
Proceedings Paper:
Herd, Benjamin, Burton, Simon orcid.org/0000-0001-9040-8752 and Zacchi, João Vitor (2024) A Deductive Approach to Safety Assurance:Formalising Safety Contracts with Subjective Logic. In: Ceccarelli, Andrea, Bondavalli, Andrea, Trapp, Mario, Schoitsch, Erwin, Gallina, Barbara and Bitsch, Friedemann, (eds.) Computer Safety, Reliability, and Security. SAFECOMP 2024 Workshops - DECSoS, SASSUR, TOASTS, and WAISE, Proceedings. 19th Workshop on Dependable Smart Embedded and Cyber-Physical Systems and Systems-of-Systems, DECSoS 2024, 11th International Workshop on Next Generation of System Assurance Approaches for Critical Systems, SASSUR 2024, Towards A Safer Systems architectur, 17 Sep 2024 Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics) . Springer Nature Switzerland , ITA , pp. 213-226.

https://doi.org/10.1007/978-3-031-68738-9_16

A deductive approach to safety assurance: Formalising safety contracts with Subjective Logic

Benjamin Herd[1][0000−0001−6439−8845], João-Vitor Zacchi[1][0009−0002−6975−6829], and Simon Burton[2][0000−0001−9040−8752]

[1] Fraunhofer Institute for Cognitive Systems (IKS), Munich, Germany
benjamin.herd@iks.fraunhofer.de
[2] University of York, York, United Kingdom
simon.burton@york.ac.uk

Abstract. The increasing adoption of autonomous systems in safety-critical applications raises severe concerns regarding safety and reliability. Due to the distinctive characteristics of these systems, conventional approaches to safety assurance are not directly transferable and novel approaches are required. One of the main challenges is the ability to deal with significant uncertainty resulting from (1) the inherent complexity of autonomous system models, (2) potential insufficiencies of data and/or rules, and (3) the open nature of the operational environment. The validity of assumptions made about these three layers greatly impact the confidence in the guarantees provided by a safety argument. In this paper we view the problem of safety assurance as the satisfaction of a safety contract, more specifically as a conditional deduction operation from assumptions to guarantees. We formalise this idea using Subjective Logic and derive from this formalisation an argument structure in GSN that allows for automated reasoning about the uncertainty in the guarantees given the assumptions and any further available evidence. We illustrate the idea using a simple ML-based traffic sign classification example.

Keywords: safety assurance · uncertainty · autonomous systems.

1 Introduction

In recent years, the rapid advancement of artificial intelligence (AI) techniques has led to an increasing adoption in safety-critical applications. One notable area where AI has proven highly useful is perception in (semi-)autonomous driving. While these advancements have brought significant improvements in performance, they also raise concerns regarding the safety and reliability of such systems. Ensuring a sufficient level of safety in these components has become a critical challenge that necessitates rigorous argumentation and assessment.

Inherent challenges of autonomous systems such as complex and open operating environments that render formal specification infeasible introduce a substantial level of uncertainty into the safety argumentation process. Unlike traditional systems, autonomous systems often lack explicit rules and rely on advanced algorithms and decision-making processes, which can lead to unexpected behavior

and novel failure modes. Safety arguments for autonomous systems must address this inherent uncertainty and provide assurance with respect to the residual risk associated with their operation. Furthermore, modern control systems are increasingly complex, require the continuous integration of new or updated components, and thus a modular and contract-based approach to safety assurance.

To structure the assurance argument for an autonomous safety-critical function, existing standards such as ISO 21448[3] (SOTIF) and the upcoming standard ISO PAS 8800[4] serve as a useful starting point. SOTIF provides a framework for assessing and managing hazards that arise from a gap between the specified and the intended functionality or from foreseeable misuse of a system. By incorporating SOTIF principles, safety arguments can address not only the known failure scenarios but also potential unforeseen risks and hazards that emerge from the complex interactions between the autonomous system and its environment. In reality, an assurance argument will typically include a mix of quantitative evidences (including constructive measures, formal analysis and testing), but also qualitative arguments. Thus, the level of residual risk that has been addressed is not always clear and often relies on expert judgment and established reasoning methods as prescribed in safety standards. Currently, there is no industry or research community consensus on which set of methods are sufficient for evaluating the performance of functions in a safety-critical context. This poses challenges for assurance argumentation, as the validity of the evidences themselves can be called into question [2]. It is thus important to acknowledge that the quality and effectiveness of the resulting argumentation requires deep scrutiny and refinement.

With this paper, we make a contribution to this rapidly advancing field of research by focussing on the problem of contract-based uncertainty quantification in safety arguments. In a recent paper, we argued for the application of Subjective Logic (SL) [12] for analysing uncertainty in the context of safety assurance [10]. Here, we expand upon this idea by (1) formulating the safety assurance problem as a conditional deduction from premises (assumptions) to conclusions (safety guarantees), and (2) deriving the structure of an assurance argument that directly corresponds to the formalisation and allows for automated reasoning about assurance confidence. More specifically:

- We formalise the notion of a safety contract introduced elsewhere [2] as a conditional deduction operation in SL.
- From this formalisation, we derive the structure of an assurance argument including sub-arguments about both the sufficiency of the assumptions and the resilience of the system in case of assumption violation.
- We provide a methodology to estimate assurance confidence in such an argument by (1) populating the argument with quantitative and qualitative evidence, and (2) applying the SL conditional deduction operator to derive an opinion about the top-level claim. The methodology is illustrated using a simple traffic sign classification example.

[3] `https://www.iso.org/standard/77490.html`
[4] `https://www.iso.org/standard/83303.html`

The paper is structured as follows. Section 2 provides a brief introduction to the problem of safety assurance for autonomous systems and to the formalism of SL. Section 3 introduces a formally derived argument structure that allows for automated reasoning, followed by the description and illustration of a methodology to reason about an argument using a simple example in Section 4. The paper concludes with an overview of related work and a summary.

2 Background

2.1 Safety assurance

The ISO/IEC/IEEE 15026:2019[5] standard defines *assurance* as grounds for justified confidence that a *claim* has been or will be achieved. A *claim* is defined as a true-false statement about the limitations on the values of an unambiguously defined property — called the claim's property — and limitations on the uncertainty of the property's values falling within these limitations. The standard also defines an *assurance argument* as a reasoned, auditable artefact that supports the contention that its top-level claim is satisfied, including systematic arguments and its underlying evidence and explicit assumptions that support the claim(s). As such, the assurance argument communicates the relationship between evidence and the safety objectives. A model-based graphical representation of the assurance argument can aid its communication and evaluation. Within this paper we make use of the Goal Structuring Notation (GSN)[6] to visualise the assurance argument. An example of a simple GSN structure is shown in Figure 1. It consists of a top-level *claim* **G1** that is to be shown to hold given *assumptions* **A1** and *context* information **C1** which defines or constrains the scope over which claim **G1** is made. **G1** is further broken down using *strategy* **S1** which describes the inferential step between **G1** and sub-goals **G2** and **G3**. The sub-goals are then substantiated by *solutions* **Sn1** and **Sn2**, i.e. by concrete evidential artifacts.

The standard ISO 21448 (SOTIF) addresses safety in terms of the absence of unreasonable risk due to functional insufficiencies of the system or by reasonably foreseeable misuse. The SOTIF approach considers hazards that are caused by latent insufficiencies of the function that are uncovered by *triggering conditions* in the operating environment at runtime. The SOTIF model describes the task of risk reduction as maximising the number of triggering conditions that are known to potentially lead to hazardous behaviour (*known unknowns*) such that they can be made safe whilst minimising the number of potentially hazardous residual unknown triggering conditions (*unknown unknowns*). In the context of AI, *known* triggering conditions could be considered as inputs that are known to reveal an insufficiency in the trained model, whilst *unknown* triggering conditions relate to inputs that were not considered within the training and test set, e.g. due to features considered irrelevant or distributional shift in the environment.

[5] `https://www.iso.org/standard/73567.html`
[6] `https://scsc.uk/GSN`

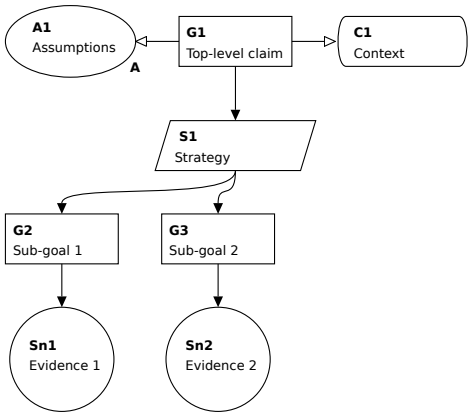

Fig. 1. Example of a simple assurance argument structure in GSN

SOTIF appears well suited as a basis for discussing the safety of AI functions where hazardous behaviours are caused by inaccuracies in the trained model itself rather than by faults during its execution.

Burton *et al.* [2] express the task of assuring the safety of machine learning (ML) according to the SOTIF model in terms of demonstrating the fulfillment of a safety contract based on the following definition.

$$\forall i \in I.A(i) \Rightarrow G(i, M(i)) \tag{1}$$

Where, for all inputs i that fulfil the set of assumptions A on the operating domain and system context, the output of model M must fulfill a set of conditions defined by guarantees G. For realistic applications, residual errors in the model will inevitably remain. Assurance thus involves demonstrating that the probability with which this contract is fulfilled is in accordance to the risk acceptance criteria. This formulation of a safety contract will form the basis for the argument structure introduced in Section 3.

2.2 Subjective Logic

Subjective Logic (SL) [12] is a framework for artificial reasoning with uncertain beliefs that combines ideas from probabilistic logic and evidence theory. The atomic building blocks of SL are *subjective opinions* and SL offers a wide range of combination operators that allow for algebraic reasoning. Subjective opinions in SL express beliefs about the truth of propositions under degrees of uncertainty. Opinions can be *binomial* ($X = \{x, \bar{x}\}$), *multinomial* (X of cardinality > 2 with singleton elements only), or *hypernomial* (X of cardinality > 2 with elements $x \in \mathcal{R}(X)$[7]). Since we aim to model binary safety claims as explained in Section 2.1, we restrict the focus to binomial opinions here.

[7] $\mathcal{R}(X)$) denotes the reduced powerset of X, i.e., the set of all subsets excluding the empty set $\emptyset$ and the full set X.

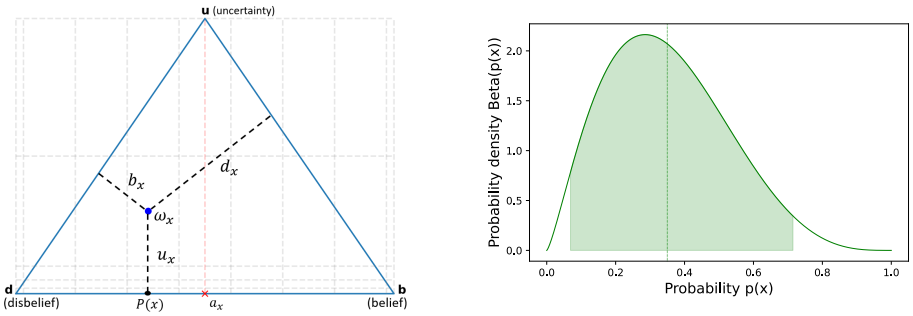

Fig. 2. Opinion triangle visualization of example opinion $\omega_x = (b = 0.2, d = 0.5, u = 0.3, a = 0.5)$ (left) and associated Beta PDF.

Definition 1 (Binomial opinion). *Let $X = \{x, \bar{x}\}$ be a binary domain. A binomial opinion about the truth of X is a tuple $\omega_X = (b, d, u, a)$ where*

- *b (belief): the belief mass in support of x being **true***
- *d (disbelief): the belief mass in support of x being **false***
- *u (uncertainty): the uncommitted belief mass*
- *a (base rate): the a priori probability in the absence of committed belief mass (often set to 0.5 for binary domains).*

The components have to satisfy $b, d, u, a \in [0, 1]$ and $b + d + u = 1$.

Binomial opinions can be visualised on an equilateral triangle as shown for the example opinion $\omega_x = (0.2, 0.5, 0.3)$ in Figure 2 (left). Any point inside the triangle represents a possible (b, d, u) triple.

Given positive evidence r (i.e. the number of positive observations) for a claim, negative evidence s (the number of negative observations) and a *non-informative prior weight*[8] $W = 2$, an opinion can be computed as follows:

$$b_X = r/(r + s + W) \tag{2}$$
$$d_X = s/(r + s + W) \tag{3}$$
$$u_X = W/(r + s + W) \tag{4}$$

Given domain $X = \{x, \bar{x}\}$, let $p : X \to [0, 1]$ be a continuous probability distribution over the same domain where $p(x) + p(\bar{x}) = 1$. A binomial opinion corresponds with a Beta probability density function (PDF) $Beta(p(x), \alpha, \beta)$

[8] The non-informative prior weight W ensures that when evidence begins to accumulate (i.e. r gets larger), uncertainty u_X decreases accordingly. W is typically set to the same value as the cardinality of the domain (2 in our binary case), thus artificially adding one "success" r and one "failure" s. Higher values of r and s require more evidence for uncertainty to decrease.

over variable $p(x)$ as shown graphically in Figure 2 (right). The α and β parameters of the PDF can be derived from the base rate a, the observation evidence r and s, and the non-informative prior weight W as follows:

$$\alpha = r + aW, \qquad \beta = s + (1 - a)W \tag{5}$$

The expectation value $E(X)$ of the Beta PDF is then derived as follows:

$$E(X) = \frac{\alpha}{\alpha + \beta} = \frac{r + aW}{r + s + W} \tag{6}$$

SL provides a wide range of combination operators [12] which provide an elegant and intuitive way to combine opinions directly instead of the underlying Beta distributions, a direct manipulation of which would be mathematically challenging. In this paper, we focus on *conditional deduction*, a realisation of Modus Ponens in SL, represented by the ternary '$\circledcirc$' operator[9].

Definition 2 (Conditional deduction). *Let $X = \{x, \bar{x}\}$ and $Y = \{y, \bar{y}\}$ be two binary domains with arbitrary mutual dependence. Let $\omega_x = (b_x, d_x, u_x, a_x)$, $\omega_{y|x} = (b_{y|x}, d_{y|x}, u_{y|x}, a_{y|x})$ and $\omega_{y|\bar{x}} = (b_{y|\bar{x}}, d_{y|\bar{x}}, u_{y|\bar{x}}, a_{y|\bar{x}})$ an agent's respective opinions about x being true, about y being true given that x is true, and about y being true given that x is false. Based on that,*

$$\omega_{y||x} = \omega_x \circledcirc (\omega_{y|x}, \omega_{y|\bar{x}}) \tag{7}$$

denotes a conditionally deduced opinion *derived from ω_x, $\omega_{y|x}$, and $\omega_{y|\bar{x}}$. It expresses the belief in y (the conclusion) being true as a function of the belief in x (the premise) and the two sub-conditionals $y|x$ and $y|\bar{x}$. It can be understood as a direct translation of the deduced probability $P(y||x)$ into SL which can be defined using the law of total probability as follows: $P(y||x) = P(x)P(y|x) + P(\bar{x})P(y|\bar{x})$.*

3 A formally grounded assurance argument structure

The conditional deduction operator of SL provides us with a basis to derive an argument structure that allows for formal deduction as follows. Referring to the constituents of a formal deduction $\omega_{y||x} = \omega_x \circledcirc (\omega_{y|x}, \omega_{y|\bar{x}})$ as per Equation 7:

- ω_x represents an opinion about the premise of the deduction. It thus expresses belief in the *assumptions* in the safety contract in (1) from which the guarantees are to be deduced.
- $\omega_{y|x}$ and $\omega_{y|\bar{x}}$ represent sub-conditionals about the *sufficiency* and the *necessity* of the premise, respectively [15]. As such, the former can be seen as an opinion about the claim that the guarantees are satisfied *given that the assumptions are satisfied*. The latter can be understood as dealing with the (invariably) existing uncertainty in the assumptions arising from unforeseen circumstances, or *unknown triggering conditions* in the context of SOTIF. It represents a defensive claim about the *resilience* of the system, i.e. its ability to still meet its requirements *even if the assumptions do not hold*.

[9] A more formal treatment of the operator is provided in the literature [12].

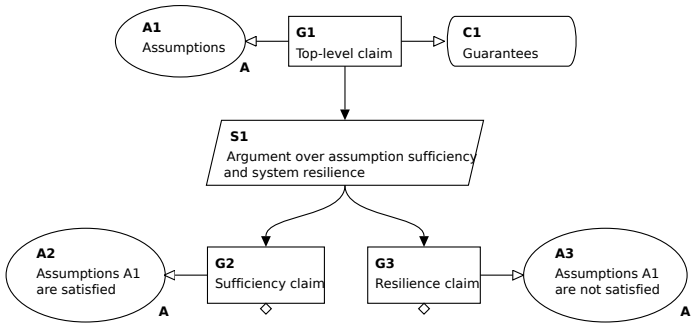

Fig. 3. Formally grounded high-level argument structure

This formalisation allows us to represent our belief in the occurrence of the four SOTIF scenarios as follows[10]:

- $E(\omega_{y|x})$: expectation about the system satisfying its guarantees when the assumptions hold. It thus reflects *known safe* scenarios.
- $1 - E(\omega_{y|x})$: expectation about the system violating its guarantees despite the assumptions holding. It thus reflects *known unsafe* scenarios.
- $E(\omega_{y|\bar{x}})$: expectation about the system satisfying its guarantees despite the assumptions not holding. It thus reflects *unknown safe* scenarios.
- $1 - E(\omega_{y|\bar{x}})$: expectation about the system violating its guarantees when the assumptions do not hold. It thus reflects *unknown unsafe* scenarios.

We can now use these building blocks to introduce an argument structure (shown in Figure 3) that directly reflects the deductive structure and thus allows for an automated assessment of assurance confidence. The argument consists of a top-level claim **G1** (corresponding to the final opinion $\omega_{y||x}$) that is broken down into sub-claims about the sufficiency of the assumptions (**G2**) and the resilience of the system in case of violated assumptions (**G3**), thus reflecting the sub-conditionals $\omega_{y|x}$ and $\omega_{y|\bar{x}}$[11]. The overall strength of the argument therefore depends on a combination of (1) our belief that the system meets its requirements when the assumptions hold, (2) our belief that the assumptions hold, and (3) our belief that the system will meet its requirements despite the assumptions not holding. The next section illustrates how this structure helps to construct and assess an argument.

4 Constructing the argument and computing confidence

We will now illustrate the process of argument construction and evaluation with a simple running example of an autonomous system function. We consider the

[10] We denote with $E(\omega)$ the expectation value of the Beta PDF of opinion ω.

[11] By a slight abuse of GSN notation, we use **A2** and **A3** to 'override' **A1**. An alternative (more GSN-compliant) representation would be to replace **A1** with **A2** and formulate **A3** as the negation of **A2**.

Input space	64x64 RGB pixel images of traffic signs
Assumptions	Camera is operating correctly and within its operating range with nominal noise, weather conditions as defined in the specification.
Top claim	The model satisfies its safety requirements.
Guarantees	The function does not miss construction signs.
Sufficiency claim	If the assumptions hold, then each construction sign image shall be correctly classified as such.
Resilience claim	Operation-time measures are in place to detect the violation of assumptions and deactivate the function.

Table 1. Safety contract for construction sign classification

case of ML-based traffic sign recognition (TSR) whose failure may have safety-relevant consequences and therefore requires safety assurance. We assume that the TSR is used as part of a highway pilot function that is enabled to automatically control the vehicle under certain conditions on a highway and should adjust its driving strategy based on prompts it receives from roadside traffic signs. We are particularly concerned with the recognition of *construction site signs* so that the highway pilot can be deactivated and control passed back to the driver.

To illustrate the methodology, we trained a Resnet-18 model on the GTSRB dataset[12] and assess its performance. The central question that we aim to answer is the following: **Given our assumptions and the available evidence, what is our confidence in the satisfaction of the top-level claim (and thus also the guarantees)?** To this end, we perform the following steps. First, we formulate a safety contract according to the formally derived argument structure described in Section 3. We then turn the safety contract into a GSN structure similar to that in Fig. 3. For each building block in the GSN argument, we then form SL opinions based on actual quantitative measurements or qualitative judgment. Finally, we combine the SL opinions using the SL deduction operator and derive an opinion ω_{G1} about top-level goal **G1**. We then determine the uncertainty in the top-level claim by calculating the *lower confidence bound (LCB)*, i.e. the lower bound of the 95% credible interval associated with ω_{G1}.

Step 1: Formulating the safety contract: The safety contract for the function is shown in Table 1. First, some *assumptions* about the system environment are made. For simplicity, we restrict the focus here to the quality of the camera. Next, the *top claim* states that the function satisfies its safety requirements which are stated as *guarantees* directly below. The *sufficiency claim* states that, if assumptions are satisfied, no construction signs shall be missed (i.e. false negatives are to be avoided). And finally, the *resilience claim* states that, any violation of the assumptions will be detected and the system will be deactivated.

[12] `https://benchmark.ini.rub.de/index.html`

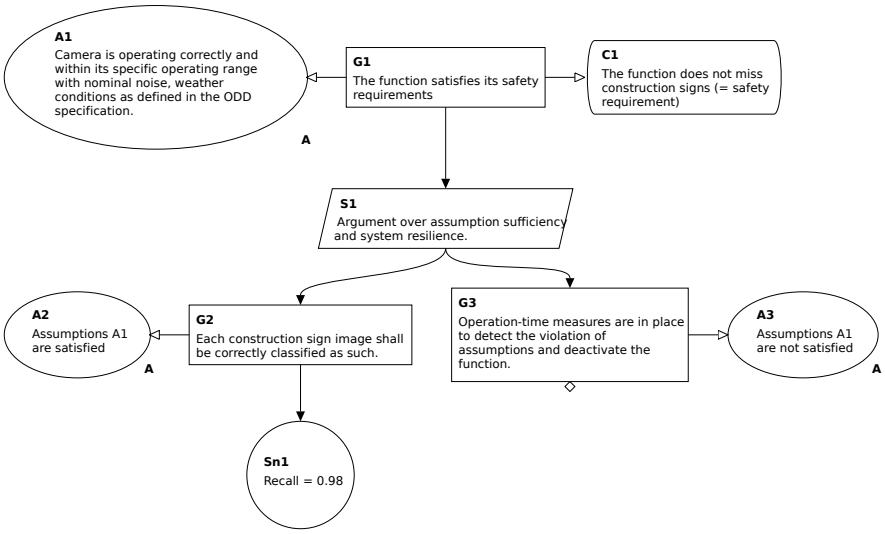

Fig. 4. Assurance argument structure for the traffic sign example

Step 2: Construction of the GSN argument structure: Adapting the structure shown in Figure 3 to our example results in the argument shown in Figure 4. Goal **G2** is substantiated with a measured recall value of 98%.

Step 3: Formulation of SL opinions: Now, we formulate SL opinions for each of the building blocks in the GSN argument in Figure 4. To illustrate the mapping, we name each opinion ω_X where X is the identifier of the respective building block. For example, ω_{A1} is the opinion associated with **A1**. We proceed bottom-up by first focussing on the evidence **Sn1** used to substantiate **G2**.

*Modelling **Sn1**:* In order to satisfy **G2**, we identified recall as the relevant metric here. Applying this measurement to the trained model yields $TP = 500$ (true positives) and $FN = 10$ (false negatives) and therefore a recall rate of $10/(500+10) = 0.98$. By interpreting TP as positive evidence and FN as negative evidence, we use Equations 2–4 to derive the belief, disbelief, and uncertainty masses of opinion ω_{Sn1} as follows (assuming $W = 2$):

$$b_{Sn1} = 500/(500 + 10 + 2) \approx 0.977 \tag{8}$$

$$d_{Sn1} = 10/(500 + 10 + 2) \approx 0.02 \tag{9}$$

$$u_{Sn1} = 2/(500 + 10 + 2) \approx 0.004 \tag{10}$$

A summary of this opinion including expectation value and LCB is given in the first line of Table 2. We see that the LCB is 96.4% and thus significantly lower

Opinion	Bel.	Disb.	Unc.	Exp.	LCB (95%)
ω_{Sn1}	0.977	0.02	0.004	0.979	**0.964**
ω_{G2}	0.977	0.02	0.004	0.979	**0.964**
ω_{G3}	0.0	0.0	1.0	0.5	**0.025**
ω_{A1}	0.874	0.051	0.075	0.911	**0.78**
ω_{G1}	**0.89**	**0.018**	**0.092**	**0.98**	**0.896**

Table 2. SL opinions for the traffic sign classification example

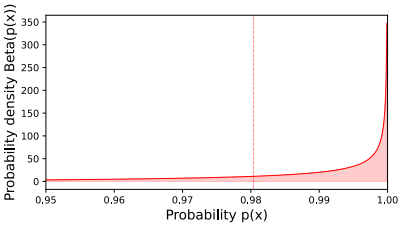

Fig. 5. Beta distribution visualisation of ω_{G1}, the opinion associated with **G1**

than the originally measured 98% recall. This reflects the uncertainty resulting from the comparatively small number of data points used to measure recall.

*Modelling **G2**:* Since **Sn1** is the only evidence used to substantiate goal **G2**, it inherits the belief, disbelief, and uncertainty masses of ω_{Sn1}, resulting in equivalent numbers in Line 2 of Table 2.

*Modelling **G3**:* **G3** represents the resilience claim and requires the existence of additional measures to detect violations of the assumptions in order to deactivate the function. Let us assume that we have no information about the existence of additional measures yet, expressed by a completely uncertain opinion $\omega_{G3} = (b = 0.0, d = 0.0, u = 1.0)$. As shown in Line 4 of Table 2, this results in a uniform Beta distribution with expectation value 0.5 and an LCB of 0.025.

*Modelling **A1**:* **A1** expresses the assumptions about the system environment that form the basis of the argument. Let us assume that no quantitative measurements are available, and that we have to rely on qualitative judgment instead. To this end, we can use the idea of qualitative opinions described in [12]. We start with a qualitative judgement of *likelihood* (on a discrete scale from *absolutely unlikely* to *absolutely likely*) and *confidence* (on a discrete scale from *no confidence* to *total confidence*) and turn that into an opinion by mapping the qualitative tuple to the corresponding area of the opinion triangle. Let us assume that the assumptions are "very likely true with high confidence" which results in opinion $\omega_{A1} = (b = 0.874, d = 0.051, u = 0.075)$. The numbers in Line 5 of Table 2 illustrate that, although the expectation value is 0.91, the opinion is associated with a fair amount of uncertainty, indicated by an LCB of 0.78.

Step 4: Computing overall uncertainty in G1: We now have all ingredients to calculate opinion ω_{G1} using deduction from ω_{A1}, ω_{G2}, and ω_{G3} according to Equation 7. The calculation yields the following results:

$$\omega_{G1} = \omega_{A1} \circledcirc (\omega_{g2}, \omega_{g3}) = (b = 0.89, d = 0.018, u = 0.092) \tag{11}$$

As shown in the bottom line in Table 2, the resulting opinion about **G1** has a high expectation value (98%), yet there is a significant amount of uncertainty associated with it which negatively impacts confidence. A Beta distribution visualisation of ω_{G1} is shown in Figure 5.

5 Related work

Assurance confidence estimation aims to reduce uncertainties associated with validity of the assurance argument itself. Qualitative approaches aim to decrease uncertainty by strengthening the argument itself, e.g. through additional confidence-specific claims, sub-claims, and evidences. Hawkins *et al.* [9] present the concept of *assured safety arguments*, an extension that separates safety argumentation from confidence argumentation. The two types of argument are connected through *assurance claim points* (*ACPs*) in the structural notation of the argument. Whilst this approach aims to provide confidence in the overall safety argument through a separate set of confidence arguments, it does not allow for the assignment of a quantitative confidence metric. A range of quantitative approaches have also been presented, e.g. using eliminative induction and Baconian probabilities [6], Bayesian inference [8, 4, 11], or Dempster-Shafer (D-S) belief functions [1, 14]. Compared with plain probabilities, D-S-based approaches allow for the representation of higher-order uncertainty (i.e. uncertainty in the probabilities themselves). However, there has been some confusion regarding the application of Dempster's rule of combination [16] which, according to Jøsang, is mainly due to a misinterpretation of the nature of situations to be modelled [12]. Subjective Logic (SL) addresses this problem and incorporates a wide range of operators to model different situations. Since we believe that a safety argument contains deductive aspects (as described in this paper), but also trust discounting relationships for the representation of defeaters [10], and situations of belief fusion, we adopt SL with its various operators for our work.

SL has also been considered by other authors. Duan *et al.* [5] build upon the work of Goodenough *et al.* [6] by quantifying the Baconian probability as a Beta distribution and visualising it in Jøsang's opinion triangle. Yuan *et al.* [15] propose an approach closely related to ours where they utilise SL operators to make formal inferences in an argument. However, whereas we aim to reason about the overall satisfaction of the safety contract (the relationship between assumptions and guarantees), they use the approach to formalise different argument types (e.g. one-to-one and alternative arguments) in the body of an argument. As such, the two approaches are essentially complementary.

On a more general level, "Assurance 2.0" (henceforth A2.0) has been proposed [13], a methodological approach that requires assurance cases to be "as deductive as possible and inductive only as strictly necessary" and suggests three principles for their assessment: (1) the *positive perspective* which considers the extent to which the evidence and overall argument make a positive case to justify belief in its claims (involving *logical soundness* and *validity*), (2) the *negative perspective* which involves active search for and resolution of *doubts* and *defeaters*, and

(3) the *residual perspective* that considers the risk of unresolved doubts and defeaters. A2.0 provides a useful framework to contextualise our work in this paper. First, structuring the argument based on the idea of a deduction of guarantees from assumptions addresses the positive perspective, in particular logical validity. Second, a subdivision of the argument into sufficiency and resilience claims makes an explicit distinction between the positive and the residual perspective. Third, using SL opinions to represent individual building blocks of the argument helps to assess the soundness of the evidence by explicitly representing belief, disbelief, and uncertainty. And, finally, as shown in [10], a consideration of different layers of uncertainty and the formulation of *trust discounting relationships* between evidences can help to identify and formally express doubts in the argument and to formulate conservative adjustments of any estimates.

6 Discussion and conclusions

Being able to reason about the uncertainty in a safety argument is an important prerequisite for increasing the trustworthiness of autonomous system functions. This, however, is exacerbated by the informal nature of many real-world safety arguments which is primarily due to a lack of formal underlying semantics. In this paper, we addressed this problem by viewing the satisfaction of a safety contract as a conditional deduction operation from assumptions to guarantees. A formalisation of this idea yields a high-level structure for safety arguments that allows for automated reasoning about the uncertainty associated with the guarantees based on the stated assumptions and the available evidence. With this, we hope to make a contribution towards a more formal treatment of safety arguments by extending the existing body of knowledge with a novel perspective.

Obviously, many questions remain open and indicate directions for future research. First, we only addressed the very high-level structure of an argument. It would be interesting to investigate how the idea presented in this paper could be combined with some of the existing approaches mentioned in Section 5, for example Yuan *et al.*'s work on the formalisation of argument types [15], to support a safety engineer with the construction of a comprehensive argument. Furthermore, we believe that the formalisation may also be a helpful starting point for the derivation of 'defeaters' which could be attached as assurance claim points to the argument in the spirit of [9]. As shown extensively by Graydon *et al.* [7], no formalism can, by itself, provide guarantees about the computed confidence and it is crucial to understand hidden implications of the chosen evidence on the confidence results. We therefore believe that doubts in evidence need to be identified and also incorporated into the argument as, e.g., addressed in our ongoing work on uncertainty-driven safety assurance [10, 3].

Acknowledgments. This work was performed as part of the ML4Safety project supported by the Fraunhofer Internal Programs under Grant No. PREPARE 40-02702.

References

1. 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*, pages 127–144, 2013.
2. S. Burton, L. Gauerhof, B. B. Sethy, I. Habli, and R. Hawkins. Confidence arguments for evidence of performance in machine learning for highly automated driving functions. In A. Romanovsky, E. Troubitsyna, I. Gashi, E. Schoitsch, and F. Bitsch, editors, *Computer Safety, Reliability, and Security*, pages 365–377, Cham, 2019. Springer International Publishing.
3. S. Burton and B. Herd. Addressing uncertainty in the safety assurance of machine-learning. *Frontiers in Computer Science*, 5, 2023.
4. E. Denney, G. Pai, and I. Habli. Towards measurement of confidence in safety cases. In *2011 International Symposium on Empirical Software Engineering and Measurement*, pages 380–383, 2011.
5. L. Duan, S. Rayadurgam, M. Heimdahl, O. Sokolsky, and I. Lee. Representation of confidence in assurance cases using the beta distribution. In *2016 IEEE 17th International Symposium on High Assurance Systems Engineering (HASE)*, pages 86–93. IEEE, 2016.
6. J. B. Goodenough, C. B. Weinstock, and A. Z. Klein. Eliminative induction: A basis for arguing system confidence. In *2013 35th International Conference on Software Engineering (ICSE)*, pages 1161–1164, 2013.
7. P. J. Graydon and C. M. Holloway. An investigation of proposed techniques for quantifying confidence in assurance arguments. *Safety science*, 92:53–65, 2017.
8. B. Guo. Knowledge representation and uncertainty management: applying Bayesian Belief Networks to a safety assessment expert system. In *International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings. 2003*, pages 114–119, 2003.
9. R. Hawkins, T. Kelly, J. Knight, and P. Graydon. A new approach to creating clear safety arguments. In *Advances in systems safety*, pages 3–23. Springer, 2011.
10. B. Herd and S. Burton. Can you trust your ML metrics? Using Subjective Logic to determine the true contribution of ML metrics for safety. *Proceedings of the the 39th ACM/SIGAPP Symposium On Applied Computing (SAC24)*, 2024.
11. C. Hobbs and M. Lloyd. The application of Bayesian Belief Networks to assurance case preparation. In C. Dale and T. Anderson, editors, *Achieving Systems Safety*, pages 159–176, London, 2012. Springer London.
12. A. Jøsang. *Subjective logic*, volume 3. Springer, 2016.
13. S. Varadarajan, R. Bloomfield, J. Rushby, G. Gupta, A. Murugesan, R. Stroud, K. Netkachova, and I. H. Wong. CLARISSA: Foundations, tools & automation for assurance cases. In *2023 IEEE/AIAA 42nd Digital Avionics Systems Conference (DASC)*, pages 1–10. IEEE, 2023.
14. R. Wang, J. Guiochet, G. Motet, and W. Schön. Safety case confidence propagation based on Dempster–Shafer theory. *International Journal of Approximate Reasoning*, 107:46–64, 2019.
15. C. Yuan, J. Wu, C. Liu, and H. Yang. A subjective logic-based approach for assessing confidence in assurance case. *International Journal of Performability Engineering*, 13(6):807, 2017.
16. L. A. Zadeh. Book review: A mathematical theory of evidence. *AI Mag.*, 5(3):81–83, 1984.