Uncertainty in Assurance Case Pattern for Machine Learning

Yassir Idmessaoud, Jean-Loup Farges, Eric Jenn, Vincent Mussot, Anthony Fernandes Pires, Florent Chenevier, Ramon Conejo Laguna

Uncertainty in Assurance Case Pattern for Machine Learning

Yassir Id Messaoud
IRT SystemX
Palaiseau, France
0009-0000-5673-0844

Jean-Loup Farges
IRT SystemX and ONERA
Palaiseau and Toulouse, France
0000-0002-0737-0640

Eric Jenn
IRT Saint Exupéry
Toulouse, France
0000-0001-9699-3497

Vincent Mussot
IRT SystemX and IRT Saint Exupéry
Palaiseau and Toulouse, France
0009-0001-8819-3163

Anthony Fernandes Pires
IRT SystemX and ONERA
Palaiseau and Toulouse, France
0000-0003-0522-3898

Florent Chenevier
IRT SystemX and Thales AVS
Palaiseau and Toulouse, France
florent.chenevier@irt-systemx.fr

Ramon Conejo Laguna
IRT SystemX and IRT Saint Exupéry
Palaiseau and Toulouse, France
0009-0007-9787-1716

***Abstract*—Assurance case (AC) patterns are structured arguments in a tree-like form in which certain choices are not frozen. By making these choices a user can determine a design, implementation, integration, verification and validation workflow that will produce artifacts supporting the argument for his/her use case. However, it is difficult to make choices in an AC pattern because of the lack of information on the consequences of these choices and the cost/effort they may require. Based on recently published results, this work proposes an uncertainty assessment that allows the user to be aware of the confidence in the argument induced by those choices. To do so, confidence features are elicited from experts. The elicitation procedure is presented and the propagation of uncertainty through the AC is analyzed. Finally, application of the method on a use case related to robustness of machine learning models demonstrates the validity of the approach.**

***Index Terms*—Assurance case, Dempster-Shafer theory, robustness, machine learning, experts' judgments elicitation**

I. Introduction

Functions designed using Machine Learning (ML) have to comply with standards and nowadays an effort is devoted to the proof of their dependability. Justification of such high-level properties can be done with structured arguments named Assurance Cases (AC). In order to streamline and normalize the design of AC, AC patterns are proposed. The objective of the research presented here is to add uncertainty or confidence to AC patterns. The final objective of uncertainty assessment in instances of AC is to provide to certification authorities an AC presenting a full belief assessment. However, intermediate steps with intermediate objectives are necessary because the product to be certified follows a design, implementation, integration, verification and validation cycle. At the beginning of the cycle, the product owner only relies, for all cycle steps, on an AC pattern that provides choices in a pre-defined tree structure. The difficulty for making decisions among choices is high when the subject of the AC is a new technology with a large number of 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 opportunity of using a specific approach. Moreover, the uncertainty assessment of each strategy in the tree structure may be performed at no cost and could be directly provided with the AC pattern. At the opposite, the evidence at some leaves of the tree is subject to dynamical 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 depending on the choice made in the AC pattern, the evidence must be provided independently from the design process by the verification and validation process with some cost.

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

The approach followed here is based on recently published results [1], [2] and brings the following contributions:

1) An uncertainty assessment based simultaneously on qualitative and quantitative uncertainty modeling,
2) an elicitation method allowing simultaneous capture of qualitative and quantitative uncertainty,
3) an analysis of uncertainty modeling and propagation on AC patterns and
4) demonstration of the approach with a use case related to robustness of ML models.

The following section is devoted to positioning the approach described above with respect to the state of the art. Then, a section presents the uncertainty assessment. Section IV is devoted to the elicitation process. Modeling and propagation are analyzed in section V. Section VI demonstrates the approach on the use case. Finally the conclusion provides a global assessment of the approach and possible improvements.

II. Background and related works

In this section, we introduce the necessary background information to facilitate a comprehensive understanding of our

work and we highlight the weaknesses of related works.

A. Structured arguments

1) Formalism: Goal Structuring Notation (GSN) [3] is a graphical way to describe AC including concepts such as *Goal*, *Solution*, *Strategy*, *Context*, *Assumption*, *Justification* and their relationships such as *Is supported by* and *In the context of*. Figure 1 illustrates some of these elements. Further versions of GSN include an extension allowing the description of argument patterns using the concepts of *Choice* and *Uninstantiated Element* and the description of confidence argument using the concept of *Assurance Claim Point* that refers to another argument for assessing the confidence [4]. The work presented here was conducted in the scope of GSN using another method to assess confidence. An alternative graphical way of describing AC is Claim Argument Evidence (CAE) [5]. More recently, Structured Assurance Case Meta-model (SACM) [6] was build upon GSN and CAE and transformations from these models to SACM were developed. SACM allows arguing the confidence in the arguments provided in the AC by using a meta-claim feature of the Assertion element. Meta-claim as its name suggests, is a Claim about an Assertion to argue the trustworthiness of the Assertion. The approach presented here is quite different from the SACM approach because here confidence is not modeled by additional claims but is grounded on uncertainty measures. Nevertheless, using the transformation GSN to SACM the results obtained here could be used in SACM.

2) AC for machine learning: Safety criteria, which if enforced would contribute to justifying the safety of neural networks, were determined and structured in an AC pattern presenting an undeveloped goal "The neural network tolerates faults in its inputs" [7]. The AC pattern for robustness of ML used in our work corresponds to a development of this goal.

Burton and Herd proposed a high level AC pattern for claiming that the ML system satisfy its allocated safety requirements within the defined context [8]. A strategy refines this goal in five sub-goals concerning specification, data sets, design, demonstration and operation. Only the sub-goal concerning specifications is detailed to the level of solutions. The AC pattern used in our work addressees design and demonstration and is detailed to the level of solutions on the design part.

B. Uncertainty modeling

Uncertainty is most of the time modeled using probabilities. Those are most of the time related to frequencies and are more suited for aleatory uncertainty than for epistemic uncertainty. T-norms and T-conorms are binary operations which generalize respectively conjunction and disjunction in valued logic [9]. The probabilistic T-norm corresponds to the product while the Zadeh's T-norm correspond to the minimum. Their associated T-conorms are the sum minus the product and the maximum. If $T(x, y)$ and $T^*(x, y)$ are a T-norm and its conorm, the distributivity property is characterized by $T(x, T^*(y, z)) = T^*(T(x, y), T(x, z))$ and $T^*(x, T(y, z)) = T(T^*(x, y), T^*(x, z))$ and the absorption property by $T(T^*(x, y), x) = x$ and $T^*(T(x, y), x) = x$. Finally the idempotency property is characterized by $T(x, x) = x$ and $T^*(x, x) = x$.

The Dempster-Shafer Theory (DST) [10] is a general framework for reasoning with uncertainty. It uses a frame of discernment and may allocate parts of an unitary mass on all non empty subsets of this frame of discernment. DST operations are extension to the cross product of frames of discernment, combination of masses from different sources managing the conflict issue and marginalization. Capacities [11], are set functions which give 0 for the empty set, 1 for the sure event and respect monotonicity with respect to inclusion.

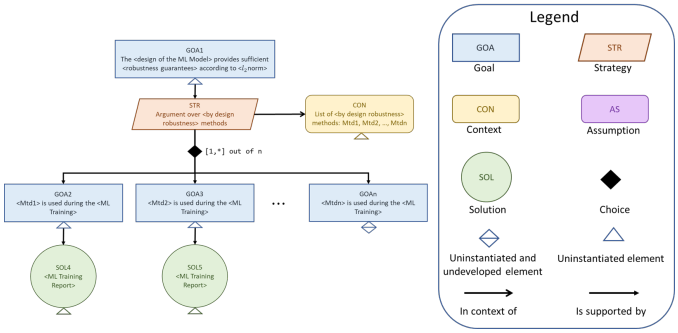

Fig. 1. An example of parts of a GSN pattern (Extract from Robustness AC pattern)

C. Uncertainty assessment in AC

1) Probabilistic approach: The question of uncertainty assessment in ACs has been the subject of a number of approaches. Some are based on *probability theory*. They use *Bayesian Networks* (BN) [12]–[15] to propagate probabilities on pieces of evidence provided by the argument up to the top-goal. Probabilities deals well with aleatory uncertainty. However, this is less the case for epistemic uncertainties due to lack of information. For this reason the work presented here is not based on BN.

2) Approaches based on DST theory: To address the issue related to BN, other approaches using DST are proposed. In addition to efficiently modeling epistemic uncertainties, these kinds of approach require less data than Bayesian approaches. First of all, those approaches assume that uncertainty is associated on the one hand to goals directly linked to solutions and on the other hand to the support relation between goals, either directly or through an explicit strategy. The other elements of GSN, such as context, assumption, justification and in the context of, provide information about uncertainty but don't carry this information. For instance, Wang et al. [16] use DST to propose models to elicit confidence values about evidence and propagate them according to the relationships between a goal and its sub-goals. The confidence on these relations is also quantified using DST. To determine their values, Wang et al. proposed to use the non linear least square method. However, this method can lead to values outside the unit interval $[0.1]$

which makes no sense. Chung-Ling et al. [17] propose to use Vector Space Model (VSM) to identify these values.

Authors in [1], [18], [19] used an approach based on experts judgment to deal with this issue. They assume that *Goals* directly supported by a *Solution* can be believed, disbelieved and epistemically uncertain, rules involved in Strategy can be believed, epistemically uncertain but cannot be disbelieved. Considered rules, $p_i \Rightarrow C$, $\neg p_i \Rightarrow \neg C$, $(\wedge_i\ p_i) \Rightarrow C$ and $(\wedge_i\ \neg p_i) \Rightarrow \neg C$, with p_i a child goal and C a father goal, provide a formal and flexible definition of Is supported by. Their corresponding belief are noted here reciprocally $B^i_{\Rightarrow}$, $B^i_{\Leftarrow}$, $B_{\Rightarrow}$ and $B_{\Leftarrow}$. Two approaches to uncertainty assessment of GSN are proposed: the quantitative approach and the qualitative approach. For the quantitative approach elicitation is performed using scales and the rankings are transformed in numbers. For the propagation child Goals with masses B^i_p, D^i_p and $1 - B^i_p - D^i_p$ on respectively itself p_i, its negation $\neg p_i$ and tautology $\top = p_i \vee \neg p_i$ lead to conclusion Goals with masses on C, $\neg C$ and $\top = C \vee \neg C$. If masses on goals linked to solutions are provided, the mass computation can be propagated from the bottom of the tree to the top of the tree and provides belief, B_C and disbelief D_C in top claim. Formulae for numeric propagation are derived from the hypotheses and the DST:

$$B_C = B_{\Rightarrow} \cdot \prod_i B^i_p(1-B^i_{\Rightarrow}) + 1 - \prod_i (1 - B^i_p \cdot B^i_{\Rightarrow}) - M_C \quad (1)$$

$$D_C = B_{\Leftarrow} \cdot \prod_i D^i_p(1-B^i_{\Leftarrow}) + 1 - \prod_i (1 - D^i_p \cdot B^i_{\Leftarrow}) - M_C \quad (2)$$

where M_C is the conflict mass on C. For its computation see [19]. For the qualitative approach elicitation is also performed using scales but there is no need to transform rankings in numbers. Formulae for qualitative propagation are derived from the hypotheses, the DST and the properties of capacities:

$$\beta_C = \max\{\min(\beta_{\Rightarrow}, \min_i \beta^i_p), \max_i \min(\beta^i_p, \beta^i_{\Rightarrow})\} \quad (3)$$

$$\delta_C = \max\{\min(\beta_{\Leftarrow}, \min_i \delta^i_p), \max_i \min(\delta^i_p, \beta^i_{\Leftarrow})\} \quad (4)$$

where β and δ are reciprocally the qualitative counterparts of B and D.

3) Criticism to propagation of uncertainty in AC: Burton and Herd indicate that these approaches depend on the availability of reliable confidence values that can be assigned to elements of the assurance argument and combined into an overall confidence score, they are themselves subject to uncertainty and subjective judgment [8]. In order to avoid this problem they propose to use locally, i.e. for each element of the AC, a first scale of uncertainty including subjective ranking, subjective probabilities, probabilities and variance combined by a second scale including ignorance, imprecise judgment, precise judgment and certainty. Those scales are quite helpful for improving locally an AC but seem inoperative for making choices in an AC pattern.

III. Qualitative and quantitative modeling shall be considered together

Requirements are proposed for uncertainty modeling and assessment: (i) The assessment shall be useful for focusing validation effort and for identifying weaknesses of AC structure, (ii) the result of the assessment of an AC tree shall not be driven by its dimension, (iii) the sensitivity of the assessment shall allow discriminating strategies and (iv) methodological choices should not be arbitrary.

A. Usefulness

The uncertainty assessment is useful for focusing validation effort on most sensitive parts of the AC because it is performed at each goal and can indicate its weakness and contradictions between proof elements. For nodes corresponding to conjunctions a procedure to focus on the most sensitive element, i.e., the one with least belief is derived. If this element corresponds to a Solution, consider means for improving its belief, for example, doing a higher number of tests. The uncertainty assessment is also useful for identifying weaknesses of AC structure and applying uncertainty reduction techniques. The proposed procedure is quite like the one for focusing validation. A Strategy associated to a node, whose uncertainty is sensitive but whose uncertainties of the children are not so, is not sufficiently convincing. Then, an alternative strategy can be considered.

B. Dimension

The result of the analysis of this requirement on a large conjunctive argument case indicates that, for the numeric approach, while the number of solutions increase the general trend is the rejection of the property corresponding to the root goal. At the opposite, for the qualitative approach the belief of the root goal cannot be lower than the belief of the 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 the qualitative approach the uncertainty of the root goal is bounded.

C. Sensitivity

Changing a strategy changes the goal supported by this strategy. This goal supports its father goal. Thus, changing a strategy changes a premise of a goal. For the numeric approach, partial derivatives of the belief and disbelief of the father goal with respect to belief and disbelief of a premise are highlighted. Thus there is a sensitivity to each premise. Concerning 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.

TABLE I
COMPLIANCE OF UNCERTAINTY MODELING WITH REQUIREMENTS

Requirement	Numeric	Qualitative
Usefulness	+	+
Result not dimension driven	–	++
Sensitivity	++	–
Not arbitrary methodological choices	+	+

D. Methodological choices

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 and 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 to linguistic qualifiers, there is consensus on their order and it is likely that two experts associate the same qualifier to the same element. However, there is gaps between the degrees of the scale and results on 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.

E. Synthesis

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

IV. ELICITING QUALITATIVE AND QUANTITATIVE UNCERTAINTY IN A SINGLE STEP

Another important result of the work is the definition of a methodology for elicitation of uncertainty associated to rules and Goals directly linked to Solutions. Following this methodology, the full tree is presented without Strategies to experts, i.e. child Goals are directly connected to father Goals by a Is supported by relation. Then a questionnaire with a form for each hidden Strategy has to be filled by experts. Figure 3 presents an extract of the form for a Goal supported by two children 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.

Answers are given by experts associating a confidence in decision on the scale { very low, low, high, very high }. For positive rules the provided decision is the acceptance of the father Goal. For negative rules it is the rejection of the father Goal. In both cases the strength of the decision is scaled on {no decision, weak, moderate, strong }. Numerical values are captured using a scroll-bar that drives the linguistic qualifier of the corresponding scale.

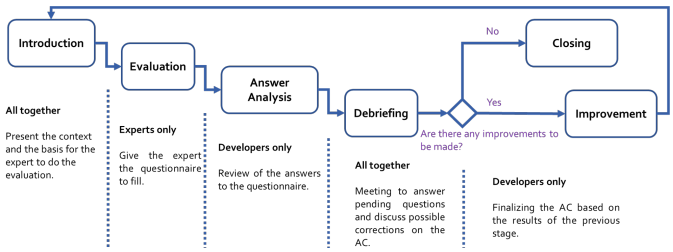

Fig. 2. Assurance case assessment process

Answers are converted to masses on belief and tautology for each rule. The quantitative approach considers the values provided by scroll-bars. The qualitative approach uses the semantic qualifiers.

It is important to know that the elicitation phase may require several round of assessment by experts. Normally during its elaboration, an AC is subject to an internal reviewing. However, Rushby et al. [20] explained that this kind of evaluation is not only insufficient, but also not very effective. This is because developers tend to justify their reasoning rather than question it, while external assessors will most likely try to criticise it. Analysis of the elicitation results provided by the external experts allowed us to improve the structure of the argument (i.e., reasoning and evidence). Hence, the necessity of reassessing the argument after each major modification, until we get a structure approved by a reasonable number of experts. Figure 2 shows this process. The answers collected during the closing phase are those that will be used to propagate confidence and uncertainty measures to the top-goal.

As shown Figure 2, the first stage after the selection of external expert(s) is to introduce the GSN standard if required, present the assurance case, and the assessment procedure. I.e., how to interpret and answer the questions in the form. Once the form is filled, answers (i.e., direct responses to questions in Figure 3 for example, and comments left by the expert(s)) are analyzed in order to detect any inconsistency or misunderstanding. A debriefing session is then scheduled to answers pending questions and discuss possible corrections on the AC. If improvements are required, the AC is modified and reassessed by a different set of experts. If this is not the case, the confidence/uncertainty measures resulting from these responses are associated to the AC so that they can be used during the propagation step.

The choice of an expert depends on his/her knowledge and competence in the fields covered by the AC. (e.g., ML, formal proof, V&V processes, etc.). Ideally the expert/assessor 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 call a set of experts. Aggregating their answers can be done through discussion by agreeing on a

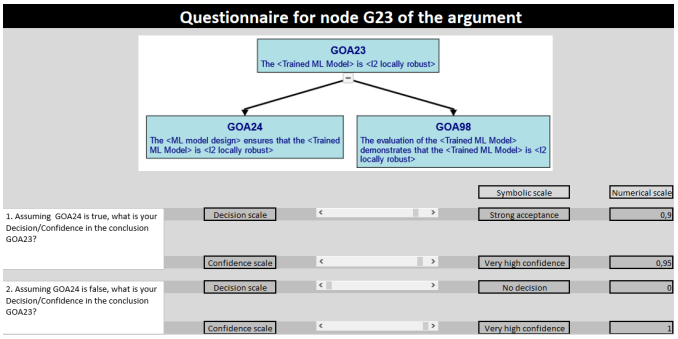

Fig. 3. Extract from the form used for elicitation

single answer, which can be difficult and time-consuming. It can also be computed using aggregation formulas. This issue is not addressed in this paper since the evaluation of "Robustness AC" was made by a single expert.

V. Analysis of uncertainty modeling and propagation in AC patterns

Results indicate that conflicts, as meant by DST, cannot be detected at single rule level because for rules mass is only on tautology $\top = r \vee \neg r$ and the rule itself r. However variation of mass between experts can be recorded. Moreover, the results indicate that 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 those results indicate that conflicts between experts are not detectable without applying the AC to a use case.

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 $1 - \epsilon$ 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.

VI. Uncertainty in the AC pattern for robustness of ML

A. AC pattern for robustness of ML

The root goal of the AC pattern for robustness of ML, i.e. goal 15 in Table II, 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

TABLE II
Goals supported by Strategies

Goal number	Wording
15	<The Trained ML model> is <robust>
17	<The Trained ML model> satisfies the <global robustness criteria>
18	<The Trained ML model> satisfies the <Global nbsample robustness criteria>
21	<The Trained ML model> is <locally robust>
23	<The Trained ML model> is $< l_2$ locally robust>
24	<The ML model design> ensures that <The Trained ML model> is $< l_2$ locally robust>
25	<The ML model design> integrates applicable <robustness reinforcement methods> and these methods allows that <The Trained ML model> is $< l_2$ locally robust>
99	<The Trained ML model> is $< l_\infty$ locally robust>
100	<The ML model design> ensures that <The Trained ML model> is $< l_\infty$ locally robust>
101	<The ML model design> integrates applicable <robustness reinforcement methods> and these methods allows that <The Trained ML model> is $< l_\infty$ locally robust>

and "<robust>" is a property defined in the AC. This goal is reformulated and then decomposed in three sub-goals, all based on 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 obtained, either by design or by validation. The tree further develop the branch dedicated to *by design* methods, splitting in sub-goals corresponding to families of methods of this category. Tables II and III present some stages of this decomposition. Note that 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 $< l_2$ locally robust>" and "The evaluation of the <Trained ML Model> demonstrates that the <Trained ML Model> is $< l_\infty$ locally robust>". Finally, as shown on Figure 4, the goal corresponding to each method is supported by a set of three goals: two which are dependent on artifacts linked to the trained ML model, and one connected to a solution referencing published research articles, cf. Table IV. The goals connected to solutions for goals 30, 42, 55, 76, 103, 126, 139, 150, and 165 are respectively goals 36, 48, 60, 82, 108, 132, 143, 154, and 168.

This structure is a pattern AC and not an AC because artifacts are not present and branches of the tree can be deleted for a specific ML model.

B. Elicitation results

One expert filled forms of the type of the one shown in Figure 3, for goals connected to articles and for nodes upper in the tree. The results are derived by gathering and analyzing the filled forms. It consists in filling the AC pattern from expert's answers.

1) Qualitative analysis: The analysis of answers to open questions and binary questions highlights the following points.

a) Too demanding expert effort: The expert indicated that he didn't analyze articles related to goals 108, 154 and 168,

TABLE III
SUPPORTING GOALS FOR GOALS SUPPORTED BY STRATEGIES

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

TABLE IV
GOALS SUPPORTED BY SCIENTIFIC ARTICLES

Goal number	Norm	Solution
36	l_2	Jacobian regularization [21]
48	l_2	Lipschitz training [22]
60	l_2	Certified robust training [23]–[25]
82	l_2	Randomized smoothing [26]–[28]
108	l_∞	Empirical robustness reinforcement [21], [29]–[32]
132	l_∞	Lipschitz training [22]
143	l_∞	Gowal certified robust training [33]
154	l_∞	Certified robust training [34], [35]
168	l_∞	Random Noising [27], [36]

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 goal/solution(s) nodes, which require the reading and processing of extensive documentation (e.g., technical reports, scientific articles, etc.).

b) Definitions: Concerning the definition of robustness, the expert indicated that the definition of robust provided by the AC is restrictive. For instance, this definition don't cover robustness with respect to distribution shift. 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 IV. 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.

c) Contexts: For the context associated to goal 101, the expert has some doubts about the applicability for l_∞ 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 l_∞ robustness

d) Relations between goal and sub-goals: The expert signaled that, for a given perturbation radius, goal 99 implies goal 23 because the l_2 ball is included in the l_∞ 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". 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) Relations between goal and 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. Some articles are subject to a deep analysis by the expert. For Jacobian regularization [21] the expert concludes that it improves l_2 robustness but doesn't ensure it. For Lipschitz training [22] he indicates that a specific loss function should be used as done in recent work [37]. For Certified robust training for l_2 robustness, the expert indicates that one article [24] is out of scope

2) Quantitative analysis:

a) 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 lx 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 4. Without a concrete use case with a specific ML model, it is not possible to assess GOA2 and GOA4 in this figure. Thus goals 30 and 36, 42 and 48... are linked for uncertainty propagation by simple arguments 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. Thus, the structure of goal 18 is changed to a simple argument with sub goal 21. Uncertainty of rules for

TABLE V
UNCERTAINTY ASSOCIATED TO RULES

Goal	Sub-goal(s)	Direct belief		Inverse belief	
		quantitative	qualitative	quantitative	qualitative
15	17	1.000	VH	1.000	VH
17	18	1.000	VH	0.915	VH
18	21	0.765	VH	0.845	VH
21	23	0.860	VH	1.000	VH
21	99	0.860	VH	1.000	VH
21	all	1.000	VH	1.000	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
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

this simple argument is derived from the answers in the form to questions concerning goal 21 alone.

b) Elicitation problems: Analysis of answer to elementary questions indicate that the expert takes sometime a decision that is excessive with respect to its confidence leading to a disrespect of Josang constraint. Moreover some inconsistency between elementary and conjunctive rules is observed. Finally, some strategies that were considered by the AC developers as pure rewording or as pure logical operators are assessed differently by the expert when the goal and sub-goals are presented without explaining the strategy, indicating that the wording of goals should be revised. This has been particularly critical for the node 21, that is a choice of a norm and that was interpreted by the expert as a competition between norms. All those elicitation problems were solved during the debriefing with the expert.

c) Elicitation of uncertainty associated to rules: Table V presents the uncertainty associated to rules after correcting the elicitation problems. In this table VH, H and L stand reciprocally for Very High, High and Low.

d) Elicitation of uncertainty for goals associated with solutions: Table IV presents the uncertainty associated to goals directly supported by solutions. In the table VL stands for Very Low. The table indicates that at the leafs of the tree the expert is confident of using Lipschitz training when considering robustness criteria based on l_2 norm, goal 48, and less confident when considering robustness criteria based on l_∞ norm, goal 132. For all other methods the belief is too low and sometime the disbelief is larger than the belief.

TABLE VI
UNCERTAINTY ASSOCIATED TO GOALS LINKED TO SOLUTIONS

Goal	Belief		Disbelief	
	quantitative	qualitative	quantitative	qualitative
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
183	0.270	L	0.110	L

TABLE VII
UNCERTAINTY PROPAGATION FOR LIPSCHITZ TRAINING BASED ON l_2 NORM

Goal	Belief		Disbelief	
	quantitative	qualitative	quantitative	qualitative
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	0.618	VH	0.014	L

C. Propagation results

Results on the use of the AC pattern by a ML model developer are obtained by making uncertainty propagation under different hypotheses for the solution directly linked to artifacts. Assuming that Lipschitz training is applicable to a specific ML model, confidence can be propagated higher in the AC assuming that goals 98 and 178 related to verification will be fulfilled with very high belief and no disbelief. A user of the AC pattern will then make a propagation up to the choice between l_2 and l_∞ norms and use the propagation results to make the choice.

1) Propagation to the choice of a norm:

a) l_2 norm: As shown in Table VII for Lipschitz training considering a robustness criteria based on l_2, the argumentation improves its initial strength, i.e. belief of 0.6 qualified as very high, because of the confidence brought by the validation, c.f. goal 23.

b) l_∞ norm: As shown by Table VIII, for a robustness criteria based on l_∞ the initial belief of 0.375, qualified as low, is also improved by the hypothesis of successful validation.

Note that for both training methods all depends on the presence of successful validation. Moreover, disbelief is reduced by propagation in the AC and reaches 0.014 and 0.003, both

TABLE VIII
UNCERTAINTY PROPAGATION FOR LIPSCHITZ TRAINING BASED ON l_∞ NORM

Goal	Belief		Disbelief	
	quantitative	qualitative	quantitative	qualitative
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 IX
POST CHOICE PROPAGATION

Goal	Belief		Disbelief	
	quantitative	qualitative	quantitative	qualitative
18	0.473	VH	0.012	L
17	0.473	VH	0.011	L
15	0.473	VH	0.011	L

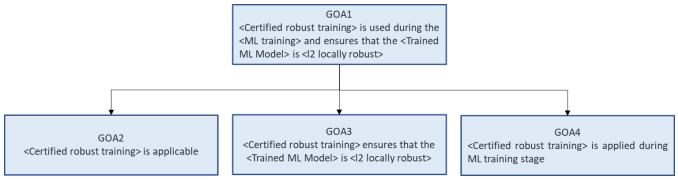

Fig. 4. Example of an argument to be adjusted for uncertainty assessment

qualified as low, for reciprocally l_2 and l_∞ Lipschitz training. Finally, all depends on the presence of a successful validation.

2) Choice of a norm: The user of the AC pattern has to make a choice on the basis of at least four criteria: quantitative and qualitative belief to maximize, quantitative and qualitative disbelief to minimize and other criteria such as cost of artifact production to minimize. Considering only the four criteria, using a ranking with Leximin the values for l_2 are (0.618, 2/3, 0.986, 1) and the values for l_∞ are (1/3, 0.430, 2/3, 0.997). 0.618 being larger than 1/3, the l_2 norm is chosen.

3) Post choice propagation: The Table IX presents the propagation from the choice to the top property of the AC. Quantitatively there is some decrease of belief at goal 18 due to a possible difference between local robustness and global robustness.

D. Lessons learned

1) Strategies shall be shown: 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.

2) Consistency shall be enforced: 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.

3) GSN format shall be adjusted for uncertainty assessment: So far, there is no systematic method to design an assurance case using GSN formalism. Moreover, uncertainty assessment procedures proposed in the state of the art are not mature enough to consider their features during the development of an AC. For example, in the literature, one can find an argument that presents a goal with a method to verify a property (defined as a top goal) and another goal that argues that this method ensures the property. However, "Is supported by" arrows, formally define by rules, already fulfill this role. I.e., saying that the application of a method m supports a property p means, according to the nature of the chosen strategy, that m ensures, demonstrates, implies, etc. p. Thus, a goal carrying the inference between a method and the property it supports, must not be considered during the uncertainty assessment. For example, questions about goal GOA3 ("`<Certified robust training> ensures that the <Trained ML Model> is` $< l_2$ `locally robust >`"), in Figure 4, will not be included in the form. In addition, solutions are either used as a reference to an artifact (e.g., a formal verification report, test results, etc.) or to a method to be applied. Remember that the assessment approach describe in this paper does not assess the inference between the solution and the goal connected to it. However, in the second case the assessment of inference is needed. To keep coherence in the approach all solutions that carry a method are transformed to goals. The artifacts resulting from the application of these methods, such as reports results, can serve as new solutions.

4) Multi criteria choice methods shall be integrated: 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. For most nodes, propagation of those four elements 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 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. Moreover, there is no total order between goals assessed following different strategies because there are four uncertainty elements and other elements such as, for instance, the cost. Thus, a multi criteria reasoning shall be performed for choosing the best solution.

5) The AC pattern shall be extended and consolidated: 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.

In consequence new AC patterns are derived: One pattern is developed for each norm and each robustness definition. Figure 5 presents the structure of the pattern devoted to the number of samples robustness criterion.

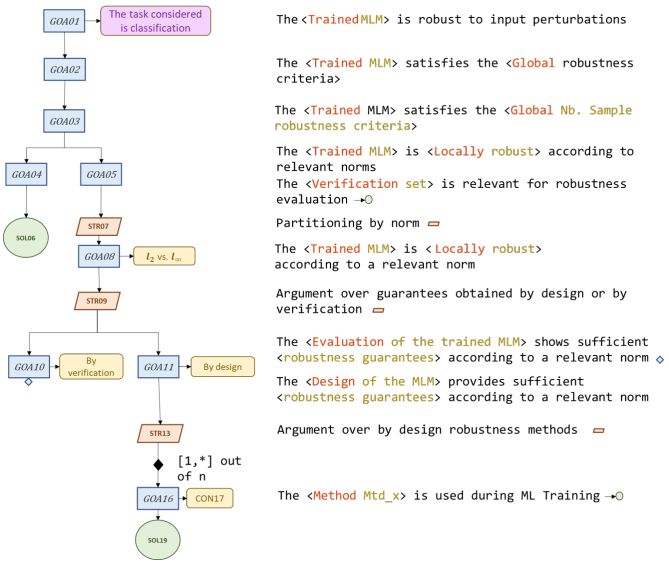

Fig. 5. Updated AC pattern

VII. Conclusion

The work presented here shows that recently proposed methods [19] can be applied to large AC patterns. However, the elicitation of masses requires a large number of questions to experts. Fortunately, the results obtained indicate that large AC don't imply large uncertainty on conclusion. The work also shows that it is useful to work with both scale and numbers and that the uncertainty in AC patterns contributes to performing design, implementation, integration, verification and validation choices and improving the AC structure. Finally, the result of this research will be integrated in the Capella system engineering environment[1].

An open question for future researches is the use of uncertainty levels in the context of certification and a possible link between the qualitative belief and disbelief of the top goal of a final AC with Safety Integrity Levels or Design Assurance Levels (DAL). For instance a belief VH and a disbelief VL could be requested for DAL A and B, a belief VH and a disbelief at most L for DAL C and D and a belief VH and a disbelief at most H for DAL E.

[1]https://eclipse.dev/capella/

Acknowledgment

This work has been supported by the French government under the "France 2030" program, as part of the SystemX Technological Research Institute.

References

[1] Y. Idmessaoud, D. Dubois, and J. Guiochet, "Uncertainty elicitation and propagation in gsn models of assurance cases," in *International Conference on Computer Safety, Reliability, and Security*. Springer, 2022, pp. 111–125.

[2] ——, "Confidence assessment in safety argument structure - quantitative vs. qualitative approaches," *International Journal of Approximate Reasoning*, vol. 165, p. 109100, 2024.

[3] T. Kelly and R. Weaver, "The goal structuring notation–a safety argument notation," in *Proceedings of the dependable systems and networks 2004 workshop on assurance cases*, vol. 6. Citeseer Princeton, NJ, 2004.

[4] Assurance-Case-Working-Group *et al.*, "Goal structuring notation community standard (version 3)," 2021. [Online]. Available: https://scsc.uk/r141C:1?t=1

[5] R. Bloomfield and K. Netkachova, "Building blocks for assurance cases," in *2014 IEEE International Symposium on Software Reliability Engineering Workshops*. IEEE, 2014, pp. 186–191.

[6] R. Wei, T. P. Kelly, X. Dai, S. Zhao, and R. Hawkins, "Model based system assurance using the structured assurance case metamodel," *Journal of Systems and Software*, vol. 154, pp. 211–233, 2019.

[7] Z. Kurd and T. Kelly, "Establishing safety criteria for artificial neural networks," in *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*. Springer, 2003, pp. 163–169.

[8] S. Burton and B. Herd, "Addressing uncertainty in the safety assurance of machine-learning," *Frontiers in Computer Science*, vol. 5, p. 1132580, 2023.

[9] M. M. Gupta and J. Qi, "Theory of t-norms and fuzzy inference methods," *Fuzzy sets and systems*, vol. 40, no. 3, pp. 431–450, 1991.

[10] G. Shafer, *A mathematical theory of evidence*. Princeton university press, 1976, vol. 42.

[11] D. Dubois, F. Faux, H. Prade, and A. Rico, "Qualitative capacities: Basic notions and potential applications," *International Journal of Approximate Reasoning*, vol. 148, pp. 253–290, 2022.

[12] E. Denney, G. Pai, and I. Habli, "Towards measurement of confidence in safety cases," in *2011 International Symposium on Empirical Software Engineering and Measurement*. IEEE, 2011, pp. 380–383.

[13] D. Nešić, M. Nyberg, and B. Gallina, "A probabilistic model of belief in safety cases," *Safety science*, vol. 138, p. 105187, 2021.

[14] C. Hobbs and M. Lloyd, "The application of bayesian belief networks to assurance case preparation," in *Achieving Systems Safety: Proceedings of the Twentieth Safety-Critical Systems Symposium, Bristol, UK, 7-9th February 2012*. Springer, 2011, pp. 159–176.

[15] J. Guiochet, Q. A. Do Hoang, and M. Kaaniche, "A model for safety case confidence assessment," in *Computer Safety, Reliability, and Security: 34th International Conference, SAFECOMP 2015, Delft, The Netherlands, September 23-25, 2015, Proceedings 34*. Springer, 2015, pp. 313–327.

[16] R. Wang, J. Guiochet, G. Motet, and W. Schön, "Modelling confidence in railway safety case," *Safety Science*, vol. 110, pp. 286–299, 2018.

[17] C.-L. Lin, W. Shen, S. Drager, and B. Cheng, "Measure confidence of assurance cases in safety-critical domains," in *Proceedings of the 40th International Conference on Software Engineering: New Ideas and Emerging Results*, 2018, pp. 13–16.

[18] Y. Idmessaoud, D. Dubois, and J. Guiochet, "A qualitative counterpart of belief functions with application to uncertainty propagation in safety cases," in *International Conference on Belief Functions*. Springer, 2022, pp. 231–241.

[19] Y. Idmessaoud, "Uncertainty assessment in safety argument structures– an approach based on dempster-shafer theory," Ph.D. dissertation, UPS Toulouse, 2022.

[20] J. Rushby, X. Xu, M. Rangarajan, and T. L. Weaver, "Understanding and evaluating assurance cases," Tech. Rep., 2015.

[21] D. Jakubovitz and R. Giryes, "Improving dnn robustness to adversarial attacks using jacobian regularization," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 514–529.

[22] C. Anil, J. Lucas, and R. Grosse, "Sorting out lipschitz function approximation," in *International Conference on Machine Learning*. PMLR, 2019, pp. 291–301.

[23] E. Wong, F. Schmidt, J. H. Metzen, and J. Z. Kolter, "Scaling provable adversarial defenses," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[24] X. Huang, M. Kwiatkowska, S. Wang, and M. Wu, "Safety verification of deep neural networks," in *Computer Aided Verification: 29th International Conference, CAV 2017, Heidelberg, Germany, July 24-28, 2017, Proceedings, Part I 30*. Springer, 2017, pp. 3–29.

[25] M. Hein and M. Andriushchenko, "Formal guarantees on the robustness of a classifier against adversarial manipulation," *Advances in neural information processing systems*, vol. 30, 2017.

[26] J. Cohen, E. Rosenfeld, and Z. Kolter, "Certified adversarial robustness via randomized smoothing," in *international conference on machine learning*. PMLR, 2019, pp. 1310–1320.

[27] H. Hong, B. Wang, and Y. Hong, "Unicr: Universally approximated certified robustness via randomized smoothing," in *European Conference on Computer Vision*. Springer, 2022, pp. 86–103.

[28] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "Certified robustness to adversarial examples with differential privacy," in *2019 IEEE symposium on security and privacy (SP)*. IEEE, 2019, pp. 656–672.

[29] A. Ross and F. Doshi-Velez, "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, 2018.

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

[31] A. Nayebi and S. Ganguli, "Biologically inspired protection of deep networks from adversarial attacks," *arXiv preprint arXiv:1703.09202*, 2017.

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

[33] S. Gowal, K. Dvijotham, R. Stanforth, R. Bunel, C. Qin, J. Uesato, R. Arandjelovic, T. Mann, and P. Kohli, "On the effectiveness of interval bound propagation for training verifiably robust models," *arXiv preprint arXiv:1810.12715*, 2018.

[34] E. Wong and Z. Kolter, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *International conference on machine learning*. PMLR, 2018, pp. 5286–5295.

[35] H. Zhang, H. Chen, C. Xiao, S. Gowal, R. Stanforth, B. Li, D. Boning, and C.-J. Hsieh, "Towards stable and efficient training of verifiably robust neural networks," *arXiv preprint arXiv:1906.06316*, 2019.

[36] M. Lecuyer, V. Atlidakis, R. Geambasu, D. Hsu, and S. Jana, "On the connection between differential privacy and adversarial robustness in machine learning," *stat*, vol. 1050, p. 9, 2018.

[37] M. Serrurier, F. Mamalet, A. González-Sanz, T. Boissin, J.-M. Loubes, and E. Del Barrio, "Achieving robustness in classification using optimal transport with hinge regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 505–514.