally define them using logical expressions. In the following, we refer to top goals as conclusions and sub-goals, particularly those linked to solutions, as premises.

B.2.1 Sources of uncertainty in an argument

Uncertainty/Confidence in an argument is measured through two parameters:

- Confidence in premises, or *trustworthiness*: where the truth and the falsity carried by a premise, assimilated to the *(sub-)goal* component of Goal Structuring Notation (GSN) formalism, is questioned. For instance, one may doubt some test results for not respecting the experimental protocol.
- Confidence in the inference or support relation between the premise(s) and the conclusion. It is also named *appropriateness*. It questions whether the validation of premise(s) leads to the validation of the conclusion or vice-versa. The support relation is assimilated to the *strategy* component in GSN formalism.

Figure 3 shows an example of these two parameters on an argument consisting in a goal GOA1 supported by two sub-goals GOA2 and GOA3. The strategy STR gives the rationale behind this decomposition into sub-goals.



- $trust_i \equiv (Bel_i, Disb_i, Uncer_i), i = \{GOA1, GOA2, GOA3\}$
- $appr \equiv \left(Bel_{(GOA2,GOA3)\rightarrow GOA1}, Uncer_{(GOA2,GOA3)\rightarrow GOA1}\right)$

Figure 3. Sources of confidence/uncertainty in AC - Example

To assign confidence/uncertainty measures to goals and strategies, they will be formally defined using logical expression. The other components existing in the GSN formalism, in addition to the two mentioned previously (e.g., contexts, solutions, justifications or defeated elements), are not formally defined in the uncertainty assessment procedure presented in this document. However, they are taken into consideration during the assessment of goals and strategies by experts.

B.2.2 Argument type definition

In the literature, authors (e.g., [Cyra & Gorski, 2011; Ayoub, 2013; Wang, 2019]) identify different types of arguments, each expressing the relation between premises in their support of a conclusion. Specifying these relations is essential to define the uncertainty/confidence propagation schema.

These arguments can be grouped into four types [Id Messaoud, 2022a]:

- **Simple argument:** describes the simplest pattern of an argument. I.e., a conclusion supported by a single premise. When this premise is true, so is the conclusion.
- **Conjunctive argument:** in which all premises (i.e., sub-goals) are needed to be valid to support the conclusion (i.e. top goal).
- **Disjunctive argument:** in which the validation of one premise is enough to support the whole conclusion.
- **Hybrid argument:** in which each valid premise supports the conclusion to some extent and their conjunction does it to a larger one.

Each argument is formally defined with logical expressions called *rules*. We have defined four types of rules:

- **Direct rules:** to propagate belief to the conclusion, when premises are true (i.e., valid).
- **Reverse rules:** to propagate disbelief to the conclusion, when premises are false (i.e., invalid). The reverse rule is the contraposed expression of a direct one.
- **Elementary rules:** express the disjunctive relation between premises supporting a conclusion. A disjunctive rule can be expressed with a conjunction of elementary rules. An elementary rule is a faithful formal translation of the informal definition of a disjunctive argument (validation of one premise leads to the validation of the conclusion).
- **Conjunctive rules:** express the conjunctive relation between premises supporting a conclusion.

The inference (support relation) between a premise(s) (p_i) and its conclusion (C) is expressed with logical implication ($\Rightarrow$). Notice that in GSN formalism the direction of the "is supported by" arrow is the opposite of the inference (from premises to the conclusion).

In the case of a conclusion (C) supported by two premises (p_1) and (p_2). A conjunctive argument type is formally defined by three rules:

- A direct conjunctive rule $[p_1 \wedge p_2] \Longrightarrow C$, all pieces of evidence are needed to support the goal (C) and
- Two elementary reverse rules $\neg p_1 \Longrightarrow \neg C$ and $\neg p_2 \Longrightarrow \neg C$, the falsity of one premise leads to the rejection of the goal (C).

Swapping the previous conjunctive and elementary rules with a negation (applied to premises (p_i) and the conclusion (C)), and we get the rules of the disjunctive argument type (i.e., two direct elementary rules and one reverse conjunctive).

Keep in mind that a disjunctive rule is equivalent to a conjunction of elementary rules: $[p_1 \vee p_2] \Rightarrow C \equiv [(p_1 \Rightarrow C) \wedge (p_2 \Rightarrow C)]$.

All six rules are used to formally define the hybrid argument type (two direct and reverse conjunctive rules, and four elementary direct and reverse rules). For n premises, we get $(2n + 2)$ rules.

The hybrid argument is a generalization of the conjunctive and disjunctive argument types, which correspond to extreme cases of the hybrid argument. The simple argument is also a particular case of the hybrid argument, where conjunctive rules are not defined.

In the literature, one can find other definitions of argument types (e.g., [Cyra & Gorski, 2011; Ayoub, 2013; Wang 2019]). However, they can be viewed as specific cases of the hybrid argument.

Example of a hybrid argument type:

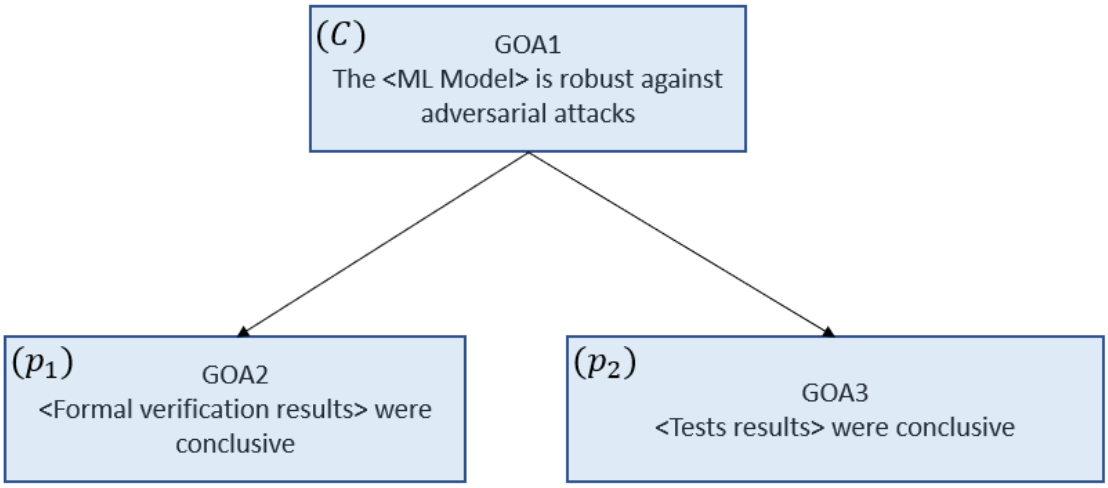


Figure 4. hybrid argument type example

In the example of figure 4, test results support the conclusion to some degree. They are necessary to prove the safety but not sufficient. Since evidence based on formal verification was also provided, which allows us to identify some unsafe states that the system will never reach, experts usually conduct limited tests (which are limited by issues such as cost, feasibility, etc.). On the other hand, tests can cover issues that formal verification might not capture. Hence, we deduce the following rules:

- $[p_1 \wedge p_2] \Longrightarrow C$
- $\neg p_1 \Longrightarrow \neg C$ and $\neg p_2 \Longrightarrow \neg C$

Conjunctive component

- $p_1 \Longrightarrow C$ and $p_2 \Longrightarrow C$
- $[\neg p_1 \wedge \neg p_2] \Longrightarrow \neg C$

Disjunctive component

Hybrid argument rules

B.3 Recent results on uncertainty and confidence assessment in AC

In this section, we present both quantitative and qualitative confidence assessment procedures, which include quantification of confidence/uncertainty in an argument in addition to the elicitation and propagation models of these metrics. Those procedures are the main results of Yassir Id Messaoud's PhD [Id Messaoud, 2022a].

B.3.1 Quantitative approach

- Uncertainty quantification

To quantify confidence and uncertainty in premises (p_i), a mass function m_p^i is associated to each premise which assigns a mass to p_i, its negation $\neg p_i$ and the tautology $(\top)$, summing to 1. I.e., $m_p^i(p_i) + m_p^i(\neg p_i) + m_p^i(\top) = 1$.

Then, a mass function is associated to each rule (r_i), such as: $m_r(r_i) = \alpha$ and $m_r(\top) = 1 - \alpha$, to quantify the confidence on it:

- $m_{\Rightarrow}$ and $m_{\Leftarrow}$: mass functions resp. of direct and reverse conjunctive rules.
- $m_{\Rightarrow}^i$ and $m_{\Leftarrow}^i$: mass functions resp. of direct and reverse elementary rules.

Remark: It is important to differentiate the logical decomposition of a goal into sub-goals and the quantification of confidence/uncertainty (i.e., mass assignment). From a pure logical perspective, the decomposition describes how premises support the conclusion, which can be expressed either with a conjunction or a disjunction. A case that combines these two concepts has no interest. Indeed, the conjunction of a conjunction and disjunction is a disjunction. At this stage, we are not yet considering the hybrid argument. Note that: $[([p_1 \wedge p_2] \Rightarrow C) \wedge ([p_1 \vee p_2] \Rightarrow C)] \equiv ([p_1 \vee p_2] \Rightarrow C)$.

However, when mass functions are assigned to rules to quantify and propagate confidence/uncertainty in an AC, several works have defined arguments that mix between conjunction and disjunction, e.g. [Cyra & Gorski, 2011; Wang et al., 2019; Id Messaoud, 2022a].

Uncertainty elicitation

To propagate confidence and uncertainty measures in an argument, an elicitation approach is defined to (1) collect confidence/uncertainty measures about premises and rules from experts and (2) transforms them into usable format (i.e., degrees of belief, disbelief and uncertainty).

To give her/his opinion, an expert provides two pieces of information. A decision to either accept or reject a premise or a conclusion (s)he assesses (noted Dec) and the degree of confidence (s)he has in it (noted $Conf$). Each piece of information (Dec or $Conf$) is related to a qualitative scale from which the expert takes her/his values. In the same way each qualitative item is mapped to a numerical value in the unit interval [0,1].

Note that numerical decision Dec can be seen as a probability measure of the acceptance of a goal. It is formally defined as follows: $c(A) = P = \frac{1+Bel(A)-Disb(A)}{2}$. The probability of rejection is $Dec(\neg A) = 1 - Dec(A)$ (the universe of all possibilities consists of two elements only, i.e., {True, False}).

The degree of confidence reflects the amount of information the expert possesses to accept (evidence for) or reject (evidence against) a conclusion. Hence, its formal definition: $Conf = Bel + Disb$.

The decision value Dec represents also the center of the interval of possible values for the probability $P \in [Bel, Pl]$ (remember that $Pl = 1 - Dis$).

Example: Let us consider the following 3-tuple: $Bel = 0.5, Disb = 0.2$ and $Uncer = 0.3$ ($Pl = Bel + Disb = 0.8$). Figure 5 represents the positioning of the uncertainty range in the unit interval $[0,1]$, where the belief and plausibility degrees represent respectively its lower and upper boundaries.

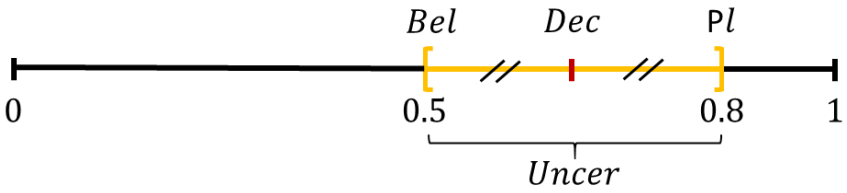


Figure 5. Uncertainty representation in the unit interval [0,1]

Transformation of (decision, confidence) pairs to (belief, disbelief, uncertainty) 3-tuples, using formal definition of Dec and $Conf$, must respect a constraint insuring that:

- a clear-cut decision (acceptable and rejectable) can be taken$Conf = 1$).
- $Bel = Disb = 0$ when the assessor has no confidence ($Conf = 0$) whatever the decision taken.

This constraint, known as Josang constraint, constrains the range of acceptable decision values for a given confidence level. Formally:

$$\frac{1 - \text{Conf}}{2} \leq \text{Dec} \leq \frac{1 + \text{Conf}}{2}$$

Uncertainty propagation

The propagation model combines all the mass (BPA) functions on the rules and premises using a conjunctive combination rule. Thus, we get the following formula for belief propagation.

$$Bel_C(C) = Bel_{\Rightarrow}([\wedge_{i=1}^{n} p_i] \Longrightarrow C) \cdot \prod_{i=1}^{n} \{Bel_p^i(p_i) \cdot [1 - Bel_{\Rightarrow}^i(p_i \Rightarrow C)]\}$$

$$+ \{1 - \prod_{i=1}^{n} [1 - Bel_p^i(p_i) \cdot Bel_{\Rightarrow}^i(p_i \Rightarrow C)]\} - m_C^{(n)}(\perp)$$

With: $m^{(n)}(\perp)$ represent the degree of conflict in an argument of (n) premises.

$$m_C^{(n)}(\perp) = Bel_C^{(n-1)}(C) \cdot m_n(\neg p_n \wedge \neg C) + Disbl_C^{(n-1)}(C) \cdot m_n(p_n \wedge C) + m_C^{(n-1)}(\perp)$$

Where:

- $m_C^{(1)}(\perp) = 0$
- $Bel_C^{(n-1)}(C) = \{1 - \prod_{i=1}^{n-1}[1 - Bel_p^i(p_i) \cdot Bel_{\Rightarrow}^i(p_i \Rightarrow C)]\} - m_C^{(n-1)}(\perp)$
- $Disb_C^{(n-1)}(C) = \{1 - \prod_{i=1}^{n-1}[1 - Disb_p^i(p_i) \cdot Bel_{\Leftarrow}^i(\neg p_i \Rightarrow \neg C)]\} - m_C^{(n-1)}(\perp)$
- $m_i(p_i \wedge C) = Bel_p^i(p_i) \cdot Bel_{\Rightarrow}^i(p_i \Rightarrow C)$
- $m_i(\neg p_i \wedge \neg C) = Disb_p^i(p_i) \cdot Bel_{\Leftarrow}^i(\neg p_i \Rightarrow \neg C)$

Note that:

- $Disb_C(C) = Bel_C(\neg C)$, i.e., to get disbelief degree replace each premise p_i with its negation $\neg p_i$ and direct rules with reverse ones.
- In the case of a conjunctive argument the beliefs in all direct elementary rules are equal to zero. On the other hand, the belief in the conjunctive rule is set to zero when it does not bring an additional

confidence to the conclusion. I.e., when $Bel_{\Rightarrow}([\wedge_{i=1}^{n} p_i] \Longrightarrow C) = max(Bel_{\Rightarrow}^{i}(p_i \Rightarrow C))$. This is the case of a disjunctive argument type.

A conjunctive argument propagates confidence of the premise with least strength (minimal belief and maximal disbelief) with an attenuation effect (or an amplification towards the rejection side). On the contrary, a disjunctive argument propagates confidence of the premise with greatest strength (maximal belief and minimal disbelief) with an amplification to the acceptance side. The hybrid argument makes a trade-off between the conjunctive and disjunctive rules according to their weights.

This amplification effect (mainly noticed in extreme situations, namely with the conjunctive and disjunctive argument types), whether towards acceptance or rejection (attenuation), is due to the use of multiplicative rules of combination which are not idempotent and assume independence of sources. For instance, in the case of a conjunctive argument, for two premises (P_1) and (P_2) with 0.9 belief value for each supporting a conclusion (C). Assuming maximal belief in their rules, the belief in the conclusion will be equal to $Bel(C) = Bel_p^1(p_1) \times Bel_p^2(p_2) = 0.9 \times 0.9 = 0.81$. In practice, it is difficult to verify the independence assumption. Few approaches of idempotent combination rules exist in the literature (e.g. [Denœux, 2006; Destercke, 2011]). But, with the exception of the minimum, they are difficult to use and have quite a few drawbacks. Note that using the idempotent combination rule, i.e. the minimum, instead of the multiplicative combination rule would lead to Bel(C) = 0.9, which is more expectable.

Conflict issue

The conflict degree measures the inconsistency between **premises** in an argument when they have opposite values (e.g., maximal belief in one and minimal disbelief in another). Note that inconsistency between rules is not considered (i.e., questioning the failure of the support relation provided by premise when they hold or vise-versa). We assume that the choice of methods aiming to support a goal follows well-established and known procedures. This assumption is made to simplify the construction of the propagation model. Thus, contradiction between **rules** will not be measured using this metric (i.e., $m_C^{(n)}(\perp)$).

Note that the conflict degree has always a null value in the case of the conjunctive argument which propagates the assessment of the premise with the least confidence values (i.e., maximal disbelief and minimal belief) and the disjunctive argument which propagates the assessment of the premise with the most confidence values (i.e., maximal belief and minimal disbelief).

B.3.2 Purely qualitative approach

Unlike the quantitative approach based on probabilities measures (BPA), the qualitative approach is based on qualitative possibility measures (BΠA). It also replaces resp. *product* and *sum* operations with minimum and *maximum* operations to propagate confidence.
The propagation model is calculated by combining all the possibility (BΠA) functions assigned to rules and premises:

$$\Upsilon_C(C) = max\{min\left[min_{i=1}^{n}\Upsilon_p^i(p_i), \Upsilon_{\Rightarrow}([\wedge_{i=1}^{n} p_i] \Rightarrow C)\right], max_{i=1}^{n}(min\left[\Upsilon_p^i(p_i), \Upsilon_{\Rightarrow}^i(p_i \Rightarrow C)\right])\}$$

Like in the quantitative model, to get the disbelief formula for the conclusion $\Upsilon_C(\neg C)$, one needs to replace premises with their negation and direct rules with reverse ones.

The propagation model takes its values from two scales:

(1) A qualitative bipolar scale for decision with $2n+1$ items: $D = \{0_D = d_{-n}, d_{n-1}, \ldots, d_0 = e, d_1, \ldots, d_n = 1_D\}$. The bottom value (0_D) expresses rejection and the top value (1_D) expresses acceptance. Where (e) expresses a neutral position between acceptance and rejection.

(2) A qualitative unipolar scale for confidence with *n*+1 items: $C = \{0_C = c_0, c_1, \ldots, c_n = 1_C\}$. The bottom value (0_C) expresses no confidence driven by no information and the top value (1_C) full confidence driven by full information.

Values are projected on a third scale L (for linguistic qualifiers), with the same length as the one of confidence $L = \{0_L = l_0, \lambda_1, \ldots, l_n = 1_L\}$, that represents the belief and disbelief dimensions. To move from $D \times C \longrightarrow L \times L$ the following translation formulas are defined:

- $If\ Dec(A) < e, \Upsilon(A) = min[v_C(Dec(\neg A)), Conf(A)]\ and\ \Upsilon(\neg A) = Conf(A)$
- $If\ Dec(A) > e, \Upsilon(A) = Conf(A)\ and\ \Upsilon(\neg A) = min[(v_C(Dec(A)), Conf(A)]$
- $If\ Dec(A) = Dec(\neg A) = e, \Upsilon(A) = \Upsilon(\neg A) = Conf(A)$

Where: v_C is an order-reversing function such that $v_C(c_i) = c_{n-i}$, and $Dec(\neg d_i) = Dec(d_{-i})$.

For instance, applying these translation formulas for the case of n=3, we get for each (decision, confidence) pair the appropriate qualitative belief and disbelief values in Table 1:

Table 1. From $D \times C$ to $(\Upsilon(A), \Upsilon(\neg A)) \in L \times L$

Dec. / Conf.	0_D	d_{-2}	d_{-1}	d_0	d_1	d_2	1_D
0_C	$(0_L, 0_L)$	$(0_L, 0_L)$	$(0_L, 0_L)$	$(0_L, 0_L)$	$(0_L, 0_L)$	$(0_L, 0_L)$	$(0_L, 0_L)$
c_1	$(0_L, \lambda_1)$	(λ_1, λ_1)	(λ_1, λ_1)	(λ_1, λ_1)	(λ_1, λ_1)	(λ_1, λ_1)	$(\lambda_1, 0_L)$
c_2	$(0_L, \lambda_2)$	(λ_1, λ_2)	(λ_2, λ_2)	(λ_2, λ_2)	(λ_2, λ_2)	(λ_2, λ_1)	$(\lambda_2, 0_L)$
1_C	$(0_L, 1_L)$	$(\lambda_1, 1_L)$	$(\lambda_2, 1_L)$	$(1_L, 1_L)$	$(1_L, \lambda_2)$	$(1_L, \lambda_1)$	$(1_L, 0_L)$

Notice that like in the quantitative elicitation formulas, when no information is available (i.e., no confidence) no matter the decision value, the pair (belief, disbelief) gets a null value $(0_L, 0_L)$. In the case of full information, we can also notice the three extreme situations: Full belief $(1_L, 0_L)$, full disbelief $(0_L, 1_L)$ and full conflict $(1_L, 1_L)$ which are represented in quantitative approach by the pair $(Bel = 0.5, Disb = 0.5)$.

C. Methodology

In a first step, the methodology proposed here for uncertainty assessment chooses either the quantitative approach described in section B.3.1 or the qualitative approach described in section B.3.2 or both. Then, in the second step, the methodology proposes a data capture method that elicits information from experts and is consistent with the choice made in the first step. Finally, the analysis and the propagation of the uncertainty to the top-goal of an AC is performed using the formulas applicable to the chosen approach(s).

This approach is illustrated using the “robustness” assurance case.

C.1 Choosing an approach

C.1.1 Derivation of requirements

Requirements are proposed for uncertainty modeling and assessment:

- (R1) The assessment shall allow comparing demonstration strategies with respect to their contribution to confidence.
- (R2) The result of the assessment of an Assurance Case shall not be driven by its dimension.
- (R3) The assessment shall allow comparing demonstration strategies with respect to the sensitivity of their contribution to confidence.
- (R4) The method for uncertainty elicitation and propagation shall be grounded.

Note that those requirements are minimal requirements, i.e. other requirements could be added.

C.1.2 Choice of a formalism complying with requirements

To choose between the quantitative and qualitative uncertainty assessment method, both approaches are analyzed with respect to the previous requirements. In this document, a **node** of an AC is defined as a goal and either its sub-goals or its solutions.

C.1.2.1 Usefulness

Usefulness is estimated by assessing the help brought by uncertainty assessment for focusing validation effort and identifying weaknesses of the argumentation.

Uncertainty assessment is useful for focusing validation effort on most sensitive parts of the AC because it is performed at each goal. So, for each goal weakness and contradictions between proof elements can be highlighted. For nodes corresponding to conjunctions, a procedure to focus on the most sensitive element, i.e., the one with lowest belief is derived. If this element corresponds to a Solution, one must consider means for increasing its belief. For example, if the solution is a set of test results, belief can be increased by making more tests.

Uncertainty assessment is also useful for identifying weaknesses of the argumentation and applying uncertainty reduction techniques. The proposed procedure is like the one for focusing validation. A strategy associated to a node, whose uncertainty of its goal is sensitive but whose uncertainties of its sub goals are not so, is not sufficiently convincing. Indeed, in that case the sensitivity of the uncertainty of the goal is due to inappropriate rules. Then, if available, an alternative strategy can be considered.

C.1.2.2 Dimension

Dependency to dimension is analyzed by applying each formalism to a perfect conjunctive argument whose number of premises varies. The limits of the belief and disbelief of the conclusion when the number of premises tends to infinite indicates whether the assessment of the AC is dimension dependent.

For the quantitative approach, the result of the analysis of this requirement on a large conjunctive argument case shows that, as the number of solutions increases, the general trend is the rejection of the property corresponding to the root goal.

For the qualitative approach, as shown in Figure 6, the belief of the root goal cannot be lower than the belief of the sub-goal directly linked to a solution with the lowest belief. Moreover, the disbelief of the root goal cannot be larger than the disbelief of the solution with the largest disbelief. With this approach, the uncertainty of the root goal is bounded.

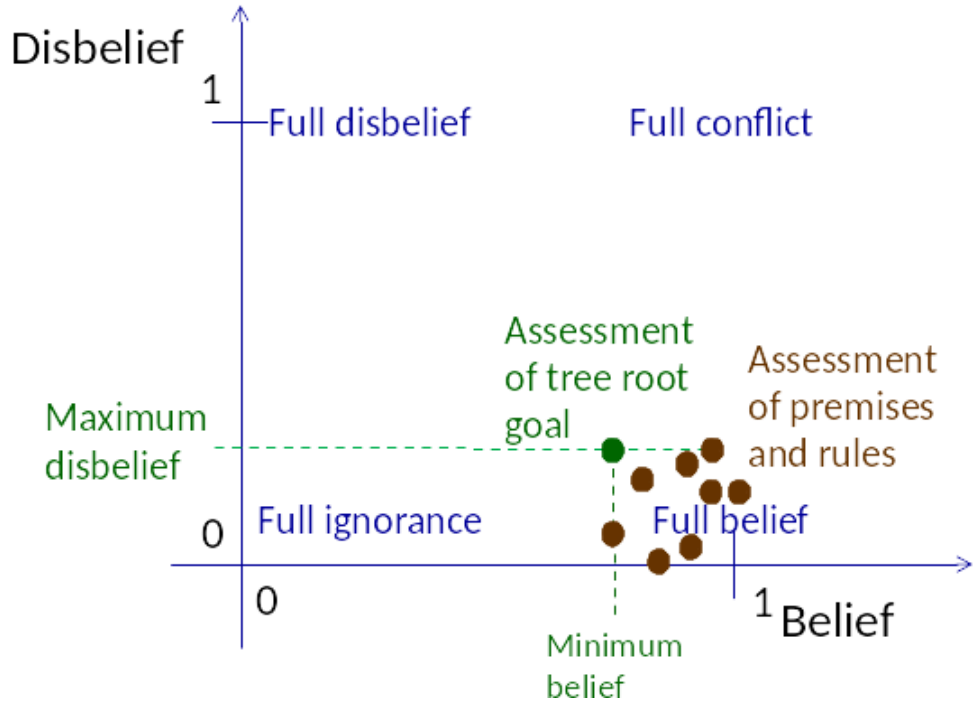


Figure 6. Illustration of the uncertainty propagation for the qualitative approach applied to a conjunctive argument

C.1.2.3 Sensitivity

Sensitivity is assessed by computing the partial derivatives and checking for their possible null value. Indeed, a null partial derivative indicates at the first order the absence of sensitivity.

Changing a strategy changes the assessment of the goal supported by this strategy. This goal supports another goal. Thus, while changing a strategy, it changes the assessment of one of the premises of this other goal.

For the numeric approach, partial derivatives of the belief and disbelief of the other goal with respect to belief and disbelief of a premise are highlighted. Thus, there is a sensitivity to each premise. Moreover, for assessing the sensitivity of arguments to disbelief in premises, a parameter ε is defined and belief and disbelief in premises are set respectively to 1 − ε and ε. Results indicate that for the conjunctive argument

belief and disbelief of conclusion are highly sensitive to ε, for the disjunctive argument belief and disbelief of conclusion are not sensitive to ε and that for the hybrid argument belief of conclusion is sensitive to ε while disbelief of conclusion is not sensitive to ε. Nevertheless, for this argument uncertainty is sensitive to ε. Additional sensitivity analysis is performed by varying the mass on individual direct rule. It indicates that the decrease of this mass reduces uncertainty and increases disbelief in conclusion. Finally, it is observed that for those cases the uncertainty is equal to the degree of conflict.

For the qualitative approach, sensitivity of goal belief to belief of premise argmin and sensitivity of goal disbelief to disbelief of premise argmax are highlighted. However, those sensitivities are valid only when argmin respectively argmax are single premise. Finally, there is no sensibility to other premises.

C.1.2.4 Methodological choices

Finally, the grounding of methodological choices is analyzed by checking mathematical properties of the T-norms associated to the numeric and the qualitative approaches and assessing the possibility of consensus between experts.

The T-norm used in the numeric approach can only be applied to numbers and is grounded on:

- Assimilating the uncertainty measure to frequencies
- Representativeness of frequencies.
- Independence of events.

The T-norm used in the qualitative approach can be applied on numbers as well as on ordered linguistic qualifiers and is the unique T-norm complying with idempotence, absorption and distributivity.

Concerning the assessment of elementary elements, the consensus on the association of a number with a linguistic qualifier is difficult. The numeric approach highlights slight differences between belief degrees. However, it is unlikely that two experts provide the same value. The scale used by the qualitative approach is associated with linguistic qualifiers, there is consensus on their order, and it is likely that two experts associate the same qualifier with the same element. However, there are gaps between the degrees of the scale and results in an extreme case highlight the negative effect of a limited number of linguistic qualifiers on sensitivity. Indeed, improving the AC implies substituting several elements in a single step.

C.1.3 Conclusion

Table 2 presents a synthesis of the compliance of uncertainty assessment methods with requirements. It indicates that to comply with all requirements it is needed **to work with both a scale and numbers and use the numeric and qualitative methods together**.

	Numeric	Qualitative
Usefulness	+	+
Result not dimension driven	-	++
Sensitivity	++	-
Reasoned methodological choices	+	+

Table 2: Compliance of uncertainty modeling with requirements

C.2 Data capture method

C.2.1 Choice of AC elements to be assessed

Another important result of the work is the definition of a methodology for elicitation of uncertainty associated with rules and goals directly linked to solutions. Following this methodology, the full tree is presented without Strategies to experts, i.e., sub-goals are directly connected to goals by a “Is supported by” relation. Experts are not asked to provide uncertainty about Strategies that are considered by AC developers as logical decomposition of their goal without uncertainty on rules or pure rewording without semantic change. Moreover, for some goals the uncertainty on rules is assessed for only one sub-goal among all sub-goals present in the AC. Indeed, some goals admit only binary assessments (full belief or full disbelief) and since the falsity of such goals may jeopardize the whole argument, they became requirements that the AC developers need to respect. For instance, in the pattern presented in figure 7, the applicability of a method (GOA2) is not questioned, we assume that the conditions of applicability of each method involved in the AC are verified and appropriate to the context of the argument structure. The same goes for GOA4 which states the application of the chosen method during the ML training stage. In this case, affecting maximal belief to these goals (GOA2 and GOA4) or simply excluding them from the confidence/uncertainty propagation scheme have the same results since the node, cf. Definition in Section C.1.2, regrouping them represents a conjunction (i.e., logical AND relation). However, it is safer to exclude such goals from the evaluation since affecting maximal belief can improve the confidence in the top-goal when the node is a disjunction (i.e., logical OR).

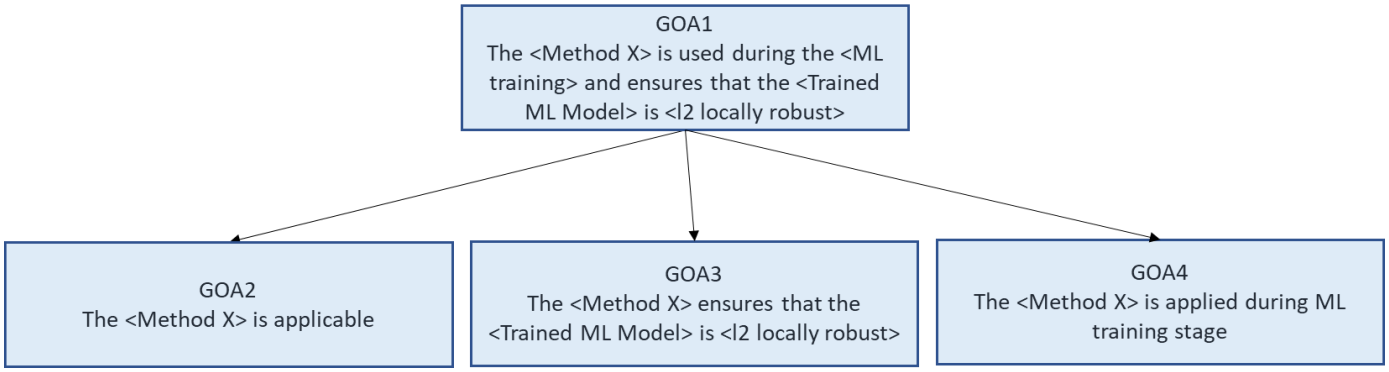


Figure 7. Pattern of local robustness demonstration methods

In addition, goals that argue support between goals with which they share the same node and the top goal are also excluded from the evaluation to avoid redundancy with the evaluation of rules that provide the same information. Figure 8 shows an example of such case. In the latter, only goal GOA2 will be assessed. The inference between the use of formal verification and robustness against adversarial attacks will be evaluated through the evaluation of the rules: $\mathrm{GOA2} \Rightarrow \mathrm{GOA1}$ and $\neg\mathrm{GOA2} \Rightarrow \neg\mathrm{GOA1}$.

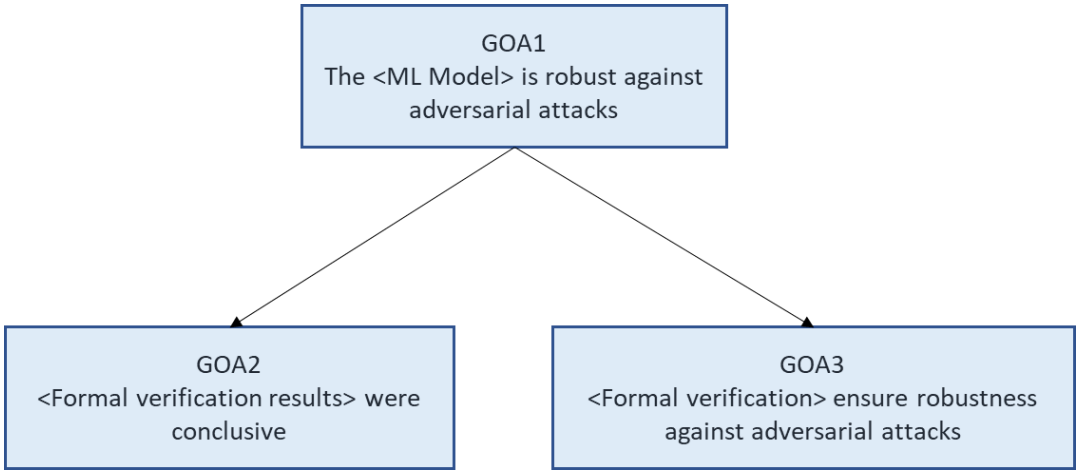


Figure 8. An example of an argument with a premise arguing the inference between a top-goal (GOA1) with its sub-goal (GOA2)

On the other hand, some goals may also be excluded from this procedure if the expert considers that he does not have sufficient knowledge to evaluate them. The robustness assurance case we developed for example (a non-instantiated AC), groups a large set of methods that may even be mutually incompatible. The aim, as mentioned before, is to enable a future user to select and apply the methods that give him/her the most confidence. And so, it's only natural that an expert might not be able to evaluate all methods proposed to him/her. For the same reasons, some solutions associated with the same goal are also not considered.

C.2.2 Elicitation of uncertainty from experts

The uncertainty assessment approach is based on expert judgments about the argument structures to provide belief, disbelief and uncertainty degrees to propagate. These judgments are collected in terms of symbolic (decision, confidence) pairs in response to multiple questions grouped in a questionnaire. For any judgement confidence is assessed using a scale of four items {Very high, high, low and very low} confidence. Each selected decision or confidence element is associated to a numerical interval which can be used to refine the assessment. Scrollbars, which drive both scales (symbolic and numerical), are provided to select the desired values.

As presented in section B2.1, two types of confidence/uncertainty need to be quantified in an argument structure: appropriateness and trustworthiness.

C.2.2.1 Appropriateness (i.e., confidence/uncertainty assessment of rules)

Experts filled a questionnaire with a form for each hidden Strategy. Figure 8 presents the form for a goal supported by two sub goals. In those forms, the number of questions per hidden Strategy is equal to the number of rules, i.e., two plus twice the number of child Goals.

In a first step, belief degree in rules which will be collected from goal/sub-goal(s) nodes, cf. definition in Section C.1.2) (e.g., Figure 8). For each node $2\mathrm{n}+2$ ($n \geq 2$, is the number of sub-goals): $2n$ questions for direct and reverse elementary rules, plus 2 questions for the conjunctive rules (direct and reverse). When one sub-goal is provided to support the goal, conjunctive rules are not defined. The expert will be asked only two questions in this case.

For direct rules the provided decision is the acceptance of the parent goal. For reverse rules, it is the rejection of the parent goal. In both cases the strength of the decision is associated to a scale of four items concerning either acceptance or rejection of a conclusion {Strong, Moderate, Weak and No decision}.

Figure 9 represents an extract of the questionnaire used to assess uncertainty in the robustness AC. The node here has two sub-goals, thus six questions (i.e., $2 \times (n = 2) + 2 = 6$) were asked to the experts to quantify confidence/uncertainty in the inference rules.

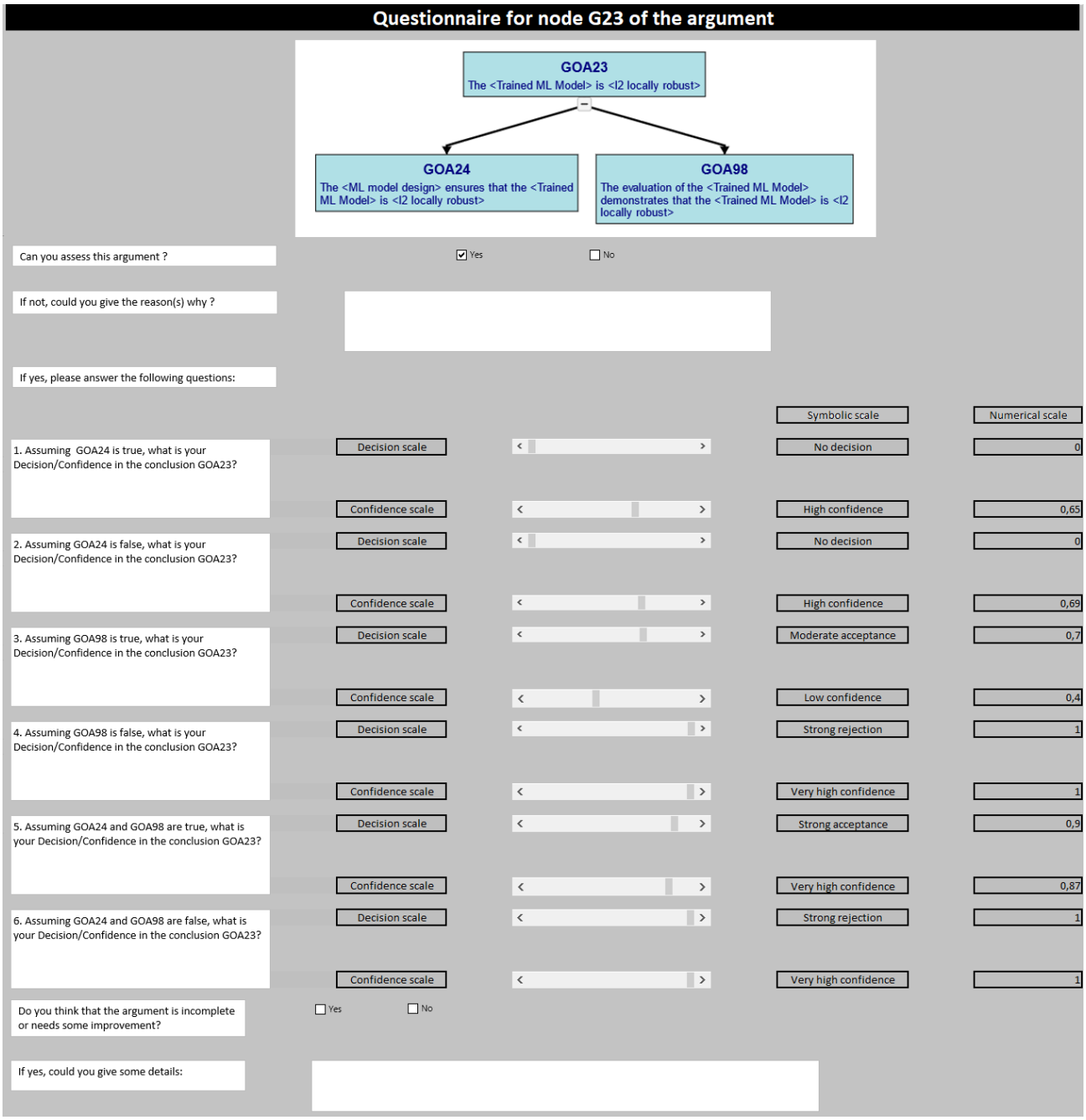


Figure 9. Questionnaire about goal/sub-goals node

C.2.2.2 Trustworthiness (i.e., confidence/uncertainty assessment of premises)

In a second step, belief and disbelief degrees in premises will be collected from the assessment of goal/solution(s) nodes (e.g., Figure 10). For each goal/solution(s) node the expert will be asked 1 question, for a total of n questions for n premises.

For goals linked to a solution decision is associated to a scale of seven items {Strong acceptance, Moderate acceptance, Weak acceptance, No decision, Weak rejection, Moderate rejection and Strong rejection}.

Figure 10 represents another extract of the questionnaire used to assess uncertainty in the robustness AC. In this case the expert will be asked to answer a single question for this node. His/her answer needs to take into consideration all components linked to the goal (e.g., contexts N°38, N°39 and solution N°37 presented in this example).

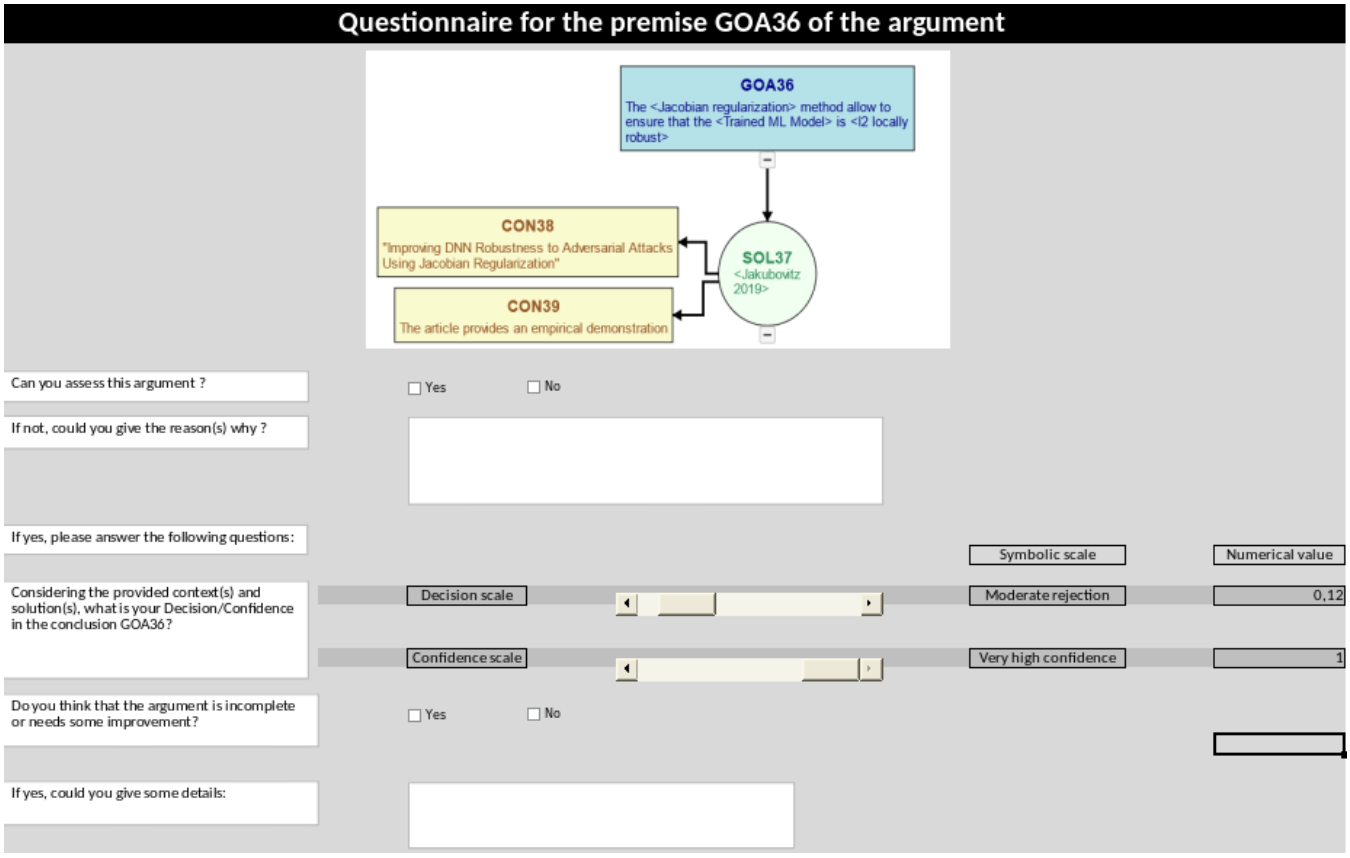


Figure 10. Questionnaire about goal/solution node

For each node (i.e. goal/sub-goal or goal/solution one), the expert will first be asked if (s)he can assess the argument. In addition, s/he is given the possibility to justify his/her answer using free text. If the expert is unable to provide answers for a node, or strongly disagrees with the decomposition or with the method proposed by the developer of the assurance case, then there is no interest in continuing evaluating the

node. At the end the expert will also be asked if (s)he thinks that the argument is complete or needs some improvement. A space dedicated to justification is also provided (see figures 9 and 10).

C.2.3 Debriefing

After a first analysis of the questionnaire filled in by the expert, a debriefing session is organized, and specific questions are asked to the expert. The topics addressed by those questions have the objective to clarify apparent paradoxes in the answers provided by the expert.

C.3 Data analysis

C.3.1 Scaling the data

The values provided by experts are transformed into beliefs and disbeliefs, using the elicitation models introduced in Section B.3.1 for the quantitative approach and in Section B.3.2 for the qualitative approach. I.e., the quantitative approach considers the numerical values provided by scrollbars while the qualitative approach uses the semantic qualifiers.

Unlike premises (goals related to solution) which can be directly transformed to masses on belief, disbelief and the tautology (to represent uncertainty), the assessments of rules (related to goal/sub-goal(s) nodes) require further adjustment before converting the collected values to masses on belief and the tautology. Indeed, for decisions related to rules, the scrollbar controls a numerical value between 0 and 1, whatever the decision is acceptance or rejection while in the theory presented in Section B.3.1 a full rejection corresponds to 0, no decision corresponds to ½ and a full acceptance corresponds to 1. Thus, each selected decision (Dec^*) is transformed following these formulas: $Dec = \frac{Dec^*-1}{2}$ for a negative decision and $Dec = \frac{Dec^*+1}{2}$ for a positive decision. Then, these rescaled decisions with their corresponding values of confidence are transformed into pairs of belief and uncertainty. The resulting disbelief values are systematically set to zero, since the latter is not defined for rules (see section B3.1).

C.3.2 Respecting Josang constraint

C.3.2.1 Quantitative approach

All expert judgments need to respect Josang constraints (introduced in section B.3.1), otherwise the collected (decision, confidence) pairs would give negative belief or disbelief values or lead to a sum of belief and disbelief larger than one which would make no sense. In addition, these metrics are probabilities, so they must belong to [0, 1]. Those constraints ensure that each decision is coherent with the confidence value selected by the expert. For instance, choosing "*strong acceptance"* or "strong *rejection"* with minimal confidence (equal to zero) has the same strength as choosing "*No decision"* with the same level of confidence (i.e., $Bel = Disb = 0$ and $Uncer = 1$ for the three cases). Those constraints can be graphically represented by a triangle in the evaluation matrix of Figure 11.

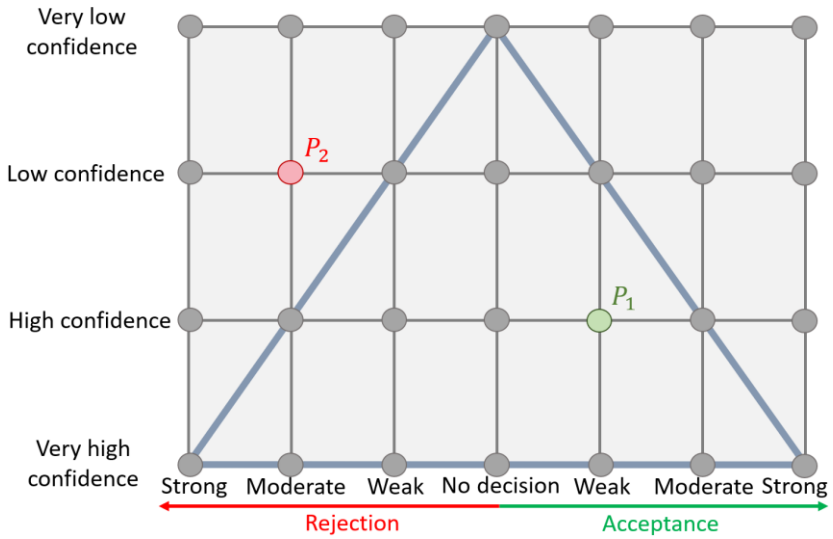


Figure 11. Josang constraints graphical representation (Josang triangle)

Example:

In this example, we present the numerical transformation of decision and confidence pairs into belief, disbelief and uncertainty measures applied for the quantitative assessment approach. Considering two (decision, confidence) pairs (p_1) and (p_2) presented in Figure 11. For each pair, we assign the following numerical values:

- $Dec(P_1) = 0.67$ and $Conf(P_1) = 0.70$
- $Dec(P_2) = 0.17$ and $Conf(P_2) = 0.34$
- Let us calculate the corresponding belief, disbelief and uncertainty degrees:
- $Bel(P_1) = \frac{Conf(P_1)-1}{2} + Dec(P_1) = 0.52, Disb(P_1) = \frac{Conf(P_1)+1}{2} - Dec(P_1) = 0.18$ and $Uncer(P_1) = 1 - Bel(P_1) - Disb(P_1) = 0.30$

The pair (p_1), located inside the triangle in Figure 11, respects Josang constraints. So, we can notice that all values are included in the unit interval [0,1] and their sum is equal to 1.

- $Bel(P_2) = -0.16, Disb(P_2) = 0.50$ and $Uncer(P_2) = 0.66$
- The pair (p_2), located outside the triangle in Figure 11, does not respect Josang constraints. So, we can notice a negative value of belief which makes no sense.

When Josang constraints are violated, we can either notify the expert to adjust his/her judgment if (s)he is available or adjust the decision value according to confidence. Two adjustments are possible: one along the Dec axis and one along the Conf axis. Adjustment along Dec seems more natural since it corresponds to the amount of information the expert uses to justify his/her decision. Thus, Dec is set to $Dec(p) \leftarrow \frac{1-Conf(p)}{2}$, when $Dec(p) < \frac{1-Conf(p)}{2}$. And $Dec(p) \leftarrow \frac{1+Conf(p)}{2}$, when $Dec(p) > \frac{1+Conf(p)}{2}$.

Back to P_2, $Dec(P_2) < \frac{1-Conf(P_2)}{2} = 0.33$, so we set $Dec(P_2) = 0.33$ so $Bel(P_2) = 0, Disb(P_2) = 0.34$ and $Uncer(P_2) = 0.66$; Notice that the projection on the constraint keeps the same level of uncertainty and the inequality $Disb(P_2) > Bel(P_2)$ is maintained.

C.3.2.2 Qualitative approach

In the qualitative approach, Josang constraint is systematically respected since the transformation formulas are designed to respect it [Dubois, 2022]. All possible transformation of decision and confidence pairs into qualitative belief and disbelief are grouped in table 1 of section B.3.2. 0_L represent the lower value of the linguistic scale (L) for decision (i.e., No decision) or confidence (i.e. very low confidence), while 1_L represent the upper value for decision (i.e., Strong acceptance or rejection) or confidence (i.e. very high confidence). This linguistic scale $L = \{0_L, \lambda_1, \lambda_2, 1_L\}$, is defined in section B3.2. Because the result is a belief or a disbelief the linguistic qualifiers of the four elements of the scale are very low, low, high and very high belief or disbelief.

C.3.3 Getting consistent rules

Another important aspect of data analysis is the consistency between rules of arguments that are not simple arguments. Indeed, the belief in the direct conjunctive rule involving all sub goals cannot be lower than the highest of the beliefs of the direct elementary rules involving individual sub goals. Moreover, the belief in the disjunctive reverse rule involving all sub-goals cannot be lower than the highest of the beliefs of the inverse elementary rules involving individual sub-goals. When such inconsistencies are found in the data, it can be a topic for debriefing with the expert and finding a way to solve those inconsistencies.

Assuming that each sub goal can be true or false independently of the truth associated to other sub goals, there are several ways for solving those kinds of inconsistencies:

- Setting the belief of the rule involving all sub goals to the largest belief of the rules involving individual sub goals.
- Setting the beliefs of the rules involving individual sub goals that are higher to the belief of the rule involving all sub goals to that last value.
- Any intermediate solution.

Note also that those inconsistencies may not matter in the case for direct rules where the hidden strategy is a choice where all sub goals cannot be selected at the same time and the case for inverse rules where the hidden strategy is a choice where all sub goals cannot be unselected at the same time.

Finally, in the case belief of rules involved in inconsistencies were corrected for respecting Josang constraint, the way the correction was made may increase or reduce those inconsistencies. For that reason, it is recommended to get a global view of the problem before choosing the adjustments.

C.4 Uncertainty propagation

Results are derived by gathering and analyzing the filled forms and by making uncertainty propagation under different hypotheses for the solution directly linked to artifacts. Three kinds of hypotheses are performed for allowing uncertainty propagation:

- Completing the AC for unassessed goals with sub goals,
- Providing an assessment for unassessed goals directly linked to solutions, and
- Making choices for choice nodes.

Unassessed goals may exist either from the will, or lack of resources, of the uncertainty assessment team, cf. Section C.2.1, or because the expert indicated, in the form of the goal, for the question “Can you assess this argument?” the answer “No”, see Figures 9 and 10. The first case shall be foreseen when designing the questionnaire. For the second case some contingent solution shall be found.

C.4.1 Completing the AC for unassessed goals with sub goals

The simplest way to complete the AC for an unassessed goal that has only one assessed sub goal among several sub-goals is to connect directly the unassessed goal to the assessed sub-goal using a simple argument with no uncertainty.

The simplest way to complete the AC for an unassessed goal that has no assessed sub goals is to make a direct hypothesis about its assessment. The most obvious hypothesis is *Bel* = 1 and *Disb* = 0, i.e. Acceptance of the goal, with a Very High Confidence. At a second order, sensitivity analysis can be performed with respect to those nodes by setting a common parameter ε and varying it as explained in Section C.1.2.3.

C.4.2 Assessing unassessed goals directly linked to solutions

For unassessed goals directly linked to solutions two strategies are possible:

- If the goal is almost directly linked to a multiple-choice node where only a few alternatives can be selected simultaneously, the goal and the corresponding alternative can be dismissed. By “almost”, it is here meant that the goal belongs to a branch that includes the multiple-choice node, and that the absence of the node invalidates the branch up to the multiple-choice node.
- If the goal is on a branch of the AC with no choice, it is possible either to make an optimistic hypothesis for the assessment of the goal or to cut the branch somewhere. The choice between those solutions is use case dependent.

C.4.3 Making choices for choice nodes

Choice for choice nodes shall be made only after uncertainty has been propagated up to all sub goals of the node. Then, a set of alternatives shall be constructed with all compatible sub goals. If sub goals are all incompatible the set of alternatives is composed of singletons, each singleton corresponding to a sub goal. If some compatibilities exist, additional sets with larger cardinal can be considered.

Singletons have already their uncertainty assessment. Uncertainty assessment of sets with larger cardinal can be performed using a disjunctive argument with the values already collected for direct rules of individual sub goals and the maximum of the values already collected for inverse rules for the inverse rule concerning all sub goals. Alternatively, the issue could be addressed with the expert during the debriefing session, see end of Section C.2.2. In that case, specific forms corresponding to specific sets of sub goals shall be produced before debriefing.

When all alternatives are valued, some of them can be dismissed because they present more disbelief than belief. the choice problem is a multi-criterion problem with the following criteria:

- Qualitative belief to be maximized,
- Qualitative disbelief to be minimized,
- Quantitative belief to be maximized, and
- Quantitative disbelief to be minimized.

Many methods from state of the art are available to solve this kind of problem: Pareto front, max Leximin, Promethee, Analytic Hierarchy Process, Electre… May be the best option is to start with a Pareto front and then to apply a max Leximin with the following correspondence for the qualitative scale: 0, 1/3, 2/3, 1.

C.5 Towards multi expert assessment

Conflicts, as meant by DST, cannot be detected at single rule level because for rules mass is only on $\top$ (tautology) and r (the rule). However, variation of mass between experts can be recorded. Moreover, conflicts cannot be detected at node level. Indeed, it is shown that if masses on rules of expert 1 and 2 respect consistency, consistency is respected by masses on rules of the fusion. Finally, conflicts cannot be detected at tree level with an optimistic leaf assignment because the propagation of an optimistic leaf assignment induces for any node of the tree a belief in [0,1] and a null disbelief. Globally, conflicts between experts are not detectable without applying the AC to a use case.

C.6 Application to the robustness of ML model

C.6.1 Presentation of the AC template

The root goal of the AC template for robustness, presented in figure 12, of ML is “<The Trained ML model> is <robust>”, where “<Trained ML model>” is an artifact resulting from the design and building stages of the life cycle and “<robust>” is a property defined in the AC. This is a template AC and not an AC because solutions that should be associated to the goals “The <verification set> is relevant for robustness evaluation”, “The <method for robustness reinforcement> is applicable” and “The <method for robustness reinforcement> is applied during the ML training stage” are not provided, and branches of the tree can be deleted for a specific ML model.

This goal is reformulated and then decomposed into three sub-goals, all referring to the concept of local robustness. Then a decomposition is performed with respect to the norms involved in the local robustness criterion and then with respect to the way robustness can be demonstrated, either “by design” or “by validation”. The tree further develops the branch dedicated to “by design” methods, splitting into sub-goals corresponding to families of methods of this category.

Finally, the goal corresponding to each method is supported by a set of three goals: two that depend on artifacts linked to the trained ML model, and one connected to a solution referencing results published in research articles.

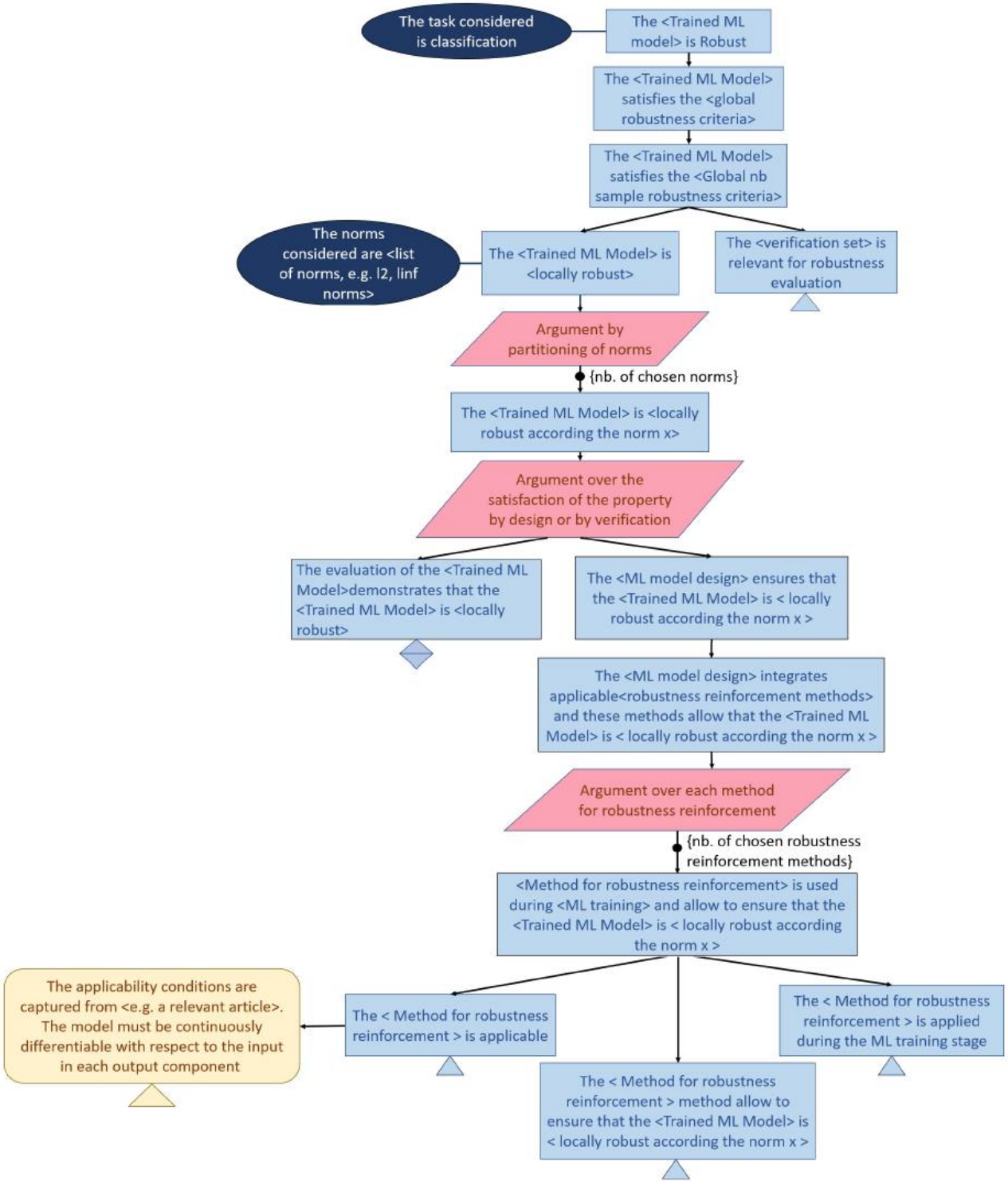


Figure 12. Template of Robustness Assurance Case (compact form)

The diamond head shape connected to a goal means that the corresponding branch needs to be completed and instantiated. While the triangular shape means that the goal needs to be instantiated only. Solutions are intentionally omitted. The extended AC template will be provided with the deliverable.

C.6.2 Assessed goals

Experts filled forms of the type of the ones shown in Figures 9 and 10, for goals connected to articles and for nodes upper in the tree.

Goals supported by Strategies investigated here are presented in Table 1.

	Wording
Goal 15	<The Trained ML model> is <robust>
Goal 17	<The Trained ML model> satisfies the <global robustness criteria>
Goal 18	<The Trained ML model> satisfies the <Global nbsample robustness criteria>
Goal 21	<The Trained ML model> is <locally robust>
Goal 23	<The Trained ML model> is <l2 locally robust>
Goal 24	<The ML model design> ensures that <The Trained ML model> is <l2 locally robust>
Goal 25	<The ML model design> integrates applicable <robustness reinforcement methods> and these methods allows that <The Trained ML model> is <l2 locally robust>
Goal 99	<The Trained ML model> is <linf locally robust>
Goal 100	<The ML model design> ensures that <The Trained ML model> is <linf locally robust>
Goal 101	<The ML model design> integrates applicable <robustness reinforcement methods> and these methods allows that <The Trained ML model> is <linf locally robust>

Table 1: Goals supported by Strategies

Because the AC is not fully developed, some goals remain open, i.e. non-developed:

- Goal 98 whose wording is “The evaluation of the <Trained ML Model> demonstrates that the <Trained ML Model> is <l2 locally robust>”
- Goal 178 whose wording is “The evaluation of the <Trained ML Model> demonstrates that the <Trained ML Model> is <linf locally robust>”.

Goals supported by Solutions investigated here are on the one hand goal 19 whose wording is “The <verification set> is relevant for robustness evaluation” and support is “<Verification set>” and on the other hand a set of goals dealing with robustness methods whose wording is “The <Method> method allows to ensure that the <Trained ML model> is <lx locally robust>” and support is a list of scientific articles. Table 2 presents those goals.

	lx	*Method*
Goal 36	l2	Jacobian regularization [Jacubovitz, 2019]
Goal 48	l2	Lipschitz training [Anil, 2019]
Goal 60	l2	Certified robust training [Wong, 2018a; Huang, 2017; Hein, 2017]
Goal 82	l2	Randomized smoothing [Cohen, 2018; Hong, 2022; Lecuyer, 2019]
Goal 108	linf	Empirical robustness reinforcement method [Ross, 2017; Jacubovitz, 2019; Tramer, 2017; Nayebi, 2017; Gao, 2017]
Goal 132	linf	Lipschitz training [Anil, 2019]

Goal 143	linf	Gowal certified robust training [Gowal, 2019]
Goal 154	linf	Certified robust training [Wong, 2018b; Zangh, 2019]
Goal 168	linf	Random Noising [Hong, 2022; Lecuyer, 2018]

Table 2: Goals supported by scientific articles

C.6.3 Supporting goals

Table 3 indicates the goals supporting assessed goals that are not supported by solutions.

	Sub goals
Goal 15	17
Goal 17	18
Goal 18	19, 21
Goal 21	23, 99
Goal 23	24, 98
Goal 24	25
Goal 25	30, 42, 55, 76
Goal 99	100, 178
Goal 100	101
Goal 101	103, 126, 139, 150, 165

Table 3: Supporting goals for goals supported by Strategies

Goal 19 corresponds to “The <verification set> is relevant for robustness evaluation”. Goals 98 and 178 are respectively “The evaluation of the <Trained ML Model> demonstrates that the <Trained ML Model> is <l2 locally robust>” and “The evaluation of the <Trained ML Model> demonstrates that the <Trained ML Model> is <linf locally robust>”. Finally, Goals 30, 42, 55, 76, 103, 126, 139, 150 and 165 have the *lx* and *Method* of Goals 36, 48, 60, 82, 108, 132, 143, 154 and 168 respectively. Their wording is “<*Method*> is used during <ML Training> and allows to ensure that the <Trained ML Model> is <*lx* locally robust>”.

D. Data collected from the expert

D.1 Expert choice

As presented in this document, this uncertainty/confidence assessment approach relies on experts' judgments to provide measures (i.e., belief and disbelief degrees) which are propagated to the overall goal of the assurance case. The choice of those assessors follows three main criteria. Firstly, the assessor needs to be an expert of the field regarding the part of the argument (i.e. node(s)) s\he assess. According to its complexity, an assurance case can use evidence from different fields. This criterion guarantees the trustworthiness of the judgments provided. Secondly, the assessor, regardless of his/her level of expertise, should not be selected from the development team. This criterion eliminates the bias associated with a subjective assessment that would attempt to defend the argument rather than question it. Finally, the assessor ideally needs to have experience from both: (1) industrial domain to judge the use case-dependent arguments, notably for the instantiated assurance cases (i.e., all required artifacts are supplied), and (2) academia to assess relatively new methods from articles used as evidence. However, since such profiles are not easy to identify, one can be limited to an expert from one of the two domains. For instance, an industrial expert when the assurance case is application-oriented (e.g. an Industrial Use-Case), or an academic expert for a generic Template based on scientific articles. For more details, deliverable **"613C – Evaluation of assurance case"** is also available.

D.2 Filled questionnaire

D.2.1 Unassessed arguments

The expert indicated that he is not able to assess goal 19. The reason provided is that no reference is given to understand what a relevant dataset for robustness evaluation is.

Goals 108, 154, and 168 were also not assessed. For those goals the expert indicated just "NOT DONE".

D.2.2 Additional comments

Table 4 presents the comments provided by the expert and its assessment about the completeness of the argument. Goals with no comment and with an empty or positive assessment of completeness are not shown in the table.

Goal	Complete	Comments
15	yes	Warning: Definition DEF14 of "robust" is restricted. Robustness can also be with respect to distribution shift for instance.
17	no	DEF8: "a subset" is unclear. Is it the verification set? Criteria of representativity of the subset is needed.

		The case where GOA18 is false could be analyzed with other criteria (max or mean) and could lead to a different conclusion?
18	yes	GOA21: I assumed that this goal is for all samples in the verification set while the definition <locally robust> indicates for 1 sample.
21	empty	Note that L2 ball is included in Linf ball.
23	empty	It's a shame that GOA98 is not detailed, because there is a lot to say about the demonstration of l2 local robustness: adversarial attacks for example only give us an upper limit of the robustness radius but do not guarantee this radius (it could be much smaller).
24	no	
25	no	The case where all goals are true seems illusory: we cannot combine all the methods. In the case where they are all false, we can imagine a new method which would emerge and which would make it possible to justify GOA25. I made my decision assuming that these were the only methods available.
100	no	I am not sure that all these methods are applicable for Linf robustness. In particular the Lipschitz even if it should strengthen it
101	empty	The case where all goals are true seems illusory: we cannot combine all the methods. In the case where they are all false, we can imagine a new method which would emerge and which would make it possible to justify GOA101. I made my decision assuming that these were the only methods available.
36	no	It will be the same conclusion for many solutions: this method adds a regularization that can help robustness. But as a regularization it cannot "ensure" that the model is l2 locally robust. A verification must be done, and no guarantee exists with only this method Maybe change "Ensure" to "Enhance" and I will change my decision.
48	no	Training with 1lipschitz constraint is not enough for robustness, a loss for have large logits is needed (even if classical loss tends to provide large logits) => add a specific loss such as "Achieving robustness in classification using optimal transport with hinge regularization" to Ensure that the NN is trained with a robustness goal. But the 1lipschitz constraint provides theoretical robustness radius ==> Acceptance (only if empirical verification is done on a validation set).
60	no	3 papers are proposed. Is it an OR on the methods, or a AND?
132	empty	I'm less confident on Linf robustness. But Anil et al seems to provide the same level of guarantees as for L2 norms. The question of having a loss for enhancing robustness should also be considered.
143	no	Not much mathematical demonstration but based on Formal Methods IBP Regularization gives no guarantee for robustness, but the paper provides a verified robustness using formal methods => better than only adversarial attack.

Table 4: Raw data on completeness of argument and associated comments

Moreover, the expert provided reading notes about some articles provided as solutions. Those reading notes can be found in Annex 1.

D.2.3 Confidence and decision

Table 5 presents the qualitative and quantitative confidence and decision provided by the expert for each considered rule.

Supported goal	Supporting sub-goal(s)	Rule type	Expert Confidence		Expert decision	
			Qualitative	Quantitative	Qualitative	Quantitative
15	17	direct	VH	1.00	S	1.00
15	17	inverse	VH	1.00	S	1.00
17	18	direct	VH	1.00	S	1.00
17	18	inverse	VH	1.00	S	0.83
18	19	direct	VH	1.00	ND	0.01
18	19	inverse	VH	0.90	M	0.74
18	21	direct	VH	0.93	M	0.60
18	21	inverse	VH	0.93	M	0.76
18	all	direct	VH	0.94	S	0.93
18	all	inverse	H	0.61	ND	0.00
21	23	direct	VH	0.94	ND	0.02
21	23	inverse	VH	1.00	ND	0.02
21	99	direct	VH	0.86	S	1.00
21	99	inverse	VH	1.00	S	1.00
21	all	direct	VH	1.00	S	1.00
21	all	inverse	VH	0.90	S	1.00
23	24	direct	H	0.65	ND	0.00
23	24	inverse	H	0.69	ND	0.00
23	98	direct	L	0.40	M	0.70
23	98	inverse	VH	1.00	S	1.00
23	all	direct	VH	0.87	S	0.90
23	all	inverse	VH	1.00	S	1.00
24	25	direct	VH	1.00	S	1.00
24	25	inverse	L	0.48	W	0.18
25	30	direct	VH	1.00	S	1.00
25	30	inverse	VH	1.00	ND	0.00
25	42	direct	VH	1.00	S	1.00
25	42	inverse	VH	1.00	ND	0.00
25	55	direct	VH	1.00	S	1.00
25	55	inverse	VH	1.00	ND	0.00

25	76	direct	VH	1.00	S	1.00
25	76	inverse	VH	1.00	ND	0.00
25	all	direct	VH	1.00	W	0.20
25	all	inverse	L	0.45	S	1.00
99	100	direct	H	0.62	ND	0.00
99	100	inverse	H	0.64	ND	0.00
99	178	direct	L	0.41	M	0.70
99	178	inverse	VH	1.00	S	1.00
99	all	direct	H	0.65	S	0.90
99	all	inverse	VH	1.00	S	1.00
100	101	direct	L	0.32	S	1.00
100	101	inverse	L	0.27	W	0.21
101	103	direct	VH	1.00	S	1.00
101	103	inverse	VH	1.00	ND	0.00
101	126	direct	VH	1.00	S	1.00
101	126	inverse	VH	1.00	ND	0.00
101	139	direct	VH	1.00	S	0.99
101	139	inverse	VH	1.00	ND	0.00
101	150	direct	VH	1.00	S	1.00
101	150	inverse	VH	1.00	ND	0.00
101	all	direct	VH	0.88	W	0.26
101	all	inverse	H	0.71	S	1.00

Table 5: Raw data on confidence and decision for rules collected with the expert; quantitative decision is Dec*, cf. Section C.3; qualitative scale for confidence is VH Very High, H High, L Low, VL Very Low; qualitative scale for decision is N No Decision, W Weak, M Moderate, S Strong acceptance for direct rules and rejection for inverse rules

Table 6 presents the qualitative and quantitative confidence and decision provided by the expert for each considered rule. Note that there are no values for goals 108, 154 and 168 because the expert had not time to read the corresponding articles, cf. Section D.1.1.1.

Goal	Expert Confidence		Expert decision	
	Qualitative	Quantitative	Qualitative	Quantitative
36	VH	1.00	MR	0.12
48	VH	1.00	WA	0.60
60	VL	0.10	ND	0.49
82	L	0.30	WR	0.29
132	L	0.49	WA	0.63
143	L	0.38	ND	0.58

Table 6: Raw data on confidence and decision for goals linked to solutions collected with the expert; quantitative decision is Dec, cf. Section C.3; qualitative scale for confidence - VH Very High, H High, L Low, VL Very Low; qualitative scale for decision - MR Moderate Rejection, WR Weak Rejection, ND No Decision, WA Weak Acceptance

Moreover, the expert indicated for solution 60 a quantitative assessment for each article:

- [Wong 2018] Confidence = 0.10 Decision = 0.50,
- [Huang 2017] Confidence = 0.70 Decision = 0.00,
- [Hein 2017] Confidence = 0.60 Decision = 0.51.

D.3 Debriefing with the expert

During the debriefing with the expert eleven specific issues were addressed.

D.3.1 Confidence level

Taking example of goals 24 and 25, the expert was questioned about the reason he *indicated low confidence for rejection of goal when sub-goal is false* and *high confidence for acceptance of goal when sub-goal is true*.

The expert indicated that another sub-goal that could invalidate the rejection may exist and that his knowledge of the state of the art may not include recent advances.

D.3.2 Weak or absent decision for low confidence

Taking example of goal 24, the expert was questioned about its interpretation of “No decision”, “Weak acceptance" or “Weak rejection” when his confidence is low.

The expert indicated that the list of methods proposed by the AC could be incomplete. He stated that it is likely that the model is not robust but something else showing robustness could exist.

D.3.3 Weak or absent decision for high confidence

Taking example of goal 25, the expert was questioned about his interpretation of “No decision”, “Weak acceptance" or “Weak rejection” when its confidence is high.

For the expert, the multiple sub-goals of goal 25 indicate a logical OR. When one single sub-goal is false another sub-goal could be true. In consequence, the expert thinks that it is not useful to take a decision for goal 25.

D.3.4 Competition between considered norms

The expert was questioned about its remark “l2 ball is included in linf ball” for goal 21.

The expert understood the decomposition of goal 21 as a logical AND, i.e. considering the assumption linked to goal 21, this goal is to be robust with respect to both norms with a common radius. For a common radius linf robustness implies l2 robustness but l2 robustness does not imply linf robustness.

D.3.5 Restructuration of the AC

The choice of a specific norm and criterion to express the robustness requirement could be application dependent. This means that we should probably have different assurance cases for different couples (norm, criterion). The number of possible combinations being potentially large, the expert was asked what the most pertinent combinations could be.

For the expert the choice of the robustness norm depends on the use case. Moreover, he indicated that for images a robustness radius is not sufficient, and that robustness radius is not relevant for badly classified samples. He thinks that the AC should be developed by use case or condition types.

D.3.6 Lower decision for conjunction of sub goals

The expert was questioned about the fact that for goal 18 he indicated a decision for the inverse rule supported by all sub-goals lower than the decision for the inverse rule supported by goal 19 or by goal 21.

The expert indicated that it was a reasoning error. He thinks that decision and confidence are the same as the decision and confidence for the inverse rule supported by goal 19. That is:

Supported goal	Supporting sub-goal(s)	Rule type	Expert Confidence Qualitative	Quantitative	Expert decision Qualitative	Quantitative
18	all	inverse	VH	0.90	M	0.74

Nevertheless, he thinks that, when the data set is not useful for assessment, no decision can be taken. Note: goal 19 is “The <verification set> is relevant for robustness evaluation”.

D.3.7 Josang constraint for acceptation

The expert was questioned about the fact that for goal 100 directly supported by goal 101 he indicated a very high acceptance, 1.00, with low confidence, 0.32. The questions of changing the assessment for respecting Josang constraint and of issuing warning in case of disrespect were raised.

The low confidence reflects the fact that the expert did not have the time to analyze all referenced articles. The very high acceptance is related to the fact that at least one article can be considered.

The assessment for respecting the Josang constraint would have been to reduce the decision to the level of confidence for safety reasons. If he had time to read all articles, he would increase confidence and reduce less decision.

Concerning issuing a warning, it would have caused some pressure to read more articles and he would have made a remark. For the expert, the difference between decision and confidence expresses the fact that the sub goals may be good practices, but he has not enough expertise to state it.

D.3.8 Josang constraint for rejection

The expert was questioned about the fact that for the inverse rule of goal 25 supported by all sub goals he indicated a very high rejection, 1.00, with a low, 0.45, confidence.

Here again the low confidence is related to an expertise not covering all articles. If all supported goals are false, he rejects the top goal. But maybe it exists another method not listed here and not known by him that can fulfill the top goal.

Concerning complying with the Josang constraint he would increase the confidence to the level of decision.

D.3.9 Showing the strategies

The expert was questioned, based on goal 21, about the change of his uncertainty assessment that is induced by displaying the strategy.

He indicated that he understood the support of goal 21 by goal 23 linked to l2 robust and goal 99 linked to linf robust as a logical “and”. But, looking at the strategy, he understood that it is a logical “or”. If the strategy were shown, he would have made the same decision for goal 23 than for goal 99.

D.3.10 Link between data and robustness

Based on his inability to assess goal 19, the expert was questioned about the link between data and robustness.

The expert indicated that it exists a lot of criteria for building a dataset, but he doesn’t know criteria devoted to robustness. Those may exist but may be use case dependent.

Difficulties are providing all samples with the same robustness radius and making trade-offs in the case of data with highly interleaved classes. May be a post assessment of the dataset is needed for verifying robustness results. In conclusion data shall be representative of the problem and the operational context of the ML model.

D.3.11 Incompatibility between robustness methods

Based on expert comments on goals 24, linked to l2, and 100, linked to linf, indicating that some methods cannot be applied simultaneously, he was questioned about compatibility between methods.

For both norms, Lipschitz training is compatible with adversary training and random noising, but it presents no interest. Moreover, Lipschitz training is incompatible with randomized smoothing and feature pruning.

For Lipschitz training, linf robustness is more difficult to obtain than l2 robustness but provides more guaranties. If linf robustness is obtained, then, for the same radius, l2 robustness is obtained. Nevertheless, it not possible to apply at the same time the Lipschitz method for linf and the Lipschitz method for l2. The choice between l2 and linf is use case dependent. Moreover, robustness is not reduced to those norms. For instance, for image processing l2 robustness is required but also robustness with respect to rotation, to luminosity variation, …

Globally, it is impossible to apply all methods at the same time.

E. Results

E.1 Completing the AC for unassessed goals

Goals 30, 42, 55, 76, 103, 126, 139, 150 and 165 are not assessed through the questionnaire. However, they have the *Ix* and *Method* of Goals 36, 48, 60, 82, 108, 132, 143, 154 and 168 respectively. In the full AC they are connected through structures like the structure of Figure 7 where, for instance GOA1 corresponds to goal 30, GOA3 corresponds to goal 36 and <Method X> corresponds to Jacobian regularization. Without a concrete use case with a specific ML model, it is not possible to assess GOA2 and GOA4 in Figure 7. Thus goals 30 and 36 are linked for uncertainty propagation by a simple argument with no uncertainty.

Goals 98 and 178 corresponding to robustness by evaluation are not assessed through the questionnaire. At the first order, it is assumed that the evaluation provided a full confidence in robustness and that their assessment is *Bel* = (1, Acceptance, with Very High Confidence) and *Disb* = (0, Rejection, with Very Low Confidence).

Despite being in the questionnaire goals 19, 108, 154, and 168 were not assessed by the expert. Goals 108, 154 and 168 are dismissed because their branches lead almost directly to a choice node with multiple incompatible alternatives. The case of goal 19, is more complex. Indeed, during the debriefing the expert suggested a quite different property than relevance for data without a clear link with robustness, cf. Section D.2.10. Thus, the structure of goal 18 is changed to a simple argument with sub goal 21. Uncertainty of rules for this simple argument is derived from the answers in the form to questions concerning goal 21 alone.

E.2 Elicitation problems

The relevance of expert's judgments is strongly dependent on the quality of the formulation of the GSN goals. Indeed, the clearer the goals are, the more precise and consistent the assessments are provided. It is not unusual for an expert to give two different assessments of the same goal, formulated in two different ways. Hence the importance and difficulty of formulating clear, comprehensible reasoning for experts from different fields.

Today, there is no systematic method to design an assurance case using GSN formalism. However, four important points can be drawn from this first application.

- First, the use of a goal arguing the inference between a goal and its sub-goal(s) should be avoided in an argument. Indeed, terms such as "assures", "demonstrates" and so on, that relate the property argued in the parent goal to the applied method(s) in sub goal(s), have the same meaning as an "*is supported by*" arrow. For instance, using the statement “Jacobian regularization method ensures that the trained ML model is l2 locally robust” to justify that “The trained ML model is l2 locally robust”. In this case the arrow between the latter statement and “Results of the application of Jacobian regularization method were satisfying” is sufficient to display the support between the application of the method and the satisfaction of the property (i.e., local robustness).

- Then, it is good practice to present the assurance case to the expert before the evaluation stage.

- Moreover, the results confirm that it is important to schedule a debriefing session to explain any potential inconsistency in the results. The key here is to provide as much information as possible on the significance of each component in the GSN formalism including strategies and choices, and to facilitate understanding of the assurance case.

- Finally, the assessment of goals associated with solutions consumes a considerable amount of time and effort because it usually relies on reading and understanding the documentation (e.g. scientific publications) referenced in the solution components. So far, no study has been carried out to find the optimal number of questions to ask the expert, particularly those concerning solutions, to guarantee a compromise between the efficiency of the evaluation approach and the ease of data elicitation.

[Rushby-15] work, cited in the deliverable 613B: Assurance case – Guidelines, differentiates between two types of nodes: Reasoning nodes and Evidential nodes. Reasoning nodes, corresponding to the pure logical decomposition of a parent node, do not require an evaluation (assuming that they are correct): confidence on these nodes is maximal (certain). On the other hand, Evidential nodes, which can relate to solution nodes, require evaluation because they may be subject to reasonable doubt. In practice, we notice that according to expert evaluation, even rules which represent the inference between goals can be questioned. Indeed, at no point did the expert identify an argument as purely conjunctive or disjunctive. For example, when it comes to arguing local robustness by design or by verification, the evaluation results characterize this node as a hybrid, when the strategy suggests it as a disjunctive node.

E.2.1 Disrespect of Josang constraint

For rules, the answers not respecting Josang constraint are reported in Table 7 together with original belief and belief bounded by Josang constraint. Note that the results presented in Table 7 are obtained by following the choice made in Section C.3.2, i.e. to decrease decision Dec*. Alternatively, Josang constraint can be respected by increasing confidence.

Supported goal	Supporting sub-goal(s)	Rule type	Confidence	Expert decision	Proposed decision	Belief	Bounded belief
21	99	direct	0.86	1.00	0.93	0.93	0.86
21	all	inverse	0.90	0.00	0.05	0.95	0.90
23	98	direct	0.4	0.85	0.7*	0.55	0.40
23	all	direct	0.87	0.95	0.935	0.885	0.87
25	all	inverse	0.45	0.00	0.275***	0.725	0.45
99	178	direct	0.41	0.85	0.705*	0.555	0.41
99	all	direct	0.65	0.95	0.825**	0.775	0.65
100	101	direct	0.32	1.00	0.66***	0.66	0.32
101	all	inverse	0.71	0.00	0.145	0.885	0.71

Table 7: Impact of considering Josang constraint on decision and belief in rules; decision is on the common [0, 1] scale for direct and inverse rule, conversion to the [0, 1] scale for acceptance and [0, 1] scale for rejection in the form shall be performed by applying 2.Dec - 1 and 1 – 2.Dec respectively, where Dec is the decision in the table. Qualitative change: * moderate to weak, ** strong to moderate, ***strong to weak

The quantitative changes proposed for respecting Josang constraint implies, for some rules, qualitative changes. For instance, for the inverse rule involving all sub goals of goal 25 the qualitative decision changes from strong rejection to weak rejection.

However, during the debriefing the expert indicated that for a rejection, i.e. for an inverse rule, he would have increased the confidence instead of decreasing the decision, cf. Section D2.8. This suggests a transformation different from the one of Table 1. Table 8 presents this suggested transformation.

Table 8. From $D \times C$ to $(Y(A), Y(\neg A)) \in L \times L$ following the expert way of thinking

Dec. / Conf.	$\mathbf{0}_D$	d_{-2}	d_{-1}	d_0	d_1	d_2	$\mathbf{1}_D$
$\mathbf{0}_C$	$(\mathbf{0}_L, \mathbf{1}_L)$	(λ_1, λ_2)	(λ_1, λ_1)	$(\mathbf{0}_L, \mathbf{0}_L)$	$(\mathbf{0}_L, \mathbf{0}_L)$	$(\mathbf{0}_L, \mathbf{0}_L)$	$(\mathbf{0}_L, \mathbf{0}_L)$
c_1	$(\mathbf{0}_L, \mathbf{1}_L)$	(λ_1, λ_2)	(λ_1, λ_1)	(λ_1, λ_1)	(λ_1, λ_1)	(λ_1, λ_1)	$(\lambda_1, \mathbf{0}_L)$
c_2	$(\mathbf{0}_L, \mathbf{1}_L)$	(λ_1, λ_2)	(λ_2, λ_2)	(λ_2, λ_2)	(λ_2, λ_2)	(λ_2, λ_1)	$(\lambda_2, \mathbf{0}_L)$
$\mathbf{1}_C$	$(\mathbf{0}_L, \mathbf{1}_L)$	$(\lambda_1, \mathbf{1}_L)$	$(\lambda_2, \mathbf{1}_L)$	$(\mathbf{1}_L, \mathbf{1}_L)$	$(\mathbf{1}_L, \lambda_2)$	$(\mathbf{1}_L, \lambda_1)$	$(\mathbf{1}_L, \mathbf{0}_L)$

It also changes the quantitative adjustment for one of the cases of violation of Josang constraint to:

When $Dec(p) < \frac{1-Conf(p)}{2}$, $Conf(p) \leftarrow 1 - 2Dec(p)$.

This would change the assessment of inverse rules in Table 7. This change is presented in Table 8.

Supported goal	Supporting sub-goal(s)	Rule type	Expert confidence	Decision	Proposed confidence	Belief	Bounded belief
21	all	inverse	0.90	0.00	1.00	0.95	1.00
25	all	inverse	0.45	0.00	1.00**	0.725	1.00
101	all	inverse	0.71	0.00	1.00*	0.885	1.00

Table 7: Impact of considering Josang constraint, as suggested by the expert, on confidence and belief in rules; decision is on the common [0, 1] scale, conversion to the [0, 1] scale for rejection in the form shall be performed by applying 1 – 2.Dec, where Dec is the decision in the table. Qualitative change: * high to very high, ** low to very high

Concerning the goals directly supported by solutions, the only disrespect of Josang constraint occurs for goal 82, with an expert decision of 0.29 for a confidence of 0.3 leading without change to a belief of -0.06 and a disbelief of 0.36. The decision value proposed for complying with Josang constraint in the most currently way is 0.35. This value leads to a belief of 0.0 and a disbelief of 0.3. This change does not impact the weak rejection qualitative assessment.

The confidence value proposed for complying with Josang constraint following the expert way of thinking is 0.42. This value leads to a belief of 0.0 and a disbelief of 0.42. This change does not impact the low confidence qualitative assessment.

E.2.2 Inconsistency between elementary and conjunctive rules

For the inverse rules related to Goal 18 the belief when all premises are considered is lower than the one when only Goal 19 or Goal 21 are considered. The origin of the problem is inconsistent answers to questions:

- Q2: Assuming Goal 19 is false, what is your Decision in the conclusion Goal 18? **Strong rejection**.
- Q4: Assuming Goal 21 is false, what is your Decision in the conclusion Goal 18? **Moderate rejection**.
- Q6: Assuming Goal 19 and Goal 21 are false, what is your Decision in the conclusion Goal 18? **No decision.**

The same problem is observed with quantitative values only for the inverse rules related to Goal 21, and for Goal 25 and 101 with quantitative values for the direct rules and for qualitative and quantitative values for the inverse rules. However, it is important to note that the function of those nodes is to decide which methods will be used for the ML model design and that the expert indicated an impossible conjunction of all methods, see “E.5.3.5 Impossible conjunction”.

E.3 Elicitation of uncertainty associated to rules

Table 8 presents the elicitation of qualitative and quantitative belief of direct and inverse rules.

Supported goal	Supporting sub-goal(s)	Direct quantitative belief	Direct qualitative belief	Inverse quantitative belief	Inverse qualitative belief
15	17	1.000	VH	1.000	VH
17	18	1.000	VH	0.915	VH
18	*19*	*0.505*	*VH*	*0.820*	*VH*
18	21	0.765	VH	0.845	VH
18	*all*	*0.935*	*VH*	*0.305*	*H*
21	23	0.480	VH	1.000	VH
21	99	0.860*	VH	1.000	VH
21	all	1.000	VH	0.900*	VH*
23	24	0.325	H	0.345	H
23	98	0.400*	L	0.345	VH
23	all	0.870*	VH	1.000	VH
24	25	1.000	VH	0.330	L
25	30	1.000	VH	0.500	VH
25	42	1.000	VH	0.500	VH

25	55	1.000	VH	0.500	VH
25	76	1.000	VH	0.500	VH
25	all	0.600	VH	0.450*	L*
99	100	0.310	H	0.320	H
99	178	0.410*	L	1.000	VH
99	all	0.650*	H	1.000	VH
100	101	0.320*	L	0.240	L
101	103	1.000	VH	0.500	VH
101	126	1.000	VH	0.500	VH
101	139	1.000	VH	0.500	VH
101	150	0.995	VH	0.500	VH
101	165	1.000	VH	0.500	VH
101	all	0.570	VH	0.710*	H*

Table 8: Belief associated with rules; VH Very High, H High, L Low, VL Very Low; * Modified for respecting Josang constraint by projecting at constant confidence; in red, inconsistent belief assignment for set of rules*; in italic, not relevant because of absence of assessment for goal 19, cf. Section E.1*

Interestingly, no rule presents a null quantitative belief or a Very Low qualitative belief. This may reflect the fact that all arguments are hybrid arguments or that it is quite difficult for the expert to use the form to indicate that he doesn't believe in an implication. Indeed, to indicate the absence of belief he would have to set at the same time 0 for its confidence and 0 for its decision Dec*.

Concerning the inconsistency for the inverse rules of goals 21, 25 and 101, it is interesting to note that, applying the adjustment for respecting the Josang constraint suggested by the expert, all rules for all sub goals are assessed quantitatively by 1.00 and qualitatively by VH, solving the consistency problem for those rules.

Concerning the inconsistency for the direct rules for goals 25 and 101, the hidden strategy is a choice between robustness methods and the expert indicated that it is not possible to use all proposed methods at the same time, cf. Section D.2.11. Thus, the assessment of the direct rule for all sub goals is not relevant.

Concerning the inconsistency for the inverse rule of goal 18, the expert indicated that the assessment of the rule for all sub goals should be equal to the assessment of the rule for sub goal 19, cf. Section D.2.6. This solves the situation from a qualitative point of view, but a problem remains from a quantitative point of view because the belief of the inverse rule for sub goal 19, 0.820, is lower than the belief of the inverse rule for sub goal 21, 0.845. We decided not to solve this slight inconsistency.

Changes performed in Table 8 are summarized in Table 9.

Supported goal	Supporting sub-goal(s)	Direct quantitative belief	Direct qualitative belief	Inverse quantitative belief	Inverse qualitative belief
18	*all*	*0.935*	*VH*	*0.820*	*VH*
21	all	1.000	VH	1.000	VH
25	all	-	-	1.000	VH

101	all	-	-	1.000	VH

Table 9: Final belief associated with rules. VH Very High; *in italic, not relevant because of absence of assessment for goal 19, cf. Section E.1*

E.4 Elicitation of uncertainty for goals associated with solutions

Table 10 presents the elicitation of qualitative and qualitative belief and disbelief for goals associated with solutions. There are no values for goals 108, 154 and 168 because the expert had no time to read the corresponding articles.

Goal	Quantitative belief	Qualitative belief	Quantitative disbelief	Qualitative disbelief
36	0.120	L	0.880	VH
48	0.600	VH	0.400	H
60	0.040	VL	0.060	VL
82	0.000*	L	0.300*	L
132	0.375	L	0.115	L
143	0.270	L	0.110	L

Table 10: Belief and disbelief for goals associated with solutions. VH Very High, H High, L Low, VL Very Low, * Modified for respecting Josang constraint

Considering Tables 8 and 10 together:

- Very Low belief corresponds to numerical belief values in [0.010, 0.300]
- Low belief corresponds to numerical belief values in [0.000, 0.450]
- High belief corresponds to numerical belief values in [0.305, 0.710]
- Very High belief corresponds to numerical values in [0.345, 1.000]

E.5 Propagation

E.5.1 Choice of a ML method to be propagated

The expert indicated that there is some incompatibility between some ML methods, therefore the propagation is performed by choosing a single method and propagating its confidence.

The methods corresponding to articles that were not read by the expert are not considered in this choice. Thus, the choice is made among goals 36, 48, 60, 82, 132 and 143. Note that in the AC, goals 36, 48, 60 and 82 are connected to one choice node, the choice of a l_2 robustness method, and goals 132 and 143 are connected to another choice node, the choice of a l_{inf} robustness method: there are two different multi-criterion decision problems. Those problems are managed as described in Section C.4.3.

For goals 60 and 82, their qualitative value for belief is the same as for disbelief, respectively Very Low and Low, but their quantitative value for disbelief is larger than the one for belief. So, those goals are not considered for propagation.

Goal 36 presents less belief and more disbelief than goal 48, qualitatively and quantitatively. So, it does not belong to the Pareto front and is not considered for propagation.

In conclusion for the choice of l2 robustness method, the only option is goal 48.

For the choice of a linf robustness method, both goals 132 and 143 belong to the Pareto frond. A maximization of Leximin with 1-*Disb* gives for 132 (1/3, 0.375, 2/3, 0.885) and for 143 (0.270, 1/3, 2/3, 0.89). 1/3 being larger than 0.270, it is possible to eliminate goal 143.

Finally, two propagation hypotheses can be used:

- Goal 48 alone: Lipschitz training [Anil, 2019] for norm l2
- Goal 132 alone: Lipschitz training [Anil, 2019] for norm linf

Note that it makes no sense to propagate goals 48 “The Lipschitz training method allows to ensure that the <Trained ML model> is <l2 locally robust>” and 132 “The Lipschitz training method allows to ensure that the <Trained ML model> is <linf locally robust>” together because in the learning algorithm either the l_2 norms of weight matrices are enforced to 1 or the l_{inf} norms of weight matrices are enforced to 1. Those two enforcements are performed by two different algorithms, described respectively in Section 4.2.1 and Section 4.2.2 of the article [Anil, 2019] provided as solution for both goals.

E.5.2 Propagation of Lipschitz training for norm l_2

The part of the robustness AC where the uncertainty on Lipschitz training for norm l_2 is propagated is presented in Figure 13.

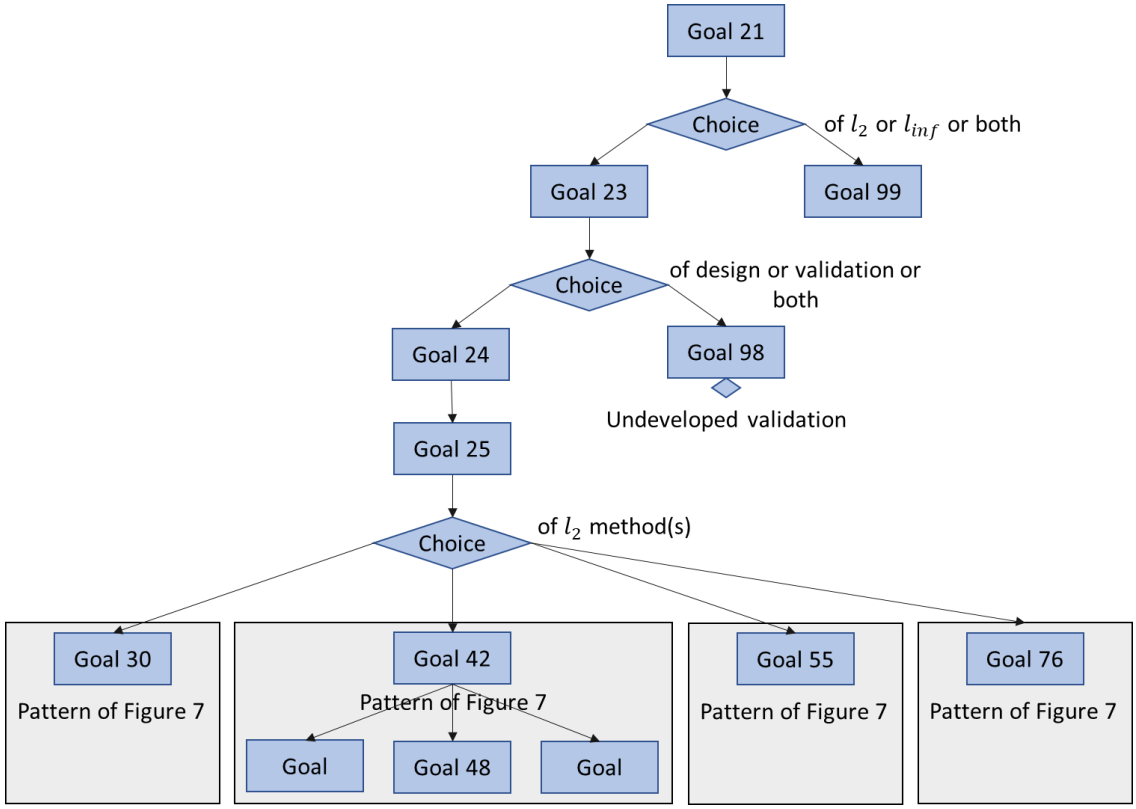


Figure 13. Part of Robustness Assurance Case where uncertainty about Lipschitz training is propagated

A summary of the propagation for Lipschitz training for norm l_2 is presented in Table 10. Goal 48 corresponds to goal 42, cf. Section C.6.3, through the pattern presented in Figure 7. With respect to this pattern goal 42 corresponds to GOA1 and goal 48 corresponds to GOA3. Thus, assuming that when the AC pattern will be used the applicability of the method is perfectly respected and the learning report is perfectly acceptable, the uncertainty of goal 48 can directly be assigned to goal 42. Goal 42 supports goal 25 as indicated in Table 8. Because goal 25 is supported by a choice, the relation between 42 and 25 can be assessed as a simple argument with belief in direct rule (1.000, VH) and belief in inverse rule (0.500, VH). Then, Goal 24 is directly supported by goal 25 through a simple argument with belief in direct rule (1.000, VH) and in inverse rule (0.330, L). Then goal 23 is supported by goal 24 and 98. Note that goal 23 corresponds to a choice node and goal 98 is an open goal, cf. Section C.6.2. Thus, the choice is to rely either only on design or on design and validation.

- If relying only on design, goal 23 can be assessed using a simple argument with belief in direct rule (0.350, H) and belief in inverse rule (0.350, H). In that case, the belief in goal 23 is (0.195, H) and its disbelief is (0.023, L).
- If relying on design and validation, some hypothesis shall be made about open goal 98. Making an optimistic hypothesis of a belief (1.000, VH) and a disbelief (0.000, VL) and assessing goal 23 as the result of a hybrid argument, leads to a belief in goal 23 of (0.719, VH) and a disbelief in goal 23 (0.014, L). Note that the level of conflict of the hybrid argument is (0.009, L). This low level indicates that there is no contradiction between relying on design and relying on validation.

Here, the comparison of the alternatives allows the ML model designer to choose to rely on design and validation. Then, goal 21 is supported by goals 23 and 99. It is a choice goal and 99 is not selected because no design method for linf robustness is used. Thus, goal 21 can be assessed through a simple argument with belief in direct and inverse rules of respectively (0.480, VH) and (1.000, VH).

Note that the expert stated that if strategies were shown he would have assessed the direct and inverse rules between 23 and 21 in the same way as between 99 and 21, cf. Section D.2.9. In that case the belief in direct and inverse rules would have been respectively (0.860, VH) and (1.000, VH) leading to an assessment for goal 21 of (0.618, VH) for belief and (0.014, L) for disbelief.

Goal	Quantitative belief	Qualitative belief	Quantitative disbelief	Qualitative disbelief
42	0.600	VH	0.400	H
25	0.600	VH	0.200	H
24	0.600	VH	0.066	L
23	0.719	VH	0.014	L
21	0.345	VH	0.014	L

Table 10: Propagation for Lipschitz training for norm l_2

E.5.3 Propagation of Lipschitz training for norm l_{inf}

A summary of the propagation for Lipschitz training for norm l2 is presented in Table 11. Goal 132 corresponds to goal 126. Thus, assuming that when the AC pattern will be used the applicability of the method is perfectly respected and the learning report is perfectly acceptable by certification authorities, the

uncertainty of goal 132 can directly be assigned to goal 126. Goal 126 supports goal 101 as indicated in Table 6. Because goal 101 is supported by a choice, the relation between 126 and 101 can be assessed as a simple argument with belief in direct rule (1.000, VH) and belief in inverse rule (0.500, VH). Then, Goal 100 is directly supported by goal 101 through a simple argument with belief in direct rule (0.320, L) and in inverse rule (0.240, L). Then, goal 99 is supported by goal 100 and 178. Note that goal 99 corresponds to a choice node and goal 178 is an open goal. Thus, the choice is to rely either only on design or on design and validation.

- If relying only on design, goal 99 can be assessed using a simple argument with belief in direct rule (0.310, H) and belief in inverse rule (0.320, H). In that case, the belief in goal 99 is (0.037, L) and its disbelief is (0.057, L).
- If relying on design and validation, some hypothesis shall be made about open goal 178. Making an optimistic hypothesis of a belief (1.000, VH) and (0.000, VL) and assessing goal 99 as the result of a hybrid argument, leads to a belief in goal 99 of (0.462, L) and a disbelief in goal 99 (0.003, L). Note that the level of conflict of the hybrid argument is (0.002, L).

Here, the comparison of the alternatives allows the ML model designer to choose to rely on design and validation. Then, goal 21 is supported by goals 23 and 99. It is a choice goal and 23 is not selected because no design method for l2 robustness is used. Thus, goal 21 can be assessed through a simple argument with belief in direct and inverse rules of respectively (0.860, VH) and (1.000, VH).

Goal	Quantitative belief	Qualitative belief	Quantitative disbelief	Qualitative disbelief
126	0.375	L	0.115	L
101	0.375	L	0.057	L
100	0.120	L	0.014	L
99	0.462	L	0.003	L
21	0.430	L	0.003	L

Table 11: Propagation for Lipschitz training for norm l_{inf}

E.5.4 Choice of a norm for Lipschitz training

Comparison of the last lines of tables 8 and 9 helps to determine which norm should be used for Lipschitz training. From a qualitative point of view, l_2 shall be chosen because it leads to a Very High belief instead of a Low belief for l_{inf} while the disbelief is Low for both norms. From a quantitative point of view, l_{inf} shall be chosen because, when compared with l_2, its belief is larger, and its disbelief is lower. However, note that the large decrease of belief from goal 23 to 21 may be related to a misunderstanding of the partitioning with respect to norms by the expert, see paragraph D.2.9. Finally, a ranking with Leximin the values for l_2 are (0.345, 2/3, 0.986, 1) and the values for l_{inf} are (1/3, 0.430, 2/3, 0.997). 0.345 being larger than 1/3, l_2 is preferred.

Moreover, considering the values the expert would have given if the strategy was shown, they would lead for l_2 to (0.618, 2/3, 0.986, 1) and the choice would be the same.

In conclusion norm l_2 is chosen and propagation continues based on last line of Table 10 and on alternative values corresponding to an expert informed about the strategies.

E.5.5 Propagation towards top goal

Table 12 presents summary of the propagation for Lipschitz training after the choice of norm l_2 for an expert not informed about strategies. Goal 18 is supported by goals 19 and 21 and the expert indicated that he was unable to assess goal 19. Thus, goal 18 is assessed using a simple argument and the values of Table 8 for goal 21 supporting goal 18. Then goal 17 and 15 are assessed as simple arguments.

Goal	Quantitative belief	Qualitative belief	Quantitative disbelief	Qualitative disbelief
18	0.264	VH	0.012	L
17	0.264	VH	0.011	L
15	0.264	VH	0.011	L

Table 12: Continuing propagation for Lipschitz training for norm l_2 for an expert not informed about strategies

Table 13 presents summary of the propagation for Lipschitz training after the choice of norm l_2 for an expert not informed about strategies. The principle is the same as before but the values for quantitative belief are different.

Goal	Quantitative belief	Qualitative belief	Quantitative disbelief	Qualitative disbelief
18	0.473	VH	0.012	L
17	0.473	VH	0.011	L
15	0.473	VH	0.011	L

Table 13: Continuing propagation for Lipschitz training for norm l_2 for an expert informed about strategies

E.5.6 Conclusion

As shown by this example, uncertainty evaluation allows the ML model designer to make an informed choice between learning methods and norms. He also has an idea of the probability that the argument holds. This probability is in the interval [0.264, 0.989] for an expert not informed about strategies and [0.473, 0.989] for an expert informed about strategies.

The approach also allows identifying the main weakness of the argument, i.e., the one characterized by the large decrease in quantitative belief for the support of goal 21 by goal 23. The ML model designer can address this weakness by avoiding mentioning l_{inf} norm to the authority or, if the authority is aware of l_{inf} norm, by claiming that the radius used for learning is much larger than the one that would have been used

in case of l_{inf} norm. Indeed, if the radius used for l_2 norm is larger or equal than the required radius for l_{inf} norm multiplied by the square root of the input space dimension, the requirement on l_{inf} norm is fulfilled.

E.6 Qualitative analysis

E.6.1 Too demanding expert effort

The expert indicated that he didn't analyze articles related to goals 108, 154 and 168, i.e., Empirical robustness reinforcement method, Certified robust training and Random Noising for linf robustness. It seems that the reason is the amount of effort needed to fill seriously the questionnaire is too large. Indeed, this evaluation procedure requires considerable time and effort to complete the questionnaire especially for parts concerning the objective/solution(s) nodes, which require the reading and processing of extensive documentation (e.g., technical reports, scientific articles, etc.).

E.6.2 Definitions

E.6.2.1 Restrictive definition

Concerning the definition of robustness, the expert indicated that the definition of robust provided by the AC is restrictive. For instance, this definition doesn't cover robustness with respect to distribution shift.

E.6.2.2 Unclear definition

The expert thinks that in the definition of <Global nbsample robustness criteria>, i.e., "the number of samples of a subset that are <locally robust> is greater than a threshold", a criterion of representativity of the "subset" is needed.

The expert found that the wording of goal 21 is incomplete because <local robustness> is defined for a single input while it supports the goals 18 that is grounded on <Global nbsample robustness criteria> that refers to several inputs. A consistent wording for goal 21 could be "<The Trained ML model> is <locally robust> for a sufficient number of inputs". The expert considered such wording. The addition of "for a sufficient number of inputs" could also be done for goals 23, 24, 25, 99, 100, 101 and for all goals of table 2.

E.6.2.3 Absence of definition

The expert stated that he was unable to assess Goal 19 whose wording is "The <verification set> is relevant for robustness evaluation" and support is "<Verification set>" because the definition of a relevant verification set is not provided. Nevertheless, he indicated values for the answers to the questions.

E.6.3 Contexts

For the context associated to goal 101, the expert has some doubts about the applicability for linf robustness of all methods among Double Backpropagation, Jacobian regularization, Saturated Network, Ensemble adversarial training, Lipschitz Training, Wong_Kolter, Universal Random Smoothing, Feature pruning and Random Noising. Moreover, the expert has specific doubt about Lipschitz Training even if he thinks that the method helps obtaining linf robustness.

E.6.4 Relations between goal and sub-goals

E.6.4.1 Dependent sub goals

The expert signaled that, for a given perturbation radius, goal 99 implies goal 23 because the l2 ball is included in the linf ball. This is true from a formal point of view, but the hidden Strategy is “Argument by partitioning of norms”. It seems that the expert has understood goal 21 as “<The Trained ML model> is <locally robust> for any norm with the same radius”.

E.6.4.2 Impossible conjunction

The expert indicated that the conjunction of goals 30, 42, 55 and 76 is impossible because the methods cannot be applied together at learning time. This also applies to goals 103, 126, 139, 150 and 165. Moreover, for the negation of the use of all methods he assumed that those methods are the only available methods.

E.6.5 Relations between goal and solutions

E.6.5.1 Goals with multiple solutions

The expert pointed out that when multiple solutions are provided for a goal, it is unclear whether the goal shall be assessed as supported by a logical “and” or by a logical “or” of solutions.

E.6.5.2 Scientific feedback

Some articles are subject to a deep analysis by the expert. For Jacobian regularization [Jacubovitz, 2019] the expert concludes that it improves l2 robustness but doesn’t ensure it. For Lipschitz training [Anil, 2019] he indicates that a specific loss function should be used as done in recent work [Serrurier, 2021]. For Certified robust training for l2 robustness, the expert indicates that one article [Huang, 2017] is out of scope.

F. Analysis of results and limitations

F.1 Comparison of qualitative and quantitative approaches

The result on comparison of approaches indicates that uncertainty modeling in AC is useful and that, when considering relevant requirements, the assessment of uncertainty shall be performed at the same time with both qualitative and quantitative approaches. This leads to a valuation of goals by four elements: the quantitative belief, the qualitative belief, the quantitative disbelief and the qualitative disbelief. A limitation associated with this result is that there is no total order between goals assessed following different strategies. Thus, a multicriteria reasoning shall be performed for choosing the best solution.

F.2 Elicitation

The choice of methodology is to hide from the expert the strategies and choices. The results show that with the information included in the strategy the expert can make a quite different uncertainty assessment of rules than without this information. Moreover, this difference may lead to a quantitative difference in the assessment of the AC property. The methodology could be revised concerning hiding or not the strategies.

The procedure and associated Excel file type for uncertainty elicitation developed here is based on scrollbars actuated by the expert. Each scrollbar drives at the same time a numerical value and a semantic qualifier. The scrollbar associated with decision is totally independent from the scrollbar associated with confidence. However, the Josang constraint must be respected. The results indicate that, when the Josang constraint is violated, the projection may depend on the context. This limitation could be addressed by asking first the question about confidence and limiting the decision scrollbar by the confidence value.

The absence of automatic enforcement of consistency between rules at elicitation time is also a serious limitation.

Finally, in case of large choices with incompatible sub goals, the question for all sub goals true and the question for all sub goals false are irrelevant. The possibilities for sub goals combinations should be assessed before making uncertainty assessment.

F.3 Uncertainty propagation

For most nodes, uncertainty propagation is quite easy and for the case study the conflict mass value is always low indicating that there is no strong contradiction inside the argument.

However, at choice node uncertainty propagation relies on building consistent sets of sub goals and on performing a multi-criterion choice among those sets. This would require a better definition of the choice and it is not sure that the propagation could be fully automatized at those nodes.

F.4 Case study

The case study highlights the benefits and some limitations of the proposed methodology. However, limited effort and time inducted additional limitations:

- Only one expert has been involved. It is impossible to distinguish between one the one hand the results that are specific to this expert and on the other hand the results that could be consolidated with a large panel of experts.
- Uncertainty has not been assessed on the whole AC for robustness of ML. Some elements, that are not purely logical were not considered, for instance the branches corresponding to two alternative definitions of robustness and the branches corresponding to verification.
- The expert had the possibility to not assess a node or to indicate that something is missing in the argument of a node and used this possibility. This induced some doubts about the structure of the AC.

G. Conclusion

A methodology for assessing the uncertainty in AC has been proposed and has been demonstrated using it as a case study. The result of the work indicates some limitations:

- For the choice nodes the methodology should be adapted to capture the incompatibility constraints between the sub goals before making uncertainty assessment.
- The AC presented to the expert should be considered as frozen and possibilities to consider a goal not assessable and to complete the argumentation should be reduced.
- The strategies and the choices should be presented to the expert to get an assessment based on as much information as possible.

Those limitations will be addressed in future works. However, globally the proposed methodology is useful and allows us to derive an indication about the probability that the argument presented in the AC holds.

At present time, uncertainty elicitation is performed using an Excel file and uncertainty propagation is performed using an ad-hoc Python script. An integration of the elicitation and propagation methods either in the Capella environment or in the Companion methodology [Adedjouma, 2022] should be carried out. AC in a GSN format may already be captured in the Capella environment. An optional capture of uncertainty features of rules associated to strategies and goals directly linked to solutions could be implemented in Capella.

H. Bibliography

[Adedjouma, 2022] Adedjouma, M., et al. "Towards the engineering of trustworthy AI applications for critical systems-The Confiance. ai program." (2022).

[Anil, 2019] Anil, C., Lucas, J., & Grosse, R. (2019). Sorting out Lipschitz Function Approximation, pages 291–301. In *International Conference on Machine Learning*. PMLR.

[Ayoub 2013] A., Ayoub, J., Chang, O., Sokolsky and I., Lee. Assessing the overall sufficiency of safety arguments. In 21st Safety-critical Systems Symposium (SSS'13), Bristol, United Kingdom, 2013. (Cited in pages v, 19, 20, 27, 46, 47, and 82.)

[Cohen, 2018] Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified Adversarial Robustness via Randomized Smoothing, pages 1310–1320. In *international conference on machine learning*. PMLR.

[Cyra and Gorski, 2011] Cyra, Lukasz et Gorski, Janusz. Support for argument structures review and assessment. *Reliability Engineering & System Safety*, 2011, vol. 96, no 1, p. 26-37.

[Dubois, 2019] Dubois, D., Faux, F., Prade, H., & Rico, A. (2019). A possibilistic counterpart to Shafer evidence theory. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* (pp. 1-6). IEEE.

[Destercke, 2011] Destercke, Sébastien et Dubois, Didier. Idempotent conjunctive combination of belief functions: Extending the minimum rule of possibility theory. Information Sciences, 2011, vol. 181, no 18, p. 3925-3945.

[Denœux, 2008] Denœux, Thierry. Conjunctive and disjunctive combination of belief functions induced by nondistinct bodies of evidence. Artificial Intelligence, 2008, vol. 172, no 2-3, p. 234-264.

[Dubois, 2022] Dubois, Didier, Faux, Francis, Prade, Henri, et al. Qualitative capacities: Basic notions and potential applications. International Journal of Approximate Reasoning, 2022, vol. 148, p. 253-290.

[Gao, 2017] Gao, J., Wang, B., Lin, Z., Xu, W., & Qi, Y. (2017). Deepcloak: Masking deep neural network models for robustness against adversarial samples. *arXiv preprint arXiv:1702.06763*.

[Gowal, 2019] Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., ... & Kohli, P. (2018). On the effectiveness of interval bound propagation for training verifiably robust models. *arXiv preprint arXiv:1810.12715*.

[Hein, 2017] Hein, M., & Andriushchenko, M. (2017). Formal Guarantees on the Robustness of a Classifier against Adversarial Manipulation. In *Advances in neural information processing systems, 30*.

[Hong, 2022] Hong, H., Wang, B., & Hong, Y. (2022). Unicr: Universally Approximated Certified Robustness via Randomized Smoothing, pages 86–103. In *European Conference on Computer Vision*. Cham: Springer Nature Switzerland.

[Huang, 2017] Huang, X., Kwiatkowska, M., Wang, S., & Wu, M. (2017). Safety Verification of Deep Neural Networks, pages 3– 29. In *Proceedings of Computer Aided Verification: 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017*. Springer International Publishing.

[Id Messaoud, 2022a] Id Messaoud, Y. (2022). Uncertainty assessment in safety argument structures – An approach based on Dempster-Shafer Theory. PhD thesis, UPS Toulouse.
[Id Messaoud, 2022b] Id Messaoud, Y., Dubois, D., & Guiochet, J. (2022) A qualitative counterpart of belief

functions with application to uncertainty propagation in safety cases. In *International Conference on Belief Functions*, pages 231–241. Springer.

[Jacubovitz, 2019] Jakubovitz, D., & Giryes, R. (2018). Improving DNN Robustness to Adversial Attacks Using Jacobian Regularization, pages 414–529. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

[Lecuyer, 2019] Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., & Jana, S. (2019). Certified Robustness to Adversarial Examples with Differential Privacy, pages 656–672. In *2019 IEEE symposium on security and privacy (SP)*. IEEE.

[Lequyer, 2018] Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., & Jana, S. (2018). On the connection between differential privacy and adversarial robustness in machine learning. *stat*, *1050*, 9.

[Nayebi, 2017] Nayebi, A., & Ganguli, S. (2017). Biologically Inspired Protection of Deep Networks from Adversarial Attacks. *arXiv preprint arXiv:1703.09202*.

[Ross, 2017] Ross, A., & Doshi-Velez, F. (2018). Improving the Adversarial Robustness and Interpretability of Deep Neural Networks by Regularizing their Input Gradients. In *Proceedings of the AAAI conference on artificial intelligence,* Vol. 32, No. 1.

[Shafer, 1976] Shafer, G. (1976) A mathematical theory of evidence, volume 42. Princeton university press.

[Serrurier, 2021] Serrurier, M., Mamalet, F., González-Sanz, A., Boissin, T., Loubes, J. M., & Del Barrio, E. (2021). Achieving Robustness in Classification using Optimal Transport with Hinge Regularization, pages 505–514. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

[Tramer, 2017] Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2017). Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*.

[Wang, 2019] Wang, Rui, Guiochet, Jérémie, Motet, Gilles, et al. Safety case confidence propagation based on Dempster–Shafer theory. International Journal of Approximate Reasoning, 2019, vol. 107, p. 46-64.

[Wong, 2018a] Wong, E., Schmidt, F., Metzen, J. H., & Kolter, J. Z. (2018). Scaling Provable Adversarial Defenses. In *Advances in Neural Information Processing Systems*, *31*.

[Wong, 2018b] Wong, E., & Kolter, Z. (2018). Provable defenses against adversarial examples via the convex outer adversarial polytope, pages 5286–5295. In *International conference on machine learning*. PMLR.

[Zangh, 2019] Zhang, H., Chen, H., Xiao, C., Gowal, S., Stanforth, R., Li, B., ... & Hsieh, C. J. (2019). Towards stable and efficient training of verifiably robust neural networks. *arXiv preprint arXiv:1906.06316*.

I. Annex 1: Detailed analysis of articles by the expert

When filling in the questionnaire, the expert analyzed deeply some articles proposed as solutions. The result of this analysis was not requested by the form, but the expert wrote the notes presented in Table 14 below the form of some goals.

Goal	Reading notes
36	Assertions and Method: The relation between the Frobenius norm and the ` 2 (spectral) norm of the Jacobian matrix has been shown in [30], and lays the justification for using the Frobenius norm of the network's Jacobian as a regularization term => should I read [30] to analyze this paper? We apply this regularization as additional post-processing training => Regularization only gives an additional objective. It does not ensure that this objective will be achieved. Mathematical justification: Eq9 is justified by the 'mean theorem' which hypothesis is 'differentiable'. But ReLU is not. => It seems that Lipschitz continuous is enough (should be ok). Lemma1: it may depend on variation of gradient of z_k1 and z_k2 in the neighborhood of x ? => only an approximation not a bound. Corrolary2: idem only an approximation can be larger or smaller. Proposition3: "the first approximation" … is "lower bounded" => no way an approximation is an approximation then we cannot conclude on the true value (even if we can lower bound the approximation). Experiments: The enhancement in robustness (larger ρ_{adv}) with adversarial attacks can either show that the network is more robust OR that the network is harder to attack => it doesn't ENSURE that the model is L2 Locally robust.
48	Mathematical justification: Theorem 1 shows at least what should not be done: using ReLU, Sigmoid, Tanh with Orthogonal layers. Eq 2: Bjorck orthogonalization is restricted to p steps => post verification of Lipschitz constant is necessary. ++ Eq 3: gives a formal robustness bound.

	Experiments: Provide both adversarial attack measures, and theoretical robustness ball => larger confidence in the theoretical one, even if it is lower than the empirical one. Lots of work in Ani et al Appendix which is not in the downloaded version!
60	SOL61: [Wong 2018]: Our work in this paper relates closely to techniques for the formal verification of neural networks systems (indeed, our approach can be viewed as a convex procedure for verification, coupled with a method for training networks via the verified bounds). Specifically, we develop a provably robust training procedure, based upon the approach in [Wong and Kolter, 2017] => should I also read this paper? Not done. Notations and demonstration 6 to 11 difficult to understand in a limited time => not verified. Theorem 2 => idem. We emphasize that in all cases we report the robust test error, that is, our upper bound on the possible test set error that the classifier can suffer under any norm-bounded attack => better than a single attack. Hard to understand what is done with these dual network. The code seems toindicate a loss and talk about eq. 14 but I was not able to find. SOL66: [Huang 2017]: This notion is also known as pointwise robustness [18,12] or local adversarial robustness [21]. We propose a general framework for automated verification of safety of classification decisions made by feed-forward deep neural networks +> Only verification not a Training procedure. We employ discretisation to enable a finite exhaustive search of the high- dimensional region η for adversarial misclassifications => finite exhaustive search? In high dimension? Our framework can guarantee that a misclassification is found if it exists => ?? Theorem 1 and 2 => difficult to understand in a limited time => not verified. Although we cannot theoretically confirm the minimality of Δ k , they are re-fined layer by layer and, in discrete settings, this process can be bounded from below by the unit step.=> not sure to understand if the method still provide guarantees ? Theorem 3 => not verified. We implement Algorithm 1 by utilising satisfiability modulo theory (SMT) solvers => formal method.

	For checking refinement by layer, we use the theory of linear real arithmetic with existen-tial and universal quantifiers, and for verification within a layer (0-variation) we use the same theory but without universal quantification. => Not sur to understand. NOT SURE TO BE ABLE TO VERIFY THE PAPER BUT SURE THAT IT DOESN'T TALK ABOUT TRAINING. SOL70: [Hein 2017] Theorem 2.1: seems correct. Define the Cross-Lipschitz Regularization functional a => Regularization only gives an additional objective. It does not ensure that this objective will be achieved. -- the theorem2.1 seems to require that the "cross-lipschitz" is low for any point in the ball, and the regularization only evaluates it on the dataset point. In all cases we compute therobustness guarantees from Theorem 2.1 (lower bound on the norm of the minimal changerequired to change the classifier decision), where we optimize over R using binary search=> Not a full guarantee?
82	SOL83 [Cohen, 2019] Since it is not possible to exactly evaluate the prediction of g at x or to certify the robustness of g around x, we will give Monte Carlo algorithms => statistical evaluation. Theorem1: seems correct. Proposition 2. With probability at least 1 − α over the randomness in CERTIFY => how to estimate α? SOL92 [Lecuyer, 2019] Lemma1 seems correct but necessary condition ε δ DP. Proposition1 seems correct. Learning with ε δ DP => as far as I know accuracy is far lower when learning with DP objective (66% on Cifar10)? Results in table III are weird since we don't know if these networks are ε δ DP? We therefore resort to Monte Carlo methods to estimate it at prediction time and develop an approximate version of the robustness certification in Proposition 1 => Statistical evaluation. We use Hoeffding's inequality [25] or Empirical Bernstein bounds [39] to bound the error in E (A(x)) => seems correct but need to verify hypothesis. Adding noise "before" the DNN in a separately trained auto-encoder=> Should robustness of autoencoder also evaluated? Making 10 draws brings it only to 0.02s, but 100 requires 0.13s, and 1000, 1.23s.... found that 300 draws were typically necessary to properly certify a prediction, implying a prediction

	time of 0.42s=> Main drawbacks of randomized smoothing need to apply many steps of inference. We share with DP ML (and most other applied DP literature) are DP theory and mechanisms. The goal of DP ML is to learn the parameters of a model while ensuring DP with respect to the training data. .....PixelDP's goal is to create a robust predictive model => Thus not sure that leamma1 holds, need verification where a small change to any input example does not drastically change the model's prediction on that example. SOL87 [Hanbin, 2022] Theorem1: seems the same as in [Cohen, 2019]? Sorry not verified
143	[Gowal, 2019] We study interval bound propagation (IBP), which is derived from interval arithmetic [14, 15, 28]: an incomplete method for training verifiably robust classifiers => why incomplete? Rely on formal method: approximation by lower and upper bound evaluation. Equation12: robustness is achieved using a regularization term => Regularization only gives an additional objective. It does not ensure that this objective will be achieved. For each example of the test set, a MIP is solved using Gurobi with a timeout of 10 minutes. Upon timeout, we fallback to solv- ing a relaxation of the verification problem with a LP [15] using Gurobi again. When both approaches fail to provide a solution within the d time, we count the example as attackable. ==> Formal verification of the attacks gives more confidence. But it may not scale for larger networks.

Table 14: Additional information provided by the expert

Title: Assurance Cases – Uncertainty Assessment

Keywords: Uncertainty, Depmster-Shafer Theory, Possibilities, Assurance Cases, Goal Structuring Notation, Elicitation.

A product to be certified follows a design, implementation, verification and validation cycle. At the beginning of the cycle, the product owner only relies, for the verification and validation aspects, on an Assurance Case (AC) template that provides choices in a tree structure. The difficulty for making decisions among choices is high when the product is based on a new technology with a large number of approaches with different levels of readiness, as it is the case for robust Machine Learning (ML). In those cases, an uncertainty assessment can be useful for making a judgment about the opportunity of using a specific approach. Based on recently published results on uncertainty elicitation and propagation in Goal Structuring Notation models of AC, the work presented here justify and implements an uncertainty assessment based simultaneously on qualitative and quantitative uncertainty modeling. Moreover, it proposes an elicitation method allowing simultaneous capture of qualitative and quantitative uncertainty and an analysis of uncertainty modeling and propagation on AC templates. Finally, it demonstrates the approach with a use case related to robustness of ML models. The result of this research will be integrated in the Capella system engineering environment.

Our partners