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

Yutaka Matsuno✉[1][0000−0001−9809−0814],
Michio Hayashi[2], 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. Real-world deployment of SAE Level 4 (L4) automated driving vehicles requires both (i) technical justification that the system is acceptably safe within its operational design domain and (ii) consensus among a broad set of stakeholders who must authorize, operate, and support the deployment. In our previous field-trial study [**?**], we combined a GSN-based assurance case with a plain-language Safety Status Report and an anonymous questionnaire, and proposed a lightweight *Consensus Score* to quantify stakeholder acceptance. However, stakeholder consensus alone cannot distinguish between insufficient technical justification and insufficient shared understanding, and many existing confidence assessment methods are too costly and opaque to serve as a shared decision instrument in time-constrained industrial pilots. This paper extends the consensus-building workflow with *Strategy-Aware Confidence*, an expert-elicited confidence propagation model that treats GSN goals, strategies, and evidence as random variables characterized by a mean and variance. Sub-goal evidence is aggregated via inverse-variance weighting, while parent goals are discounted by the assessed validity of the argument strategy, allowing uncertainty in reasoning patterns to propagate to top-level claims. We keep organizational acceptance (Consensus) and technical validity (Confidence) as separate dimensions and combine them in a 2D decision dashboard that supports targeted interventions (evidence strengthening vs. risk communication). We present the elicitation protocol, the propagation algorithm, and a case-study design for an industrial L4 minibus pilot.

Keywords: assurance case · GSN · confidence assessment · expert judgment aggregation · consensus building · automated driving

1 Introduction

SAE Level 4 (L4) automated driving aims to operate without a human driver within a defined operational design domain, but real-world pilots are inherently *open systems*, where environments, operational assumptions, and hazards evolve over time [16,12]. This openness amplifies uncertainties in both technical and operational risks, making assurance a moving target rather than a one-time artifact. As a consequence, deploying L4 systems requires not only a structured safety argument but also an organizational process for repeatedly negotiating what evidence is sufficient for the next operational step.

Assurance cases (e.g., GSN) provide a disciplined way to structure safety claims, argument strategies, and supporting evidence, and to record counterarguments as *defeaters*. To make such arguments more actionable, the assurance-case community has also developed *confidence assessment methods* (CAMs) that quantify the strength of claims using expert scoring, probabilistic models (including Bayesian networks), or eliminative argumentation [6,11,9]. These methods help reduce overconfidence by exposing missing evidence and weak reasoning steps. In practice, however, two persistent challenges arise in industrial L4 pilots. First, communicating a detailed GSN-based argument to non-safety experts (e.g., business, product, executive stakeholders) is difficult, especially when residual risks and operational constraints must be understood under time pressure. Second, even a technically well-structured assurance case does not directly tell whether the diverse stakeholders *accept* the argument as a basis for go/no-go decisions.

In our SafeComp 2025 case study of an industrial L4 minibus pilot, we therefore paired a GSN-based assurance case with a *Safety Status Report* that re-expresses the argument in plain language and explicitly highlights residual risks and limitations. We then conducted an anonymous questionnaire that rated key goal and strategy nodes on an ordinal (0–3) scale and collected open-ended comments. To summarize organizational acceptance in a lightweight manner, we introduced a *Consensus Score* that combines top-down (holistic) and bottom-up (strategy/sub-goal) ratings into a single indicator. This workflow enabled stakeholder feedback—including critical comments—to be systematically incorporated into the assurance case as defeaters and tracked through iterative revisions.

Feedback from Prof. Lorenzo Strigini highlighted that the aggregation of ordinal judgments is a long-standing problem at the intersection of voting/social choice and expert judgment aggregation. In particular, social choice theory shows that no aggregation rule can satisfy all desirable fairness and rationality axioms simultaneously [1], and the expert judgment literature debates when, and how, experts should be weighted based on performance, calibration, or information [4,3,5]. This implies that a practical consensus metric must be positioned explicitly as a *decision-support heuristic*, with its purpose and assumptions made transparent.

More importantly, the feedback motivated a sharper separation between two distinct questions:

- **Organizational acceptance:** “Do stakeholders, from their respective perspectives, accept this safety claim as a basis for action?”

- **Technical validity:** "Is the argument (structure + evidence) technically persuasive, and how uncertain are we about it?"

Our Consensus Score primarily addresses the former. The latter is the scope of CAMs, but many existing approaches either require heavy probabilistic modeling and extensive parameterization, or treat the argument structure—especially GSN *strategy* nodes—implicitly. In practice, however, the decomposition strategy (i.e., the reasoning pattern that connects sub-goals to a parent goal) can be a critical weak link: a flawed argument strategy cannot be compensated by stacking more evidence underneath it.

This paper proposes *Strategy-Aware Confidence* as a complementary, expert-focused assessment that explicitly models strategy nodes as uncertain. Each GSN node X is treated as a random variable in $[0, 1]$ characterized by a mean μ_X and variance σ_X^2, elicited from a safety/development expert team. Sub-goal confidences are aggregated via inverse-variance weighting, reflecting that more precise (lower-variance) assessments should contribute more. Parent goals are then *discounted* by the assessed validity of the argument strategy, so that uncertainty about the reasoning pattern propagates upward. As a simple illustration, even strong aggregated evidence can be substantially reduced when the argument strategy is judged questionable.

We integrate this expert-based confidence with stakeholder-based acceptance using a *two-dimensional Consensus–Confidence dashboard*. Each assurance-case node is mapped onto (Consensus, Confidence), yielding four decision-relevant quadrants: (i) high–high nodes ready to proceed, (ii) high confidence but low consensus nodes indicating a communication gap, (iii) low confidence but high consensus nodes indicating potential organizational over-optimism, and (iv) low–low "red nodes" requiring prioritized attention. This separation enables targeted interventions: strengthening evidence and argumentation where technical confidence is low, and strengthening risk communication where consensus is low despite strong technical justification.

Research questions. This paper is organized around the following research questions:

- **RQ1:** How can we quantify confidence in GSN-based assurance cases while explicitly accounting for uncertainty in argument strategies?
- **RQ2:** How does separating organizational acceptance (Consensus) from technical validity (Confidence) improve prioritization and decision support in an industrial L4 deployment workflow?

Contributions. Our main contributions are:

- A reframing of the existing Consensus Score as a voting-like indicator of organizational acceptance, explicitly distinguished from technical confidence.
- Strategy-Aware Confidence, an expert-elicited confidence propagation model that treats GSN strategy nodes as first-class uncertain entities.

– An integrated consensus-building process and a 2D dashboard that connects acceptance and validity to concrete intervention strategies.
– A case-study design for applying the method to an industrial L4 minibus pilot, including data to be collected and analyses to answer RQ1–RQ2.

Paper organization. The remainder of this paper is structured as follows. Section 2 reviews related work on confidence assessment and judgment aggregation. Section 3 describes the proposed workflow and the Strategy-Aware Confidence model. Section 4 presents the case-study design and planned analyses. Section 5 discusses limitations and threats to validity. Section 6 concludes and outlines future work.

2 Related Work

Assurance cases provide a structured way to justify a top-level dependability or safety claim by explicitly linking claims, arguments, evidence, and assumptions. Goal Structuring Notation (GSN) is one of the most widely used notations for this purpose, and its community standard clarifies the roles of goals, strategies, solutions, and contextual information in building an argument structure [10]. In safety-critical domains, assurance cases are often developed alongside process and product oriented standards (e.g., ISO 26262 for road vehicles and ISO 21448/SOTIF for intended functionality) [13,14].

For open and evolving systems such as automated driving, assurance cases must support continuous updates and effective communication across diverse stakeholders with varying levels of safety expertise. Industrial and regulatory communication artifacts provide important context for such systems, including engineer-centric templates and frameworks (e.g., UL 4600 and SAFAD) and policy-oriented guidance documents (e.g., NHTSA ADS guidance) [18,2,15]. However, these artifacts primarily focus on structuring and disclosing safety activities and rationales, and they provide limited mechanisms to quantify whether different stakeholder groups actually accept the stated safety claims.

A substantial body of work has studied how to assess and communicate the strength of an assurance case beyond purely qualitative review. Denney et al. propose early steps toward measuring confidence in safety cases, highlighting the need for explicit metrics grounded in evidence and expert judgment [6]. Several Confidence Assessment Methods (CAMs) operationalize this idea through probabilistic or quasi-probabilistic reasoning. Bayesian-network-based approaches model uncertainty in claims and evidence using expert elicitation and conditional probability structures; Fenton and Neil provide methodological foundations and practical guidance for BN-based risk assessment and decision analysis [7]. Other probabilistic treatments include Baconian probabilities for assurance case confidence measurement [19] and models for safety case confidence assessment proposed in the assurance and safety communities [11]. Alternative uncertainty formalisms such as Dempster–Shafer theory have also been used to represent belief and evidence combination when probabilities are difficult to elicit precisely [17].

Eliminative argumentation (and related defeater-based approaches) provides another path: instead of assigning a single probability-like value to a claim, it focuses on systematically identifying, refining, and addressing potential defeaters, thereby strengthening the argument through structured rebuttal and evidence planning [9]. Across these CAMs, a recurring practical challenge is that rigorous parameterization and interpretation can be demanding for non-experts and costly under time constraints, especially in early deployment phases or when empirical datasets are still limited.

When assurance decisions must be made under uncertainty and with incomplete data, eliciting and aggregating judgments from experts and stakeholders becomes essential. Structured expert judgment provides principled methods for elicitation and aggregation, including performance-based weighting approaches such as Cooke's classical model [4]. More broadly, group-based judgmental forecasting research studies how to combine opinions and how to design group processes; it also highlights open questions about when equal weighting is sufficient and when differential weighting is beneficial [21,3,5,20].

From a theoretical standpoint, aggregating preferences or ratings is related to social choice and voting theory, where classical results (e.g., Arrow's impossibility theorem) show that no single aggregation rule can satisfy all desirable fairness criteria simultaneously [1]. In dependability, voting and adjudication have been studied in the context of diverse-redundant systems, where the aggregation rule itself affects system-level reliability and fault tolerance [8]. These strands motivate treating aggregation not as a purely technical detail but as a design choice that should be transparent, justifiable, and tailored to the decision context.

Our previous work [?] introduced a stakeholder-oriented communication process (Safety Status Report + survey) and a lightweight Consensus Score to quantify organizational acceptance of safety claims in an SAE L4 automated driving field demonstration. The present work builds on that foundation by positioning consensus (acceptance) and confidence (argument validity under uncertainty) as complementary yet distinct dimensions, and by integrating them in a decision-support view. This positioning responds to limitations of applying existing CAMs alone when stakeholder groups are broad and heterogeneous, and it leverages insights from both assurance-case confidence research and judgment-aggregation theory to support practical, iterative assurance for complex open systems.

3 Strategy-Aware Confidence Assessment method

In this section, we propose a quantitative framework for assessing confidence in Goal Structuring Notation (GSN) arguments. Our approach explicitly models the *validity of the inference strategy* as a stochastic variable and separates the evidence aggregation process from the strategy assessment. This separation ensures that the final confidence reflects both the quality of evidence and the soundness of the reasoning strategy.

3.1 Modeling Inference Strategy as a Random Variable

In a GSN structure, a parent goal G_{parent} is supported by a set of sub-goals $\{G_1, G_2, \dots, G_n\}$ through a strategy S. We model the strategy node S not merely as a connector, but as a governing variable representing the validity of the inference logic.

The confidence in the strategy is derived from expert elicitation. We define the strategy variable S with mean μ_S and variance σ_S^2 as follows:

$$\mu_S = \mu_{\text{survey}}, \quad \sigma_S^2 = \sigma_{\text{survey}}^2 + \epsilon \tag{1}$$

where μ_{survey} and σ_{survey}^2 are the sample mean and variance obtained from expert elicitation, and ϵ is a regularization term that ensures numerical stability and avoids zero-variance dogmatism, in accordance with Cromwell's rule.

3.2 Sub Goal Aggregation

To aggregate heterogeneous sub-goals $\{G_i\}$ into a single *total evidence* variable E, we employ Inverse Variance Weighting (IVW). Unlike simple averaging, this method assigns greater weight to evidence with lower uncertainty.

Let μ_{G_i} and $\sigma_{G_i}^2$ denote the mean confidence and variance of the i-th sub-goal, respectively. The weight w_i for each sub-goal is defined as the inverse of its variance:

$$w_i = \frac{1}{\sigma_{G_i}^2} \tag{2}$$

The aggregated evidence E is characterized by a weighted mean μ_E and variance σ_E^2:

$$\mu_E = \frac{\sum_{i=1}^{n} w_i \cdot \mu_{G_i}}{\sum_{i=1}^{n} w_i} \tag{3}$$

$$\sigma_E^2 = \frac{1}{\sum_{i=1}^{n} w_i} \tag{4}$$

This formulation ensures that low-quality evidence (i.e., evidence with high $\sigma_{G_i}^2$) contributes minimally to the aggregated result, thereby preventing the artificial inflation of confidence through the accumulation of weak evidence.

3.3 Confidence Discounting via Strategy Validity

Finally, we determine the confidence in the parent goal G_{parent}. We posit that the reliability of the conclusion is conditional on the validity of the strategy used to derive it. Accordingly, we apply a multiplicative discounting model in which the aggregated evidence E is modulated by the strategy variable S:

$$G_{\text{parent}} = E \times S \tag{5}$$

Assuming independence between the inherent quality of evidence and the validity of the strategy, the expected confidence μ_G is given by:

$$\mu_G = \mu_E \cdot \mu_S \tag{6}$$

This formulation implies that even if the evidence is perfect ($\mu_E \approx 1$), a flawed strategy ($\mu_S < 1$) will inevitably reduce the final confidence.

To capture the full spectrum of uncertainty, we apply Goodman's formula for the variance of the product of two independent random variables:

$$\sigma_G^2 = \mu_E^2 \sigma_S^2 + \mu_S^2 \sigma_E^2 + \sigma_E^2 \sigma_S^2 \tag{7}$$

This variance σ_G^2 accounts for three sources of uncertainty:

1. Uncertainty in the strategy, scaled by the evidence strength ($\mu_E^2 \sigma_S^2$);
2. Uncertainty in the evidence, scaled by the strategy validity ($\mu_S^2 \sigma_E^2$); and
3. A higher-order interaction term ($\sigma_E^2 \sigma_S^2$), representing the compound risk when both the strategy and the evidence are uncertain.

This approach provides a conservative and rigorous estimation suitable for safety-critical contexts.

Acknowledgments

This research is based on results obtained from a project subsidized by the New Energy and Industrial Technology Development Organization (NEDO).

References

1. Arrow, K.J.: Social Choice and Individual Values. Yale University Press, 2 edn. (1963), originally published in 1951; introduces Arrow's impossibility theorem
2. Automated Driving Safety Consortium: Safety first for automated driving (2019)
3. Bolger, F., Rowe, G.: The aggregation of expert judgment: do good things come to those who weight? Risk Analysis **35**(1), 5–11 (2015). https://doi.org/10.1111/risa.12272, discussion of equal vs. differential weighting for expert judgment aggregation
4. Cooke, R.M.: Experts in Uncertainty: Opinion and Subjective Probability in Science. Oxford University Press (1991), discusses structured expert judgment, scoring rules, calibration/information and the classical model
5. Cooke, R.M.: The aggregation of expert judgment: do good things come to those who weight? Risk Analysis **35**(1), 12–15 (2015). https://doi.org/10.1111/risa.12353, commentary/response regarding performance-based weighting (classical model)
6. Denney, E., Pai, G., Habli, I.: Towards measurement of confidence in safety cases. In: ESEM 2011. pp. 380–383. IEEE (2011)
7. Fenton, N., Neil, M.: Risk Assessment and Decision Analysis with Bayesian Networks. CRC Press (2012), foundational text on the Ranked Nodes method and expert elicitation in BNs

8. Giandomenico, F.D., Strigini, L.: Adjudicators for diverse-redundant components. In: SRDS 1990. pp. 114–123. IEEE Computer Society (1990). https://doi.org/10.1109/RELDIS.1990.93957, `https://doi.org/10.1109/RELDIS.1990.93957`
9. 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)
10. GSN contributors: GSN community standard version 1.0 (2011), `http://www.goalstructuringnotation.info`
11. Guiochet, J., Hoang, Q.A.D., Kaâniche, M.: A model for safety case confidence assessment. In: SafeComp 2015. pp. 313–327 (2015)
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. NHTSA: Automated driving systems 2.0: A vision for safety. Tech. rep., U.S. Department of Transportation (2017)
16. SAE International: Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles. SAE Standard J3016_202104 (2021)
17. Shafer, G.: A Mathematical Theory of Evidence. Princeton University Press (1976), the seminal work on Dempster-Shafer theory and belief discounting
18. UL: UL 4600: Standard for Safety for the Evaluation of Autonomous Products (2023)
19. Weinstock, C.B., Goodenough, J.B., Klein, A.Z.: Measuring assurance case confidence using baconian probabilities. In: ASSURE 2013. pp. 7–11. IEEE (2013)
20. Winkler, R.L.: Equal versus differential weighting in combining forecasts. Risk Analysis **35**(1), 16–18 (2015). https://doi.org/10.1111/risa.12302
21. Wright, G., Rowe, G.: Group-based judgmental forecasting: An integration of extant knowledge and the development of priorities for a new research agenda. International Journal of Forecasting **27**(1), 1–13 (2011). https://doi.org/10.1016/j.ijforecast.2010.05.012, editorial for the special issue on group-based judgmental forecasting