Strategy-Aware Confidence Assessment with Stakeholder Consensus in Automated Driving Assurance Cases

Yutaka Matsuno✉[1][0000−0001−9809−0814],
Michio Hayashi[2][0000−0002−5702−1334], and Tomoyuki Tsuchiya[3]

[1] Nihon University, Japan
matsuno.yutaka@nihon-u.ac.jp
[2] TIER Ⅳ North America, Inc., USA
[3] TIER Ⅳ, Inc., Japan
{michio.hayashi.2,tomoyuki.tsuchiya}@tier4.jp

Abstract. Deploying SAE Level 4 (L4) automated vehicles requires both a technically credible safety argument and acceptance of it by diverse stakeholders. While our previous work [17] introduced a lightweight *Consensus Score* to quantify this acceptance within a Goal Structuring Notation (GSN) framework, measuring consensus alone is insufficient. It cannot distinguish between a fundamentally weak technical justification and simple misunderstandings caused by poor communication. On the other hand, existing Confidence Assessment Methods (CAMs) that measure technical validity are often too complex or costly for fast-paced industrial projects [5]. This paper therefore extends our consensus-measuring workflow with *Strategy-Aware Confidence*, a lightweight confidence propagation model based on expert ratings. In this model, sub-goal confidence is aggregated using inverse-variance weighting, while parent-goal confidence is discounted based on the assessed validity of the strategy. This ensures that uncertainty in the reasoning pattern propagates upward. By keeping organizational acceptance (Consensus) and technical validity (Confidence) as separate dimensions, we provide a two-dimensional dashboard to analyze the overall maturity of an assurance case. The entire workflow, from GSN editing through questionnaire generation to dashboard visualization, is implemented in D-Case Communicator, a web-based GSN tool. We report practical experience applying this framework to a SOTIF (ISO 21448) assurance case for a planned L4 automated driving service and the practical challenges encountered in the process.

Keywords: assurance case · GSN · confidence assessment · stakeholder consensus · SOTIF · automated driving

1 Introduction

An SAE Level 4 (L4) automated driving system [20] functions as an "open system" [12] whose environment, and therefore whose technical and operational

risks, evolve continuously. A static safety argument is thus insufficient; a continuous assurance argument [15] capable of evolving with the system is required. The open nature of L4 automated driving also broadens the set of stakeholders who must agree on the state of assurance, making consensus building a practical necessity.

Two challenges arise. First, communicating GSN-based assurance cases [21] to non-experts is a longstanding difficulty [18]. Second, while existing confidence assessment methods (CAMs) [3,8] can assess argument validity, they do not directly indicate whether diverse stakeholders *accept* the argument as a basis for decision-making [5].

In previous work [17], we addressed these challenges during an L4 minibus field trial led by TIER IV. We complemented a GSN-based assurance case with a plain-language *Safety Status Report*, distributed an anonymous questionnaire, and introduced a *Consensus Score* that blends top-down and bottom-up ratings into a single acceptance measure. The Consensus Score, however, primarily captures stakeholder acceptance of safety claims rather than the argument's technical soundness. This calls for a clearer distinction between two questions:

- **Organizational acceptance:** "Do stakeholders accept this safety claim as a basis for decision-making?"
- **Technical validity:** "Is the argument technically persuasive, and what is the degree of uncertainty?"

Assessing technical validity requires an explicit reasoning structure, yet published safety cases often leave the decomposition rationale implicit: some existing safety arguments are written entirely without formal notation [7] or use GSN but omit or underspecify strategy nodes. Most existing CAMs propagate confidence from leaf evidence upward without separately assessing the decomposition logic. In our previous study [17], we attempted to apply existing CAMs but found them impractical given the project's time and resource constraints, motivating us to seek an approach that evaluates the reasoning pattern directly.

We therefore propose *Strategy-Aware Confidence*, which measures the technical validity of an assurance-case argument by accounting for both evidence quality and the validity of the decomposition strategy. Each assessed Goal and Strategy node is characterized by a sample mean and variance computed across expert $[0, 1]$ ratings. Sub-goal confidences are combined by inverse-variance weighting (IVW), and parent goals are *discounted* by the assessed validity of the argument strategy, preventing weak reasoning from being hidden by strong evidence. We integrate this confidence metric with stakeholder acceptance in a *two-dimensional Consensus–Confidence dashboard* that reveals where each assurance argument stands: low *confidence* calls for a different decomposition strategy or stronger evidence, while low *consensus* may reflect poor communication or unmet stakeholder concerns rather than weak evidence.

This paper makes four contributions:

1) **Strategy-Aware Confidence:** An expert-elicited confidence propagation model that treats GSN strategy nodes as explicit sources of uncertainty. Unlike CAMs that rely on analyst-specified conditional probability tables or

belief tuples, our approach requires only a single $[0, 1]$ score per node, lowering a practical barrier to adoption [5]. A strategy exponent α, calibrated against expert holistic ratings, controls discounting strength and yields diagnostic gaps at intermediate goals for targeted improvement.

2) **Two-dimensional decision dashboard:** A view that separates organizational acceptance from technical validity and links each score combination to a recommended follow-up action. To the best of our knowledge, no prior work visualizes these two dimensions in a single diagnostic view.
3) **Tool support:** D-Case Communicator 2.0[4], a full reimplementation of the web-based GSN editor [16], supports the workflow from GSN editing through questionnaire generation to Consensus–Confidence computation and visualization.
4) **SOTIF case study:** An industrial case study applying quantitative confidence assessment to a SOTIF assurance case for a planned L4 automated driving service, the first such application to the best of our knowledge.

The remainder of this paper is organized as follows. Section 2 reviews related work. Section 3 describes the proposed Strategy-Aware Confidence model. Section 4 presents the case study and discusses findings and limitations. Section 5 concludes the paper.

2 Related Work

Assurance cases are increasingly used in automated driving. UL 4600 [22] explicitly requires a structured safety case. ISO 26262 [13] and ISO 21448 (SOTIF) [14] define lifecycle processes whose work products serve as safety-case evidence. GSN [21] is widely adopted for structuring safety arguments, but communicating GSN-based arguments to non-experts remains difficult [18], and existing industrial frameworks offer limited guidance on measuring stakeholder acceptance.

Several papers address quantitative assessment of assurance-case strength. Denney et al. [3] proposed mapping a GSN structure onto a Bayesian Network (BN) to propagate confidence, identifying strategy and context validity as distinct uncertainty sources but requiring analyst-specified conditional probability tables. Nešić et al. [19] extended the BN approach by making inference validity an explicit factor, but both share the elicitation burden of specifying conditional probability tables.

Alternative formalisms address epistemic uncertainty where classical probabilities are insufficient. Yuan et al. [24] defined four argument types (one-to-one, alternative, conjunction, disjunction) with type-specific Subjective Logic propagation rules, requiring sufficiency, necessity, and confidence for each node; Herd et al. [11] applied Subjective Logic to quantitative assurance arguments and extended it to handle defeaters; Diemert and Weber [6] proposed Certus, a fuzzy-set-based language for expressing confidence in linguistic terms. These methods

[4] `https://d-case.matsulab.org/`

can represent epistemic uncertainty more flexibly but typically require multiple parameters per node. Beyond uncertainty formalisms, Goodenough et al. [8] proposed eliminative argumentation, which identifies and resolves defeaters. Bloomfield and Rushby [2] discuss Assurance 2.0, which separates logical soundness from probabilistic confidence.

Graydon and Holloway [10] showed that most existing quantitative confidence techniques can yield unrealistic results, such as confidence inflation from weak evidence. Conversely, Barrett et al. [1] demonstrated that multiplicative propagation can cause the opposite problem: achieving 95% top-level confidence in a seven-component fragment required approximately 99.3% confidence in every leaf, leaving realistic calibration as an "open question." Diemert et al. [5] found that only 2 of 19 practitioners had ever applied a quantitative CAM to a real system. Published real-world applications are scarce: most proposals have been validated only on illustrative examples [10]. A common barrier is the elicitation burden: specifying conditional probability tables, Dempster–Shafer mass functions, or multi-valued belief tuples for every node is impractical under industrial constraints.

On the tooling side, CLARISSA/ASCE [23] implements Assurance 2.0 confidence and Socrates [4] uses Bayesian networks with defeater support. Neither tool generates questionnaires from the argument structure or derives propagation parameters from stakeholder ratings.

Our model builds on Denney et al. [3] but addresses two concerns separately: structural validity (via a strategy variable S) and sub-goal aggregation (via IVW), requiring only a single $[0, 1]$ score per node. As in Nešić et al. [19], inference quality affects parent confidence, but through a direct multiplicative discount. IVW also reduces confidence inflation from weak evidence [10]. A strategy exponent α, calibrated against expert holistic ratings, addresses the calibration problem identified by Barrett et al. [1]. To the best of our knowledge, no prior CAM provides a mechanism to calibrate propagated confidence; our approach uses holistic expert ratings as the alignment target. Keeping organizational acceptance and technical validity on separate axes addresses the gap noted by Diemert et al. [5]: practitioners need CAMs that support stakeholder communication, not merely score computation. The model does not yet formally incorporate defeaters [2,6,11]; each questionnaire item includes a free-text field that collects defeater-like feedback informally.

3 Strategy-Aware Confidence Assessment

Consider a parent goal G_{parent} in a GSN structure, supported by sub-goals $\{G_1, \ldots, G_n\}$ through a strategy S. We model S as a variable representing the validity of the decomposition logic, aggregate sub-goal confidences into an evidence variable E via IVW, and discount E by S.

Expert $[0, 1]$ ratings yield a sample mean μ and variance σ^2 per node. IVW assigns greater weight to sub-goals with lower uncertainty (ϵ=10^{-6} prevents

division by zero):

$$w_i = \frac{1}{\sigma_{G_i}^2 + \epsilon}, \qquad \mu_E = \frac{\sum_{i=1}^{n} w_i \cdot \mu_{G_i}}{\sum_{i=1}^{n} w_i}, \qquad \sigma_E^2 = \frac{1}{\sum_{i=1}^{n} w_i} \tag{1}$$

Parent-goal confidence is obtained by discounting the evidence score with strategy validity. Multiplicative discounting across multiple GSN layers can drive top-level confidence to unrealistically low values; Barrett et al. [1] identify realistic calibration of such propagation as an open question. To address this, we introduce a strategy exponent $\alpha \geq 0$ as a calibration parameter that controls discounting strength:

$$\mu_G = \mu_E \cdot \mu_S^{\alpha} \tag{2}$$

When α=1 the model reduces to direct multiplication; α<1 weakens the strategy discount; α>1 strengthens it. The variance of μ_S^{α} is approximated by the delta method ($\sigma_{S'}^2 = (\alpha\, \mu_S^{\alpha-1})^2\, \sigma_S^2$), and the parent-goal variance follows Goodman's formula [9], which assumes that E and S are approximately independent:

$$\sigma_G^2 = \mu_E^2\, \sigma_{S'}^2 + (\mu_S^{\alpha})^2\, \sigma_E^2 + \sigma_E^2\, \sigma_{S'}^2 \tag{3}$$

This computation is applied recursively up the GSN tree. Section 4 describes how α is calibrated from the data.

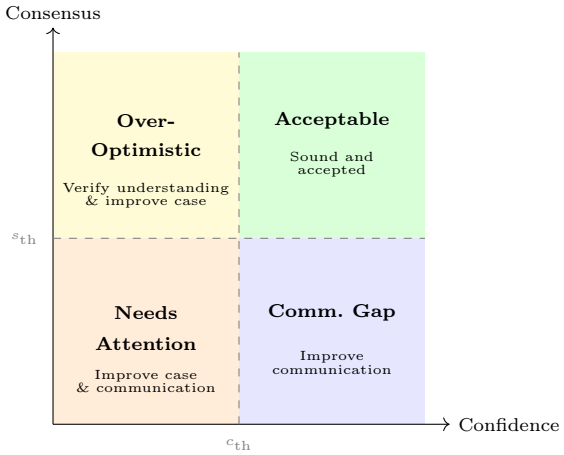

Fig. 1. Consensus–Confidence quadrant interpretation.

The Consensus Score, introduced in our previous work [17], quantifies organizational acceptance. For each GSN node, every respondent's rating is normalized to $[0, 1]$. A *top-down* score averages these ratings directly, while a *bottom-up* score propagates child-node scores upward; the two are blended with equal weight to produce the final Consensus Score. Unlike Confidence, which is derived only from

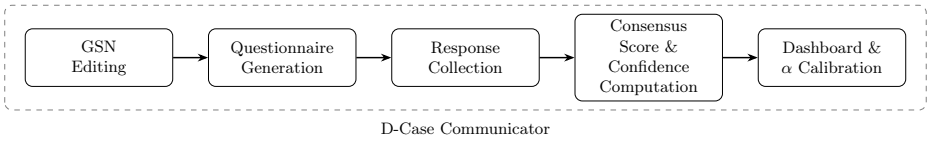

Fig. 2. End-to-end assessment workflow implemented in D-Case Communicator.

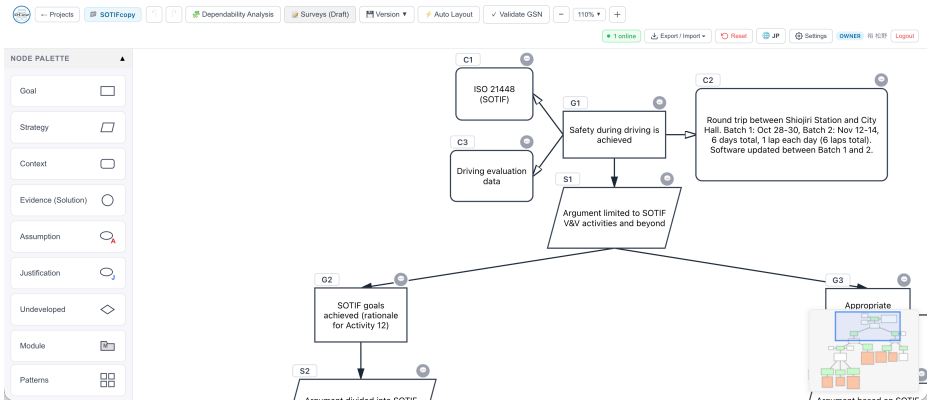

Fig. 3. Screenshot of D-Case Communicator.

expert ratings, the Consensus Score incorporates both expert and non-expert responses with uniform weighting.

Each node is then mapped onto a (Consensus, Confidence) plane, yielding four quadrants (Fig. 1): *Acceptable* (both scores above their respective thresholds), *Communication Gap* (high Confidence, low Consensus), *Over-Optimistic* (high Consensus, low Confidence), and *Needs Attention* (both below their thresholds). The thresholds c_{th} (Confidence) and s_{th} (Consensus) are project-specific.

Fig. 2 shows the end-to-end workflow implemented in D-Case Communicator (Fig. 3). The tool generates role-specific questionnaires from a GSN model. Experts rate leaf goals and strategy nodes on a continuous $[0, 1]$ scale; these ratings drive the confidence propagation defined above and also enter the Consensus Score. For intermediate goals, experts use a four-point Likert scale (0–3, normalized to $[0, 1]$); these ratings contribute only to the Consensus Score, since intermediate-goal confidence is computed by upward propagation rather than direct elicitation. Non-experts rate all nodes on the same Likert scale and contribute exclusively to the Consensus Score. Evidence nodes are not rated directly; instead, experts account for evidence quality when rating the corresponding leaf goals, because evidence quality is meaningful only in relation to the goal it supports. Undeveloped goals (e.g., G7) have no sub-goals or evidence; experts rate them directly, and the resulting mean and variance reflect confidence without supporting evidence. The two-dimensional dashboard also supports interactive α calibration (Section 4.2).

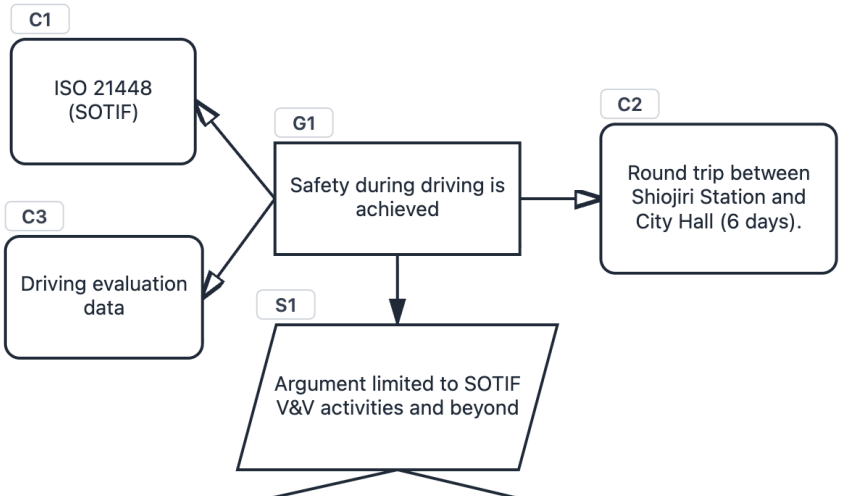

Fig. 4. Top-level GSN for the Shiojiri L4 automated driving safety case.

4 Case Study

TIER IV is preparing an SAE Level 4 automated driving service in Shiojiri, Japan. As part of this preparation, we collected driving evaluation data in two batches over six days on the route between Shiojiri Station and City Hall (Batch 1: October 28–30; Batch 2: November 12–14, 2025). A software update for feature improvements was applied between batches. We applied our framework in this context. The objective was not to obtain safety certification or prove full system safety, but to assess current consensus and confidence in the SOTIF portion of the assurance case and to identify areas needing improvement before the start of operations. Thus, the target assurance case was based on SOTIF [14] with specific focus on verification and validation (V&V) activities (SOTIF Sections 9–12), and was developed collaboratively using D-Case Communicator.

Fig. 4 illustrates the top level of the GSN model. The top goal G1, "Safety during driving is achieved," is contextualized by the SOTIF standard (C1), the operational route and evaluation schedule (C2), and the driving evaluation data (C3). Strategy S1 decomposes G1 into two sub-goals: G2 (achievement of the SOTIF goal, basis for Activity 12) and G3 (management of specific SOTIF issues).

Fig. 5 details the G2 branch. Strategy S2 splits the argument along SOTIF Activities 10 and 11. Goal G4 (Activity 10) asserts that risks from known scenarios are minimized, supported by the known scenario list (C5) and a scenario design guide (C6). Strategy S4 further decomposes G4 into three leaf goals:

– G8: classifying functional scenarios by encounter rates (supported by Sn4);
– G9: verifying the Operational Design Domain (ODD) for each automated driving feature using field data (supported by Sn5);

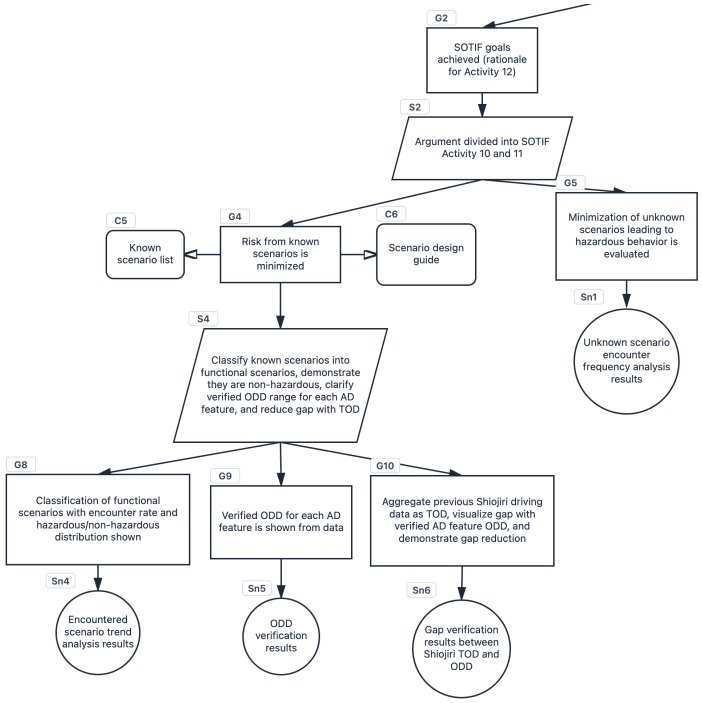

Fig. 5. Sub-GSN rooted at G2: SOTIF goal achievement via Activities 10 and 11.

– G10: visualizing the gap between Target ODD and verified ODD to demonstrate gap reduction (supported by Sn6).

Goal G5 (Activity 11) claims that unknown scenarios that could lead to hazardous behavior have been evaluated for minimization (supported by Sn1).

Fig. 6 depicts the G3 branch. Strategy S3, derived from SOTIF Activity 7, decomposes G3 into two sub-goals: identification and risk evaluation of triggering conditions (G6, supported by Sn2 and Sn3) and remediation of unacceptable SOTIF issues (G7), still undeveloped (U1), awaiting remediation confirmation. The resulting GSN comprises 10 goals, 4 strategies, 6 context nodes, 6 evidence (solution) nodes, and 1 undeveloped node. Strategy nodes S1–S4 each correspond to a specific SOTIF activity, making the reasoning pattern explicit and independently assessable.

Following the workflow in Fig. 2, D-Case Communicator distributes questionnaires via separate URLs for expert and non-expert respondent groups. To account for cumulative discounting and anchor expert ratings, the questionnaire provided the verbal anchors shown in Table 1.

The verbal anchors align expert expectations: they are set relatively high because multiplicative discounting across two to three GSN layers drives down top-level confidence. Accordingly, we set the Confidence threshold for the dashboard

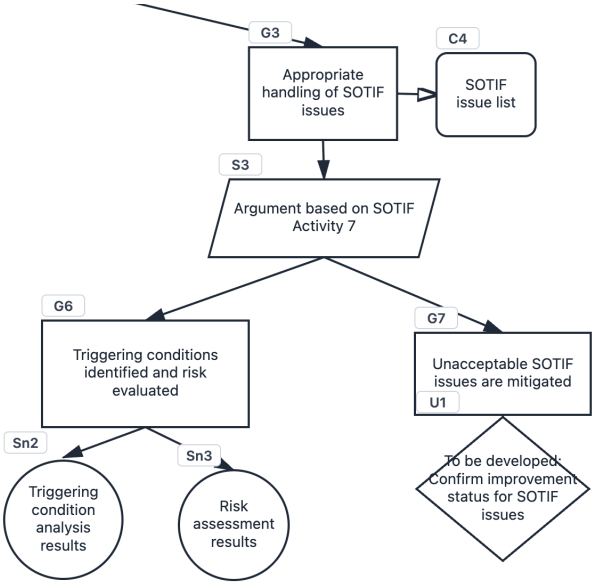

Fig. 6. Sub-GSN rooted at G3: SOTIF issue handling via Activity 7.

Table 1. Verbal anchors for expert $[0, 1]$ confidence ratings.

Value	Guidance
0.98	Near-textbook. Virtually no gaps or open questions.
0.95	Generally sound. Minor concerns but practically acceptable.
0.85	Sound but with weaknesses. Assumption dependence or coverage gaps.
0.70	Considerably weak. Missing key assumptions or unanticipated scenarios.
0.55	Substantially insufficient. Core reasoning gaps requiring reinforcement.
0.40	Fundamentally questionable. Argument or evidence needs restructuring.

quadrants at 0.7: since the anchors define this value as "considerably weak," a node at or below this level cannot be considered acceptable. Because no similar anchors were provided for the Consensus questionnaire, its threshold is set at the neutral midpoint 0.5.

4.1 Results

We collected responses from 15 respondents from TIER Ⅳ: 5 safety experts with SOTIF assessment experience and 10 non-expert respondents involved in the Shiojiri project. Table 2 shows the raw expert $[0, 1]$ ratings for directly assessed nodes (leaf goals and strategies).

Two patterns are notable in the raw ratings. First, expert ratings were higher for strategies than for leaf goals on average (strategy mean 0.76 vs. leaf-goal mean 0.56). Second, expert disagreement is largest for strategy nodes S1 (σ^2=0.069)

Table 2. Expert $[0, 1]$ confidence ratings for directly assessed nodes.

Node	A	B	C	D	E	μ	σ^2
G5 (unknown scenarios)	0.70	0.55	0.50	0.80	0.40	0.59	0.026
G6 (triggering cond.)	0.70	0.40	0.50	0.70	0.70	0.60	0.020
G7 (SOTIF remediation)	0.40	0.55	0.40	0.50	0.40	0.45	0.005
G8 (scenario classif.)	0.55	0.45	0.70	0.70	0.40	0.56	0.019
G9 (ODD verification)	0.55	0.55	0.50	0.80	0.70	0.62	0.016
G10 (Target ODD gap)	0.75	0.30	0.40	0.70	0.40	0.51	0.041
S1 (SOTIF V&V scope)	0.98	0.85	0.40	0.60	0.40	0.65	0.069
S2 (Act. 10 & 11)	0.85	0.90	0.85	0.99	0.98	0.91	0.005
S3 (Act. 7)	0.85	0.95	0.55	0.75	0.95	0.81	0.028
S4 (known-scenario decomp.)	0.64	0.90	0.40	0.99	0.40	0.67	0.075

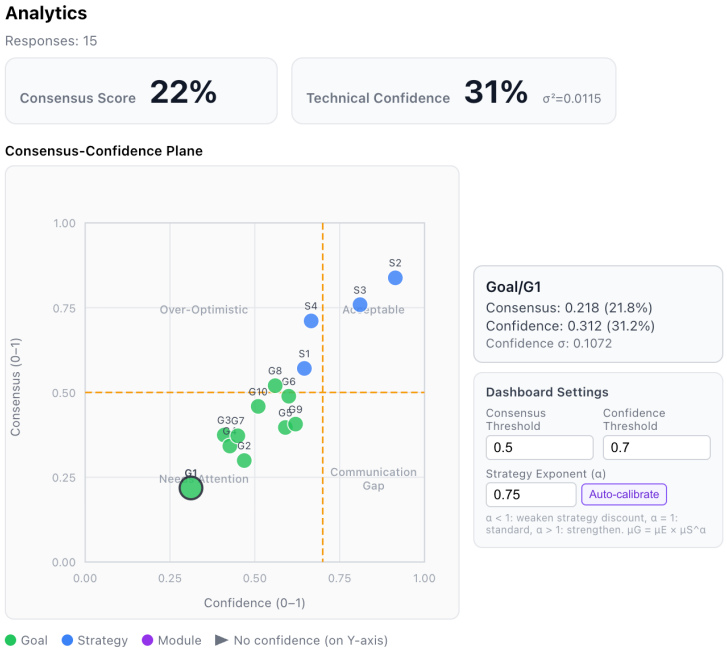

Fig. 7. Consensus–Confidence dashboard with threshold and α settings (right panel).

and S4 (σ^2=0.075): Experts C and E each rated both at 0.40, citing scope limitations (S1) and insufficient scenario coverage (S4). Among leaf goals, G10 (Target ODD gap) shows the highest variance (σ^2=0.041). D-Case Communicator propagates these ratings upward through the GSN tree. Fig. 7 shows the resulting Consensus–Confidence dashboard computed from all 15 responses.

The top-level goal G1 receives a Consensus Score of 0.22 and a Strategy-Aware Confidence of 0.31 (σ=0.107) after calibration (α=0.75; see Section 4.2),

placing it in the *Needs Attention* quadrant. Strategy nodes S2 and S3 fall in the *Acceptable* quadrant, with S2 the strongest (confidence 0.91). However, S1 and S4 fall in *Over-Optimistic*: their consensus exceeds the 0.5 threshold but their confidence (0.65 and 0.67) falls below 0.7, reflecting disagreement among experts on the SOTIF V&V scope limitation (S1) and the known-scenario decomposition (S4). Nine of the ten goal nodes fall in *Needs Attention*; the exception is G8 (scenario classification), which lands in *Over-Optimistic* (Consensus 0.52, Confidence 0.56). G7 (SOTIF issue remediation), which remains undeveloped (U1), has the lowest mean (μ=0.45) among directly assessed goals. The results indicate that both evidence strength and argument structure need improvement.

The same asymmetry between strategies and goals appears in the non-expert Likert ratings. Non-experts gave the lowest rating (0) to goal nodes in 36% of responses (36/100), but to strategy nodes in only 5% (2/40); strategies S2–S4 received no zero ratings at all. The average non-expert score for goals was 1.04/3, compared with 2.10/3 for strategies. Experts showed the same relative pattern on their $[0, 1]$ scale, though the difference was less pronounced.

Written comments from both groups cite the same concerns: insufficient evaluation data (six days of driving data), missing acceptance criteria, and incomplete scenario coverage. One non-expert noted that the GSN structure was adequate but the content lacked persuasiveness, with no data to support the claims. Expert C, the most critical assessor, argued that, without established risk acceptance criteria, the assessment cannot determine whether identified risks are tolerable. Expert D endorsed the GSN reasoning structure (rating three of four strategies above 0.75) but questioned the exclusion of functional safety and cybersecurity from the scope.

Overall, the results aligned with expectations given the stage of preparation TIER IV was in and the small data sample available. TIER IV's main objective was to assess *strategy* validity and consensus among stakeholders, and to later strengthen evidence with more data if the strategy was vindicated.

4.2 Strategy Exponent Calibration

Realistic calibration of propagated confidence remains an open question [1]. Because experts also rate intermediate and top-level goals holistically on a Likert scale for the Consensus Score, these ratings provide an independent reference point: they reflect the expert's direct assessment of a goal as a whole, not a value derived from sub-goal ratings or the propagation formula. We calibrate α by minimizing the sum of squared errors between propagated and holistic ratings across all four goals G1–G4, yielding α=0.75. With α=1.0 propagated confidence is systematically lower than expert holistic ratings; the calibrated α<1 corrects for the overly aggressive multiplicative discount that direct multiplication (α=1) imposes in this two-to-three-layer GSN. D-Case Communicator automates this calibration (Fig. 7, right panel). The residual gaps between propagated confidence and holistic ratings serve as diagnostic observations (Table 3).

When propagated and holistic scores agree closely (G1, G4), the GSN decomposition and its evidence capture the expert's overall assessment well, lending

Table 3. Diagnostic gap (propagated − holistic) at goals (α=0.75).

Goal	Propagated	Holistic	Gap	Interpretation
G1	0.31	0.33	−0.02	Decomposition matches holistic view
G2	0.47	0.33	+0.14	Holistic rating lower; concerns beyond sub-goals
G3	0.41	0.53	−0.12	Undeveloped node (G7) depresses propagation
G4	0.43	0.40	+0.03	Decomposition matches holistic view

credibility to the argument structure at those nodes. When they diverge, the gap indicates that the expert's holistic judgment reflects factors not captured by the current decomposition. For G2 (+0.14), individual leaf nodes score reasonably, but experts see concerns not documented in the current sub-goals and rate G2 lower. For G3 (−0.12), the undeveloped goal G7 and the strategy discount from S3 drag down the propagated score, but experts rated G3 higher, possibly because they viewed SOTIF issue handling as progressing adequately despite the missing evidence. These gaps suggest concrete next steps: for G2, refine the decomposition by adding sub-goals that address the experts' holistic concerns; for G3, develop the undeveloped node G7 with supporting evidence to raise the propagated score.

Note that α is a project-specific alignment parameter, not a universal constant; its value depends on the GSN depth, scale design, and expert panel, so the resulting scores should not be compared across projects.

4.3 Discussion

Communication barriers were evident: undefined abbreviations (e.g., "VRU"), misidentification of GSN notation as TIER IV-specific terminology, and questions about evidence credibility arising from misunderstandings. Tracking how scores shift across iterative rounds can verify that improvements move in the expected direction. The two-dimensional view also guides resource allocation: *Over-Optimistic* nodes (S1, S4) signal that the decomposition needs revision before collecting additional evidence, but their high Consensus also warrants scrutiny; non-experts may agree without fully understanding the technical weaknesses, so the quality of their understanding should be verified alongside strengthening the case. *Needs Attention* leaf goals (e.g., G7) call for developing undeveloped nodes with concrete evidence. The gap analysis in Section 4.2 then pinpoints which nodes need revision and what kind of improvement is required.

We note four limitations:

1. **Aggregation operator.** The model uses IVW uniformly. IVW is well suited to multi-legged arguments, where independent evidence lines support the same claim (e.g., S4). For conjunctive decompositions (e.g., S1–S3), a minimum or geometric-mean operator [24] would better reflect the logical structure; a preliminary analysis with minimum and geometric-mean operators for S1–S3 shows that diagnostic gap directions are unchanged and RMSE shifts

only from 0.093 to 0.087–0.080, suggesting that the choice of aggregation operator has limited impact in this case study.
2. **Independence assumption.** Goodman's formula (Eq. 3) assumes that evidence quality and strategy validity are independent. In this case study, strategies received higher ratings than leaf goals on average (0.76 vs. 0.56); while this does not formally establish independence, it suggests that experts assessed strategy and evidence quality as conceptually distinct factors.
3. **Shared input data.** Expert $[0, 1]$ ratings feed both Confidence and the Consensus Score, and the Likert ratings collected for the Consensus Score are reused as the alignment target for α calibration. The $[0, 1]$ scale uses verbal anchors (Table 1) that guide ratings upward, whereas the Likert calibration target has no such anchors; this scale mismatch may bias the calibrated α.
4. **Scale design.** The Likert scale lacks a "cannot judge" option, making it impossible to distinguish genuine disagreement from insufficient understanding. As noted in Section 4.1, 36% of non-expert goal ratings were zero; some may reflect inability to judge rather than a negative assessment.

5 Concluding Remarks

This paper presented Strategy-Aware Confidence, a propagation model that treats GSN strategy nodes as explicit sources of uncertainty via expert-elicited per-node mean and variance. A strategy exponent α, calibrated from expert holistic ratings, controls the discounting strength and yields diagnostic gaps at intermediate goals. Combined with the Consensus Score, the model produces a two-dimensional dashboard that separates stakeholder acceptance from technical validity. We demonstrated the framework on a SOTIF assurance case for a planned L4 automated driving service in Shiojiri, Japan.

Future work includes formal defeater integration [11,2], iterative assessment cycles to track score shifts as the case matures, and extending our consensus-confidence framework to external stakeholders. Although this study targets L4 automated driving, the approach applies to other safety-critical domains where diverse stakeholders must agree on assurance arguments.

References

1. Barrett, S., Fox, P., Krook, J., Mondal, T., Mylius, S., Tlaie, A.: Assessing confidence in frontier AI safety cases (2025), arXiv preprint arXiv:2502.05791
2. Bloomfield, R., Rushby, J.: Confidence in Assurance 2.0 Cases. In: The Practice of Formal Methods: Essays in Honour of Cliff Jones, Part I. LNCS, vol. 14780, pp. 1–23. Springer (2024)
3. Denney, E., Pai, G., Habli, I.: Towards measurement of confidence in safety cases. In: ESEM 2011. pp. 380–383. IEEE (2011)
4. Diemert, S., Millet, L., Joyce, J., Weber, J.H.: Including defeaters in quantitative confidence assessments for assurance cases. In: SASSUR 2024 (SAFECOMP Workshops). LNCS, vol. 14989, pp. 267–282. Springer (2024)

5. Diemert, S., Shortt, C., Weber, J.H.: How do practitioners gain confidence in assurance cases? Information and Software Technology **185**, 107767 (2025)
6. Diemert, S., Weber, J.H.: Certus: A domain specific language for confidence assessment in assurance cases. In: SAFECOMP 2025 Workshops. LNCS, vol. 15955. Springer (2025)
7. Favaro, F., Fraade-Blanar, L., Schnelle, S., Victor, T., Peña, M., Engström, J., Scanlon, J.M., Kusano, K.D., Smith, D.: Building a credible case for safety: Waymo's approach for the determination of absence of unreasonable risk. arXiv preprint arXiv:2306.01917 (2023)
8. Goodenough, J.B., Weinstock, C.B., Klein, A.Z.: Eliminative argumentation: A basis for arguing confidence in system properties. Tech. rep., Software Engineering Institute, Carnegie Mellon University (2015)
9. Goodman, L.A.: On the exact variance of products. Journal of the American Statistical Association **55**(292), 708–713 (1960)
10. Graydon, P.J., Holloway, C.M.: An investigation of proposed techniques for quantifying confidence in assurance arguments. Saf. Sci. **92**, 53–65 (2017)
11. Herd, B., Kelly, J., Zacchi, J.V., Heinemann, C., Diemert, S.: Integrating defeaters into subjective logic-based quantitative assurance arguments. In: EDCC 2025. pp. 141–149. IEEE (2025)
12. IEC: IEC 62853:2018 Open systems dependability (2018)
13. ISO: ISO 26262:2018 Road vehicles - Functional Safety - (2018)
14. ISO: ISO 21448:2022 Road vehicles — Safety of the intended functionality (2022)
15. Kodama, H., Matsuno, Y., Takai, T., Ota, H., Okada, M., Tsuchiya, T.: A case study of continuous assurance argument for level 4 automated driving. In: SafeComp 2024. LNCS, vol. 14988, pp. 150–165. Springer (2024)
16. Matsuno, Y.: D-Case Communicator: A web based GSN editor for multiple stakeholders. In: ASSURE 2017. LNCS, vol. 10489, pp. 64–69. Springer (2017)
17. Matsuno, Y., Hayashi, M., Tsuchiya, T.: Consensus building in level 4 automated driving field trials through assurance cases. In: Gallina, B., Törngren, M., Bitsch, F. (eds.) SAFECOMP 2025. LNCS, vol. 15954, pp. 33–47. Springer (2025)
18. Matsuno, Y., Takai, T., Yamamoto, S.: Facilitating use of assurance cases in industries by workshops with an agent-based method. IEICE Trans. Inf. Syst. **103-D**(6), 1297–1308 (2020)
19. Nešić, D., Nyberg, M., Gallina, B.: A probabilistic model of belief in safety cases. Saf. Sci. **138**, 105187 (2021)
20. SAE International: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. SAE Standard J3016_202104 (2021)
21. SCSC Assurance Case Working Group: Goal structuring notation community standard version 3. Tech. Rep. SCSC-141C, Safety-Critical Systems Club (2021)
22. UL: UL 4600: Standard for Safety for the Evaluation of Autonomous Products (2023)
23. Varadarajan, S., Bloomfield, R., Rushby, J., Gupta, P., Murugesan, A., Stroud, R., Netkachova, K., Wong, E.: CLARISSA: Foundations, tools and automation for assurance cases. In: DASC 2023. pp. 1–10. IEEE (2023)
24. Yuan, C., Wu, J., Liu, C., Yang, H.: A subjective logic-based approach for assessing confidence in assurance case. International Journal of Performability Engineering **13**(6), 807–822 (2017)