

Assessing Microscopic Traffic Simulation Confidence Level Through Expert-Weighted AHP and Belief Theory: An Application to CAV Testing

Hugues Blache^{1,*}, Pierre-Antoine Laharotte^{2,*}, Nour-Eddin El Faouzi^{2,3}

Cite this article: <https://doi.org/10.26599/JICV.2026.9210107>

ABSTRACT: This paper addresses the challenge of assessing experts' confidence in simulation tools for reproducing safety-related indicators in traffic scenario-based testing. While simulation offers major advantages in terms of safety, cost, and repeatability, its outputs do not equally reflect reality across all traffic situations. The objective is therefore to identify which scenarios can be considered high confidence in simulation. The proposed approach relies on several Analytic Hierarchy Process (AHP) approaches combined with an expert survey on microscopic traffic simulation. Scenarios are represented through ontology-based attributes, and confidence weights are assigned through pairwise comparisons. Three fusion strategies are compared: classical AHP, weighted AHP based on expertise level, and AHP combined with belief theory. The results show that accounting for expertise improves the confidence scores and leads to more consistent rankings of scenarios. Dynamic traffic and environmental conditions are generally associated with lower confidence than static road-related attributes. Scenario score distributions and Gaussian mixture analysis further reveal distinct groups of low-, medium-, and high-confidence scenarios. These results provide a decision-support framework for distinguishing scenarios associated with higher expert-assigned confidence from those that may require real-world testing.

KEYWORDS: Scenario-based testing, Analytic Hierarchy Process, simulation confidence, microscopic traffic simulation, expert survey, belief theory.

1 Introduction

Scenario-based validation approaches for connected and automated vehicle systems typically involve two main steps: generating a comprehensive scenario universe and selecting a relevant subset for testing. This selection step highlights the importance of scenario representation (e.g., traffic situation coverage and diversity) and the influence of the selection strategy (probabilistic or non-probabilistic) on the types of scenarios retained. The resulting subset is then used for acceptance testing and system evaluation. This evaluation can be performed through two complementary approaches: field testing and simulation-based testing. Simulation offers significant advantages, including the ability to reproduce scenarios multiple times without physical risk and reduced costs, depending on the X-in-the-Loop (XiL) level. As a result, industrial and regulatory strategies increasingly rely on simulation to maximise test coverage. Simulation tools remain simplified representations of reality and typically rely on a limited set of indicators. They therefore cannot fully capture real-world complexity, which can affect the reliability of simulation outputs. A key challenge is therefore to determine which scenarios can be reliably assessed in a virtual environment and which require real-world testing. This raises the need to identify scenarios for which simulation results can be considered high-confidence. Like scenario

generation, this problem can be addressed through two main approaches:

- A *data-driven approach*, which involves comparing simulation outputs with real-world data to derive confidence metrics. This approach allows for a quantitative assessment of simulation reliability using objective performance functions (Ni, 2004). However, it is computationally intensive and requires significant effort in terms of calibration and data collection.
- An *expert-based approach*, which assigns confidence levels to scenarios based on domain expertise, often using multi-criteria decision-making methods. This approach is faster and more practical for large scenario sets but may introduce subjective biases.

In this work, we adopt the expert-based approach. This raises the following research question: *What strategies can be implemented to assign confidence levels to simulation tools to reliably reproduce output indicators based on expert judgment?* Following the methodology established in a previous study (Blache et al., 2025), the objective is to assign “confidence labels” or weights to each scenario (layer-based micro traffic situations) reflecting the reliability of the simulation outputs. In

¹ Research Centre for Integrated Transport Innovation (rCITI), School of Civil and Environmental Engineering, University of New South Wales, Sydney NSW 2052, Australia. ² EMob-Lab, École Nationale des Travaux Publics de l'État (ENTPE), Université Gustave Eiffel, Bron 69675, France. ³ ISAP+D, Mohammed VI Polytechnic University, Ben Guerir 43150, Morocco.

✉ Corresponding authors. E-mail: H. Blache, h.blache@unsw.edu.au; P. Laharotte, pierre-antoine.laharotte@univ-eiffel.fr

Received: May 4, 2026; Revised: August 28, 2026; Accepted: October 10, 2026

© The Author(s) 2026. This is an open access article under the terms of the Creative Commons Attribution 4.0 International License (CC BY 4.0, <http://creativecommons.org/licenses/by/4.0/>).

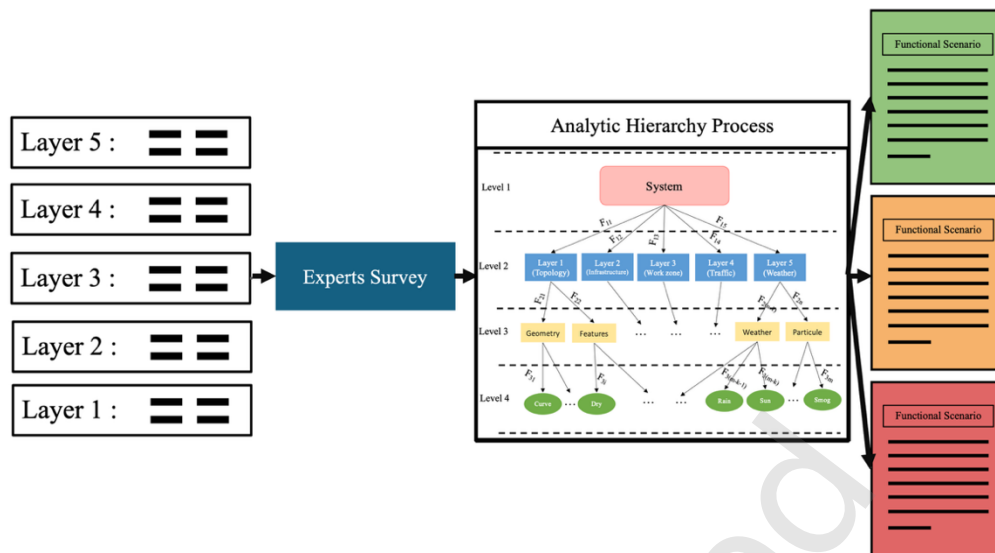


Figure 1. Overview of the Methodology of Attribution of Confidence Weight of Scenarios

this study, we focus on safety indicators (e.g., Time-To-Collision). To estimate these confidence levels, an survey was conducted involving researchers and experts in microscopic traffic simulation. The responses were integrated into a decision-making framework based on the Analytic Hierarchy Process (AHP), which provides confidence levels for the ability of simulation tools to reproduce output indicators for each scenario, as illustrated in Figure 1. The main contributions of this work are:

- A structured expert-survey framework for assigning simulation confidence to ontology-based CAV test scenarios,
- a comparative evaluation of three aggregation strategies on the same fixed dataset, classical AHP, expertise-weighted AHP, and AHP coupled with Dempster-Shafer belief theory, clarifying how each formulation of the hybrid method handles respondent heterogeneity and epistemic uncertainty;
- a decision-support classification of scenarios into confidence groups, supporting the identification of scenarios suitable for simulation-based versus real-world testing; and
- open replication materials (survey data, AHP/belief-theory code, and sensitivity analyses) on GitHub repository.

2 Literature Review

The question of how far simulation outputs can be confident to reflect real-world traffic behaviour has been investigated since the early development of microscopic traffic simulators. Traditional approaches rely on calibration and validation frameworks that compare simulated outputs with empirical observations, typically using statistical fit measures and parameter adjustment procedures to reproduce traffic indicators (Ni et al., 2004; Punzo & Simonelli, 2005; Barceló, 2010). More recent work has incorporated sensitivity analysis to identify influential parameters and reduce model complexity (Ciuffo et al., 2014; Punzo et al., 2014). While effective for reproducing aggregate traffic behaviour, these data-driven approaches require extensive empirical datasets and remain computationally demanding, limiting their scalability to the large range of scenarios required for Connected and Automated Vehicles (CAV) acceptance testing. In the context of scenario-based validation, frameworks such as PEGASUS have been developed to support systematic and large-scale testing of automated systems (Kalra & Paddock, 2016; Menzel et al., 2018). These approaches emphasise scenario coverage and criticality but do not explicitly quantify the confidence that can be assigned to simulation outputs across diverse traffic situations. Moreover, simulation confidence

remains an open challenge, as simulation environments may fail to capture the physical complexity of rare or extreme conditions (Riedmaier et al., 2020). This limitation motivates the development of complementary approaches, including expert-based methods, to assess simulation confidence prior to large-scale testing.

In this context, and in the absence of exhaustive data-driven validation, decision-support methods are required to structure and formalise expert knowledge for assessing simulation reliability. Multi-Criteria Decision-Making (MCDM) provides a relevant framework. According to (Taherdoost et al., 2023), MCDM is a decision-making approach that aims to determine the best alternative by considering multiple criteria in the selection process. More broadly, MCDM addresses problems involving several, often conflicting, objectives and supports decision-makers in identifying the most suitable compromise solution. The generic steps of MCDM involve identifying and selecting relevant criteria, determining their relative importance (weights), and ranking alternatives using appropriate decision-making methods.

To structure these steps, MCDM approaches are generally divided into two main categories: Multiple-Objective Decision-Making (MODC) and Multiple-Attribute Decision-Making (MADM) (Gebre et al., 2021). MODC is suited for continuous optimisation problems with an infinite set of alternatives, typically solved using mathematical programming or heuristic optimisation techniques. In contrast, MADM addresses discrete problems with a finite set of alternatives, where options are evaluated, ranked, or prioritised based on multiple criteria. Within MADM, several families of methods can be identified (Aruldoss et al., 2013). *Value-based approaches* rely on utility functions (Vargas, 1987) to evaluate alternatives. *Distance-based methods*, such as TOPSIS, rank alternatives according to their proximity to an ideal solution. *Outranking methods*, such as ELECTRE, eliminate less favourable options based on pairwise dominance relations. Among these approaches, *pairwise comparison methods*, particularly the AHP, are widely used to incorporate expert judgment. AHP enables the comparison of attributes within a hierarchical structure and the derivation of relative weights through judgment matrices. This makes it particularly suitable for problems involving qualitative criteria and expert-driven evaluations. However, a key limitation of classical AHP lies in its sensitivity to uncertainty and variability in expert judgments, which may affect the robustness of the resulting weights.

Beyond classical MCDM methods, several studies have proposed coupling AHP with uncertainty-management

frameworks to address the variability and imprecision inherent in expert judgments. Rikhtegar et al. (2014) combine AHP with the Dempster-Shafer theory of evidence for environmental impact assessment, showing that belief functions allow a more principled treatment of conflicting expert opinions than simple averaging. In the transportation domain, (El Faouzi et al., 2009, 2011) have demonstrated the effectiveness of Dempster-Shafer fusion for combining heterogeneous data sources with different reliability levels, drawing a productive analogy between sensor reliability and respondent expertise. More recently, Çelikbilek et al. (2026) applied grey AHP to autonomous vehicle adoption, incorporating interval judgments and stakeholder weights. This highlights AHP's compatibility with uncertainty modelling, while addressing imprecision rather than explicit ignorance or source conflict. Building on these works, the present study extends this fusion paradigm to the specific problem of aggregating expert judgments on simulation confidence, explicitly modelling respondent expertise as a discounting factor within the belief framework.

In our context, detailed in (Blache et al., 2025), functional scenarios are defined through an abstract linguistic description of traffic situations (named scenario for the rest of the article). This characteristic makes continuous parameter analysis difficult, and consequently, a Multiple-Objective Decision-Making (MODC) approach is not well-suited for obtaining meaningful results. Furthermore, as shown in (Blache et al., 2025), the universe of scenarios is relatively large, making the rapid evaluation of attribute combinations challenging. This complexity limits the applicability of distance-based and outranking approaches, which can become computationally intensive in such contexts. Recent studies have applied AHP-based frameworks to the evaluation of autonomous driving test scenarios. Gao et al. (2019) propose a five-layer AHP model to evaluate scenario complexity, highlighting the role of hierarchical representations in structuring scenario-based assessment. Zhang and Tak (2024) propose a modified AHP method to assess the risk level of test scenarios for autonomous vehicles, combining pairwise comparisons with fuzzy logic to handle uncertainty in expert judgments. Li et al. (2025) extend this line of work by integrating AHP with entropy-weighted TOPSIS to perform a multidimensional evaluation of test scenarios, accounting for both qualitative expert input and quantitative performance data.

This trend persists in recent work, using AHP-based risk indices to generate critical test scenarios (Zhang et al., 2025) and compare automated with conventional traffic via aggregated KPIs (del Cacho Estil-les et al., 2026). These contributions demonstrate the relevance of AHP for scenario evaluation in the CAV domain. They primarily focus on scenario complexity or risk, rather than the confidence that can be placed in simulation outputs for reproducing safety-critical indicators. To the best of our knowledge, applying AHP to quantify simulation confidence in the context of CAV acceptance testing therefore represents a new contribution. Existing approaches target the evaluation of scenarios rather than the reliability of the simulation environment and do not account for variability in respondent expertise or uncertainty in expert judgments. No studies simultaneously address (i) heterogeneity in respondent expertise through differential weighting, (ii) epistemic uncertainty via belief theory, and (iii) the challenges associated with a large ontology-based scenario space.

In contrast, the proposed approach use a pairwise comparisons at the attribute level within the scenario ontology, enabling the assignment of confidence levels to simulate conditions (e.g., comparing the realism of simulations under different weather conditions). These comparisons can then be used to reconstruct scenarios and derive confidence weights associated with simulation outputs. To this end, we develop an AHP-based methodology supported by an survey, in which responses are

weighted according to the expertise of participants. This survey-based approach enables broader expert participation while improving the robustness and representativeness of the results, despite the inherent variability in respondent backgrounds, three response fusion techniques are considered:

- A classical AHP approach, which does not differentiate between respondents' expertise levels, denoted Ω_{sf}^{C-AHP}
- A weighted AHP approach, which assigns weights based on respondents' expertise, denoted Ω_{sf}^{W-AHP}
- An AHP approach combined with belief theory, which incorporates degrees of belief based on expertise, denoted Ω_{sf}^{B-AHP}

The following section presents the methodological framework developed to operationalise these principles, covering survey design, response fusion strategies, and scenario scoring.

3 An Analytic Hierarchy Process Methodology

Developed by Professor Saaty, the AHP is a branch of multi-criteria decision-making. This technique can be divided into different stages (Saaty, 1987; Vaidya & Kumar, 2006). For this article, these stages can be presented in four steps, as defined in Figure 2:

- **Step 1 - Problem Definition:** This step defines the problem framework by identifying the elements to be compared.
- **Step 2 - Survey Construction:** This step identifies the population to be selected to address our problem and constructs a survey accordingly.
- **Step 3 - Model Construction:** This step develops the mathematical model to address the problem, including all equations and the fusion of different responses.
- **Step 4 - Scenario Score:** This final step determines the confidence score of each scenario, noted $\mathcal{S}_f$.

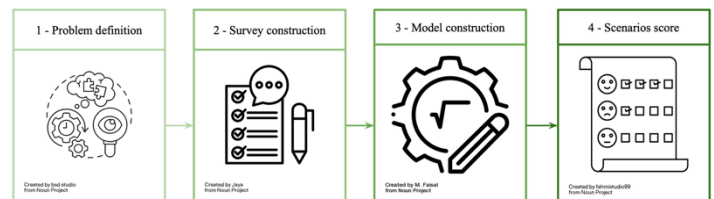


Figure 2. Steps for constructing the methodology to qualify scenarios based on their confidence in simulation tools using AHP

3.1. Problem Definition

Formulating the AHP problem requires defining three key elements: the study objective, the criteria used to evaluate the scenarios, and the method for assigning weights to these criteria. The objective of this study is to assign confidence indicators to each scenario to assess the reliability of simulation tools in reproducing criticality indicators. Within the context of acceptance testing for CAV validation, the goal is to categorise functional scenarios according to their level of confidence when evaluated in microscopic traffic simulation environments, distinguishing between high, medium, and low-confidence scenarios. The evaluation criteria correspond to the attributes (i.e., descriptive words) defining each scenario. These attributes are structured into hierarchical layers that reflect different aspects of traffic situations.

As illustrated in Figure 3, attributes belong to distinct categories, and their influence on simulation confidence may vary across layers. For instance, environmental conditions (Layer 5) may impact simulation reliability differently than road topology (Layer 1). Following previous work (Gao et al., 2019), these attributes can be organised into a hierarchical tree, where higher-level factors influence lower-level attributes.

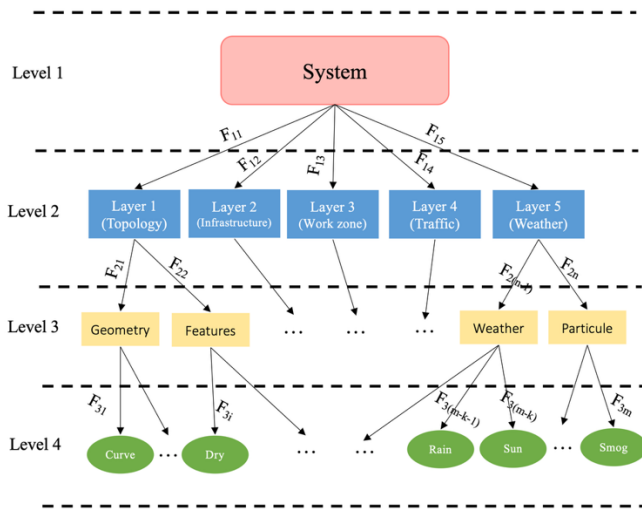


Figure 3. BSI PAS 1880:2020 framework concept adapted into a five-layer structure.

Formally, a factor can be expressed as a set of sub-factors across layers:

$$F_{ij} = \{W_{i,k}, k \in \Omega_{ij}\}, (i, j) \in \Omega_{end}$$

where F_{ij} denotes the j^{th} factor of layer i , and Ω_{ij} represents the set of subordinate sub-factors. At the attribute level (end-node), factors correspond to end-node attributes, which define the values composing a scenario. To assign weights to these attributes, the AHP relies on pairwise comparisons between elements at the same hierarchical level. These comparisons are performed using a judgment scale reflecting the relative importance of attributes with respect to the objective, here defined as confidence in simulation outputs. For example, respondents may be asked to compare the reliability of simulation results under different environmental conditions, such as *sunny* versus *snowy* weather. Although (Saaty, 2004) originally proposed a scale ranging from 1 to 9, previous studies (Schwartz, 2015) have shown that large scales may introduce cognitive bias. Therefore, a reduced scale is adopted in this study (Table 1) to improve the consistency and reliability of responses.

Table 1: Judgment Scale for the Pairwise Comparison between Factors

Importance value	Linguistic scale	Explanation
1	Equal Importance	Both attributes have equal importance for the objective
3	Moderate Importance	The 1st attribute has some importance for the objective compared to the 2nd attribute

5	Strong Importance	The 1st attribute has strong importance for the objective compared to the 2nd attribute
7	Extremely Importance	The 1st attribute has extreme importance for the objective compared to the 2nd attribute

These three questions establish the framework of the problem to be investigated. The next step is to construct our questionnaire to collect the answers to our questions.

3.2. Survey Construction

The principle of AHP is to perform pairwise comparisons between elements within each layer. However, as presented in (Blache et al., 2025), our ontology is too extensive to be fully evaluated through pairwise comparisons in a practical and parsimonious way for respondents. Indeed, many questions can negatively affect response quality, whereas shorter questionnaires tend to improve response reliability (Sharma et al., 2022). To address this issue, we divided the ontology into smaller subsets and designed multiple shorter questionnaires. The overall survey is structured into three main steps and illustrated in the Figure 6:

Step 1 - Expertise Assessment: In the first step, respondents are asked to evaluate their level of expertise in microscopic traffic simulation tools (see Figure 4). The targeted participants include individuals from academia and research and development institutions worldwide, with varying levels of experience in using such tools. To preserve anonymity while still capturing expertise levels, respondents are asked the following question: “Regarding microscopic traffic simulation tools, what is your level of expertise?”

Figure 4: Expertise Levels Question

A second question is then asked to guide respondents

toward a specific branch of the questionnaire in the next step.

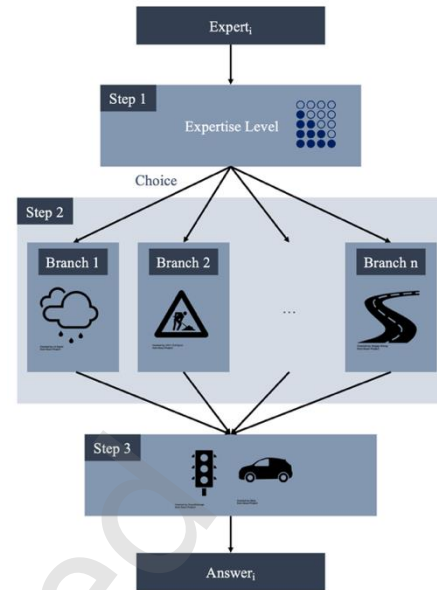
- **Step 2 - Attribute Evaluation and Pairwise Comparison:** The second step consists of assignment a shorter questionnaire to a participant depending on the branch they selected. Each questionnaire corresponding to a specific subset of the ontology. Although the content differs across branches, the structure remains consistent. In each branch: The respondent assigns a value (β) to each attribute based on their level of confidence. The respondent performs pairwise comparisons between attributes. A final question involves comparing two predefined scenarios to capture higher-level comparisons within the ontology. Seven branches were selected for the study: Driver's vision (Illuminance scenario), Sunny (Weather scenario), Type of user (Vehicle type and crossing elements), Intersection (Intersection scenario) Specific Lane (Lane type scenario), Exit (Convergence scenario), Curved Road (Road geometry scenario). Figure 5 presents an example of a question from the Sunny branch, illustrating how the value (β) is assigned to evaluate the scenario. After completing this step, the respondent proceeds to the final stage. Direct confidence value β : respondents assign β on a normalised [0,1] scale (0 = no confidence, 1 = full confidence that simulation reproduces safety indicators for the attribute condition). β encodes attribute presence in a scenario, whereas AHP weights W_i reflect relative simulation confidence among attributes. Pairwise comparisons assess relative simulation confidence, not generic importance.

	Externely confident	Strongly confident	Moderately confident	Low confidence	Not confident
No disturbance	☐	☐	☐	☐	☐
Windy	☐	☐	☐	☐	☐
Rainy	☐	☐	☐	☐	☐
Snowy	☐	☐	☐	☐	☐

Figure 5: Example of a Question Asked to the Respondent Concerning the Sunny Branch to Find a β value

- **Step 3 - Scenario Comparison:** The final step requires respondents to compare four pairs of scenarios: Highway vs. Urban Car-following vs. Lane-changing Free-flow vs. Congested Straight Road vs. Turning

Figure 6. Flowchart of the survey in three steps for a single respondent



Once the survey has been designed and deployed, the next step is to construct the mathematical model used to assign weights to each attribute

3.3. Model for Constructing and Determining Attribute Values

MCDM methods often require the intervention of experts to achieve a given objective. They rely on opinions and comparisons between elements, sometimes with a high level of precision, to ensure confidence in the decision-making process. However, achieving such precision and granularity in comparisons can be challenging for humans, as it may involve significant cognitive biases, leading to uncertainties. According to (Klir and Yuan, 1995), and further elaborated by (Sadiq et al., 2009), these uncertainties mainly arise from two sources: vagueness and ambiguity:

- **Vagueness (imprecision):** Perceptions lack clear and precise distinctions.
- **Ambiguity (belief):** There is an unclear distinction between alternatives and choices. This ambiguity can be divided into two components: *Conflict (discord)*: Disagreement among multiple alternatives. *Non-specificity*: Situations where alternatives are not clearly specified.

By analogy with data fusion approaches in the transport literature, respondents can be treated as sensors. Their expertise is then interpreted as the reliability of each sensor (Figure 7). To address this issue, three approaches are compared: Classical averaging (AHP), denoted Ω_{sf}^{AHP} , Weighted averaging (W-AHP), denoted Ω_{sf}^{W-AHP} , Belief theory (B-AHP), denoted Ω_{sf}^{B-AHP}

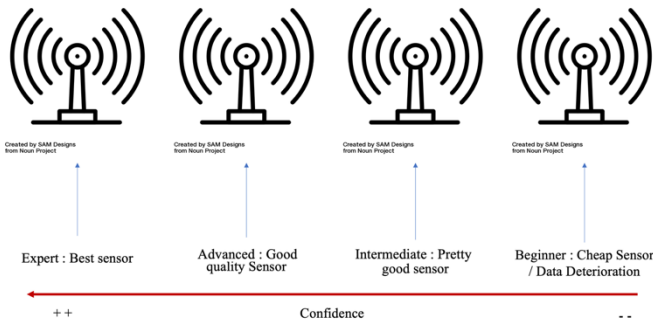


Figure 7. Analogy between levels of expertise and sensor reliability

3.4. Classical AHP Approach – Ω_{sf}^{C-AHP}

The pairwise judgment values defined in the first step and collected through the survey are used to construct a judgment matrix $M^{(k)}$ for each respondent k :

$$M^{(k)} = \begin{bmatrix} m_{11}^{(k)} & m_{12}^{(k)} & \cdots & m_{1n}^{(k)} \\ \vdots & \vdots & \ddots & \vdots \\ m_{n1}^{(k)} & m_{n2}^{(k)} & \cdots & m_{nn}^{(k)} \end{bmatrix}$$

where $m_{ij}^{(k)}$ represents the judgment value between attribute i (i.e., word i) and attribute j , with the reciprocal property:

$$m_{ij}^{(k)} = \frac{1}{m_{ji}^{(k)}}$$

The geometric mean of each row i is defined as:

$$U_i^{(k)} = \left(\prod_{j=1}^n m_{ij}^{(k)} \right)^{1/n}$$

The local weight $w_i^{(k)}$ of attribute i for respondent k is then computed as:

$$w_i^{(k)} = \frac{U_i^{(k)}}{\sum_{j=1}^n U_j^{(k)}}$$

In the classical AHP formulation, all respondents are considered equally reliable, meaning that no weighting is applied based on expertise. The overall weight of each attribute is therefore computed as the geometric mean of all respondents' weights:

$$W_i^{C-AHP} = \left(\prod_{k=1}^n w_i^{(k)} \right)^{1/n}$$

where $w_i^{(k)}$ is the weight assigned to attribute i by respondent k and W_i^{C-AHP} represents the average confidence level associated with attribute i .

3.5. Weighted AHP Approach – Ω_{sf}^{W-AHP}

In this approach, the responses are weighted according to the expertise level of each respondent. The first question of the survey evaluates respondent expertise, allowing the assignment of an influence weight α_k to each respondent k , as shown in Table 2.

Table 2. Assigned weights α_k according to expertise level

Expertise Level	α_k
No expertise	0.1
Beginner level	0.3
Intermediate level	0.5
Advanced level	0.7
Expert level	1.0

A sensitivity analysis of alternative expertise weights α_k is provided in the GitHub replication package (sensitivity_expertise_weights.py); scenario rankings remain highly correlated across the tested schemes. The confidence level of each attribute i is then computed using a weighted geometric mean of respondents' weights:

$$W_i^{W-AHP} = \left(\prod_{k=1}^n (w_i^{(k)})^{\alpha_k} \right)^{\frac{1}{\sum_{k=1}^n \alpha_k}}$$

where: $w_i^{(k)}$ is the weight assigned to attribute i by respondent k , α_k represents the expertise level of respondent k , and W_i^{W-AHP} is the weighted confidence level of attribute i .

3.6 AHP Coupled with Belief Theory Approach – Ω_{sf}^{B-AHP}

Belief theory, also known as Dempster–Shafer theory, aims to account for uncertainty in responses using belief functions (Shafer, 1992). This theory has been widely applied in various fields, particularly in transportation for data fusion (El Faouzi et al., 2011; El Faouzi et al., 2009). Let R be the set of possible responses to a question. The power set 2^R represents all subsets of R , including the empty set $\emptyset$. For example, if $R = \{r_1, r_2, r_3, \dots, r_n\}$, then:

$$2^R = \{\emptyset, \{r_1\}, \dots, \{r_n\}, \{r_1, r_2\}, \dots, R\}$$

A mass function (or Basic Belief Assignment, BBA) is defined as:

$$m: 2^R \rightarrow [0,1]$$

such that:

$$\sum_{A \subseteq R} m(A) = 1$$

where A is a subset of R , and $m(A)$ represents the belief assigned to A . For a respondent k , this is denoted as $m_k(A)$. However, not all responses are equally reliable. To account for this, Shafer introduced a discounting factor ξ_k , which reflects the reliability of respondent k :

$$\begin{cases} m_k^{\xi_k}(A) = (1 - \xi_k) m_k(A) \\ m_k^{\xi_k}(R) = (1 - \xi_k) m_k(R) + \xi_k \end{cases}$$

where $\xi_k \in [0,1]$. A fully reliable respondent corresponds to $\xi_k = 0$, while a completely unreliable respondent corresponds to $\xi_k = 1$.

In this study, the discounting factor is derived from the expertise level α_k of respondent k , such that:

$$\xi_k = 1 - \alpha_k$$

For each survey question, the frame of discernment is $\Theta = R$, where $R = \{r_1, \dots, r_n\}$ denotes the set of possible attribute responses. Respondent k 's initial belief prior to discounting is categorical: $m_k(\{r_i\}) = 1$ for the selected response r_i . Applying the discounting scheme above with $\xi_i = 1 - \alpha_k$ yields $m_k^{\xi_i}(\{r_i\}) = (1 - \xi_k) = \alpha_k$ (belief in the chosen response, weighted by expertise) and $m_k^{\xi_k}(\Theta) = \xi_k = 1 - \alpha_k$ (residual uncertainty allocated to the whole frame). Dempster's rule then combines the discounted BBAs of all respondents within the same branch; the pignistic transformation converts the combined mass into BetP probabilities used as attribute-level confidence weights. Unlike weighted AHP, which applies α_k as an exponent in a geometric mean, B-AHP discounts less expert respondents by allocating residual mass to Θ . To combine two sources of information, for example $m_1^{\xi_1}$ and $m_2^{\xi_2}$, Dempster's rule of combination is used:

$$(m_1^{\xi_1} \oplus m_2^{\xi_2})(A) = \begin{cases} \frac{\sum_{B \cap C = A \neq \emptyset} m_1^{\xi_1}(B) m_2^{\xi_2}(C)}{1 - \sum_{B \cap C = \emptyset} m_1^{\xi_1}(B) m_2^{\xi_2}(C)} & \text{if } A \neq \emptyset \\ 0 & \text{if } A = \emptyset \end{cases}$$

For n sources, the combined mass function is:

$$m(A) = \frac{\sum_{B_1, \dots, B_n | \bigcap_{i=1}^n B_i = A} \prod_{i=1}^n m_i^{\xi_i}(B_i)}{1 - \sum_{B_1, \dots, B_n | \bigcap_{i=1}^n B_i = \emptyset} \prod_{i=1}^n m_i^{\xi_i}(B_i)} \quad \text{for all } A \subseteq R, A \neq \emptyset$$

For each question, responses are combined to capture the interaction between all sources. To obtain usable weights, the mass functions are transformed into probabilities using the pignistic transformation, denoted *BetP*. For an attribute w_i :

$$\text{BetP}(w_i) = \frac{1}{1 - m(\emptyset)} \sum_{w_k \in A} \frac{m(A)}{|A|}$$

These probabilities are then interpreted as confidence values for each attribute:

$$\text{BetP}(w_i) \sim W_i^{B-AHP}$$

3.7. Consistency Verification

Once the three approaches for defining confidence weights are established, AHP requires a consistency check to validate the coherence of the responses.

The Consistency Index (CI) is defined as:

$$CI = \frac{\lambda_{\max} - n}{n - 1}$$

where $\lambda_{\max}$ is the largest eigenvalue of the comparison matrix. The Consistency Ratio (CR) is then computed as:

$$CR = \frac{CI}{RI}$$

where RI is the random index defined by (Saaty, 1987), shown in Figure 8. A consistency ratio below 10% is considered acceptable. If this threshold is exceeded, responses are deemed inconsistent and should either be excluded or the questionnaire should be revised.

n	3	4	5	6	7	8	9	10	11	12	13	14	15
RI	0.58	0.90	1.12	1.24	1.32	1.41	1.45	1.49	1.51	1.48	1.56	1.57	1.59

Source: Adapted from Saaty (1987).

Figure 8. Random index defined by Saaty (1987)

Once the weights are established and consistency is ensured, the final step consists of assigning scores to the functional scenarios.

3.8. Determination of the Scenario Score

Initially, conditional probabilities are combined through multiplication when propagating weights from higher-level factors (parent layers) to lower-level attributes. For example, the weight of the attribute *sunny* is conditioned on its belonging to *Layer 5*. However, during the fusion process, i.e., when calculating $\mathcal{S}_{f_i}^{trust}$, we assume independence between parameters in the absence of additional information. Once the weights for each attribute and each layer are determined, the final step consists of assigning a score to each functional scenario. Since higher layers already carry information about confidence, the weights of parent factors influence the weights of their corresponding child factors. Each parent factor is therefore multiplied by its associated child factors, starting from the highest levels of the hierarchy. The value of an attribute i is defined as:

$$V_i = W_i \times \prod_l W_l$$

where W_l represents the weights of the higher-level layers. A confidence score of a scenario can then be defined as:

$$\mathcal{S}_{f_i}^{conf} = [V_{i,j}]_{1 \times N}$$

where $V_{i,j}$ corresponds to the value of the j^{th} factor in the i^{th} scenario, with:

- $i = 1, \dots, M$ (number of scenarios)
- $j = 1, \dots, N$ (number of attributes)

If we define $V^{attribute}$ as the vector of attribute values, the scenario score can be expressed as:

$$\mathcal{S}_{f_i}^{conf} = (v_1, v_2, \dots, v_n) \cdot \begin{pmatrix} \beta_{w_1}^{(i)} \\ \beta_{w_2}^{(i)} \\ \vdots \\ \beta_{w_n}^{(i)} \end{pmatrix}$$

The final score of a functional scenario is computed as:

$$\mathcal{S}_{f_i}^{conf} = \sum_{j=1}^n v_j \times \beta_{w_j}^{(i)}$$

We denote: $\mathcal{S}_{f_i}^{C-AHP}$ as the score obtained using the classical AHP approach. $\mathcal{S}_{f_i}^{W-AHP}$ as the score obtained using the weighted AHP approach. $\mathcal{S}_{f_i}^{B-AHP}$ as the score obtained using AHP combined with belief theory. The steps presented in this section define the overall methodology.

4 Result

The survey responses are anonymous and freely accessible on the GitHub repository (HuguesBlache). For the present study, the data were collected between *January 1, 2024* and *September 1, 2024*, corresponding to a total of 42 respondents. The types of respondents are illustrated in Figure 9a, and the distribution of respondents across the different scenario branches is shown in Figure 9b. According to these results, most respondents have experience with simulation tools: more than 78% of participants report at least an intermediate level, and 26.2% consider themselves experts (i.e., possessing near-complete expertise of traffic microsimulation tools). However, the number of respondents is unevenly distributed across the different branches of the questionnaire. Nevertheless, each branch includes at least one respondent with an intermediate level of expertise (e.g., Driver's Vision), and some branches include respondents with expert-level knowledge (e.g., Specific Lane). It should be noted that certain branches, such as *Curved Road*, have fewer responses. Geographical location and institutional affiliation were not recorded because the survey was anonymous. Expertise levels are self-assessed and may introduce subjective bias, lower-expertise responses are discounted through the weights rather than excluded. Following Saaty (2008), AHP validity depends primarily on expert quality rather than sample size; the study should therefore be interpreted as an exploratory AHP application based on the fixed sample of 42 respondents. However, according to Saaty, AHP does not rely on the same assumptions as traditional survey methods that require large sample sizes (Kotrlík & Higgins, 2001; Hill, 1998). Instead, the validity of AHP depends more on the level of expertise of respondents than on the number of participants (Saaty, 2008). Therefore, a large sample size is not strictly required.

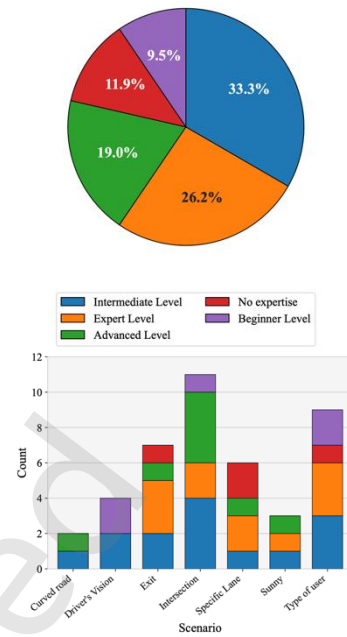


Figure 9. General information about respondents' expertise

(a) Type of expert in the survey

(b) Distribution of responses by scenario branch

Table 6 reports n , mean CR, and median CR for each questionnaire branch. Most branches show acceptable consistency, only the Intersection branch slightly exceeds the mean CR threshold (0.112), while its median CR remains below 0.10 (0.057), consistent with inconsistency concentrated in a subset of responses in the largest branch.

Tab 6. Consistency ratio (CR) by questionnaire branch.

Questionnaire branch	n	Mean CR	Median CR
Intersection	11	0,112	0,057
Type of user	9	0,050	0,053
Exit	7	0,048	0,052
Specific Lane	6	0,087	0,084
Driver's Vision	4	0,022	0,014
Sunny	3	0,052	0,022
Curved road	2	0,099	0,099

The results of this study are divided into two main parts. First, the weights of each attribute are analysed according to the different approaches. Secondly, the distribution of scenario scores is examined across these approaches. This analysis enables the identification of relationships between attributes and the confidence levels of scenarios

4.1. Individual Attribute Weight

Figure 10 presents the weights of each attribute in the ontology tree, distinguishing between the highest layer (Figure 10.a) and the lowest layer (Figure 10.b). The comparison is carried out across three approaches: belief theory (blue), weighted AHP (red), and classical AHP (green). Figure 10.a first highlights a clear distinction between the belief-based and weighted approaches compared to the classical AHP regarding which layers exhibit higher confidence levels. The dynamic parameter layers, namely Layer 5 (weather) and Layer 4 (dynamic road elements), receive lower confidence scores in the weighted and belief approaches than the static layers (Layers 1 to 3). In contrast, the classical AHP

shows the opposite trend. From an expert perspective, this result is coherent: dynamic elements in simulations are more complex and tend to introduce greater uncertainty compared to static elements such as road topology (e.g., their influence on trajectory conflicts). The reversed trend observed in the classical approach may be explained by the inclusion of non-expert respondents, who may misinterpret the impact of dynamic traffic elements (e.g., vehicle behaviour models), overestimating their reliability compared to structural elements. Additionally, Layers 4 and 5 appear to have similar weights in terms of confidence, suggesting that temporal dynamics and road actors have comparable impacts on simulation reliability. Regarding the static layers, and excluding the classical approach, Layer 3 exhibits lower confidence than Layers 1 and 2. This result is consistent with its definition, as it includes work zones, which often involve lane reductions and increased complexity in vehicle interactions (e.g., lane-changing behaviour). At the attribute (end-node) level (Figure 10b), differences between approaches become more pronounced. Some results are intuitive: higher confidence is associated with *sunny* weather conditions, *free-flow* traffic, and *simple geometric road structures*. However, notable discrepancies appear for certain attributes. For instance, confidence is higher for the presence of road signs than for traffic lights. This can be explained by limitations in simulation tools such as SUMO, where vehicles rarely come to a complete stop (speed strictly equal to 0 m/s), potentially affecting the realism of traffic light behaviour. Furthermore, intersection modelling in such tools often relies on predefined trajectories or conflict zones, which may deviate from real-world dynamics. Another divergence between the belief and weighted approaches versus the classical approach is observed in Layer 3 attributes. A clear distinction appears between *no temporary changes* and *lane closure* scenarios. Expert respondents tend to assign lower confidence to lane closures, as they involve increased complexity (e.g., reduced capacity and lane-changing behaviour). In contrast, less experienced respondents may incorrectly assume that fewer lanes imply fewer interactions and thus higher confidence. This simplified reasoning highlights the importance of incorporating expertise weighting in the analysis.

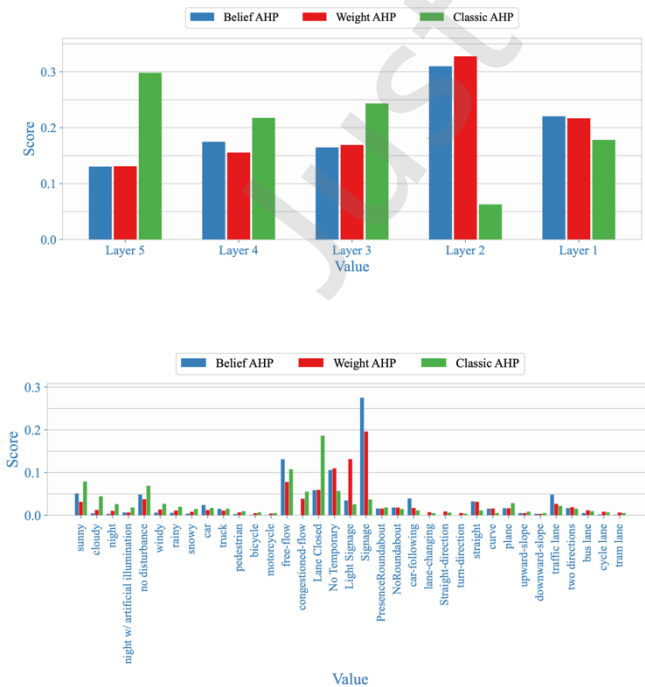


Figure 10. Differences in attribute weights across ontology layers

(a) Ontology layer-level wights (Layer 1-5) (b) Attribute-level (end-node) weights

After analysing the weights of individual attributes, the next step consists of examining the distribution of scenario scores across the different approaches

4.2. Weight Distributions

Figure 11 presents the distribution of scenario scores for each approach: classical AHP (green), weighted AHP (red), and belief theory (blue). The scores are normalised between 0 and 1. A score close to 0 indicates low confidence in simulation outputs, while a score close to 1 indicates high confidence. Across all approaches, the distributions show relatively low densities at the extremes, although values near 0 are more frequent than those near 1. In addition, the distributions' peaks are all below 0.5. This suggests that, overall, confidence in simulation outputs remains limited, which is consistent with the initial hypothesis of this article: *simulation tools, due to their simplified models, tend to diverge from reality*. However, it is important to note that a significant proportion of scenarios still exhibit confidence levels above 0.5, as shown in Table 3. For instance, the first quartile (25%) corresponds to values around 0.28 for the belief and classical approaches, and 0.33 for the weighted approach. Similarly, the median values (50%) are approximately 0.44-0.45 across all methods. The general trends remain comparable between approaches. The boxplot representation (Figure 11.b) highlights that the weighted AHP approach tends to assign slightly higher scores overall compared to the other methods. Despite differences in distribution shapes, the belief and classical approaches exhibit similar coverage levels

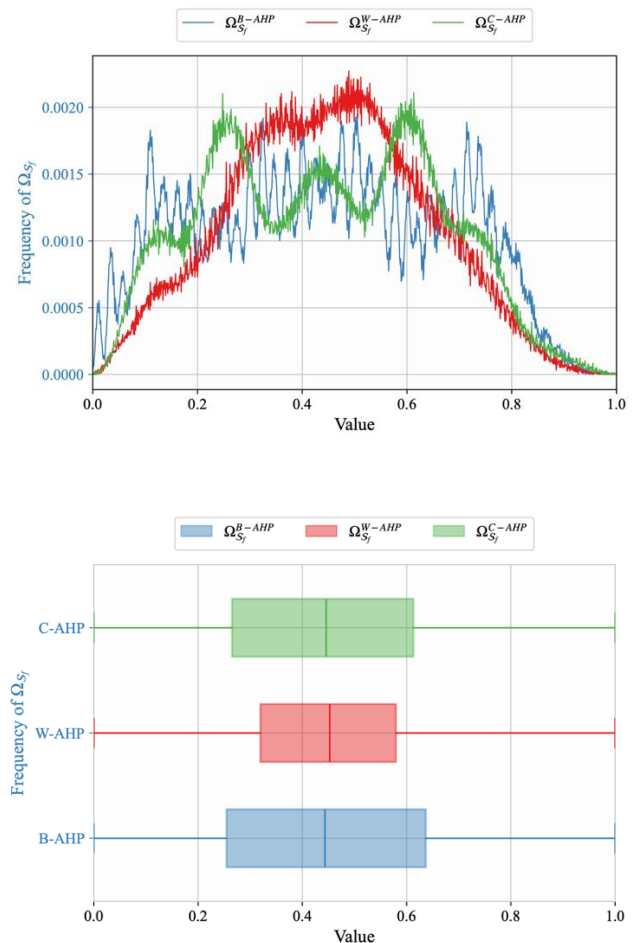


Figure 11. Comparison of score distributions across approaches

- (a) Score distribution for each approach
(b) Boxplot representation of score distributions

Table 3. Confidence level distribution (quartiles)

Percentage of scenarios	10%	25%	50%	75%
Belief	0.17	0.28	0.44	0.61
Weighted	0.22	0.33	0.45	0.58
Classical	0.16	0.28	0.44	0.61

Furthermore, Table 3 suggests the existence of distinct groups of scenarios with varying confidence levels. Some scenarios consistently exhibit high confidence, while others belong to lower-confidence groups. Given the distribution of the classical AHP approach, which does not consider respondent expertise, it is likely that similar patterns exist in the other approaches. The distributions may therefore be interpreted as mixtures of underlying sub-distributions, potentially resembling Gaussian mixture models. This indicates the presence of clusters of scenarios with different confidence levels, where certain groups exhibit higher reliability than others.

4.3. Gaussian Mixture

Gaussian mixture models are a subset of the broader family of mixture models. These models assume that a dataset, represented by a probability density function, is composed of several underlying distributions. When these distributions are Gaussian, the model is called a Gaussian Mixture Model (GMM). This approach is commonly used for clustering, as it enables the identification of groups within a population. In this study, these groups correspond to sets of scenarios with similar confidence levels. Let $X_s = \{x_s^{(1)}, \dots, x_s^{(n)}\} \in [0,1]$ be the sample of scenario scores. Each score is assumed to belong to one of several groups $C_i \in \{C_1, \dots, C_n\}$, each following a normal distribution characterised by a mean μ_i and a covariance matrix Σ_i . The parameters of each distribution are denoted $\theta_i = (\mu_i, \Sigma_i)$, while the global mixture model is defined by parameters Ω , including the mixing proportions π_i . The mixture model is expressed as:

$$\phi(x, \Omega) = \sum_{k=1}^n \pi_k f(x, \theta_k)$$

where $f(x, \theta_k)$ follows a multivariate normal distribution defined as:

$$f(\mathbf{x}, \theta_i) = \frac{1}{((2\pi)^{\frac{n}{2}} |\Sigma|)^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right)$$

To estimate the optimal parameters, the likelihood function is maximised using the Expectation-Maximisation (EM) algorithm:

$$\mathcal{L}(x, \Omega) = \sum_{i=1}^n \log \left(\sum_{k=1}^n \pi_k f(x_i, \theta_k) \right)$$

Each scenario is then assigned to a group based on posterior probabilities:

$$\mathbb{P}(x_i \in C_h) = \frac{\pi_h f(x_i, \theta_h)}{\sum_{k=1}^n \pi_k f(x_i, \theta_k)}$$

The group corresponding to the highest probability is selected. In this study, the GaussianMixture module from the *scikit-learn* library was used for clustering, while *SciPy* was used for visualisation. The optimal number of Gaussian components was determined using the elbow method. The results indicate different optimal numbers of clusters depending on the approach and two trends, the approach giving an optimal number of 3 distributions (Belief approaches), 4 distributions (Weighted approaches) and the approach giving a number of 5 distributions (Classic-AHP). This can be interpreted as: (i) for three Gaussians, we can group scenarios into categories of high confidence, medium confidence, and low confidence. (ii) in the case of 4 Gaussians, we can group scenarios with very high, high, low, and very low confidence. (iii) in the case of 5 Gaussians, we can group scenarios with very high, high, medium, low, and very low confidence. The three values of $\$k\$$ are now explicitly reported as method-specific choices derived from the elbow analysis, and we acknowledge that they do not define a strictly equivalent confidence taxonomy across methods. The clusters are therefore interpreted as low-, medium-, and high-confidence groups based on the ordering of component means, rather than as a harmonised statistical classification common to all three approaches. Beyond this analysis, one can interpret this group as levels of confidence, commonly seen as white, grey, and black confidence zones. Once the groups have been assigned to each scenario, it's useful to look at the nature of the groups by identifying the word groups present in each scenario.

4.4. Word Classification

In this section, the objective here is to examine how clusters are distributed across the attributes (words) used to define the scenarios. Following the same logic (Blache et al., 2025), a scenario can be represented as a bag of words, denoted as $\mathcal{S}_f^{bag} = \{w^1, \dots, w^n\}$, and each scenario $\mathcal{S}_f$ belongs to a cluster C_i . This cluster-attribute analysis is a post-hoc descriptive characterisation of confidence-derived groups, not an independent validation of the weighting framework. Since a given word may appear in multiple scenarios, it is possible to define the probability of occurrence of a word within each cluster, denoted as $\mathbb{P}(word_i | C_j)$. The overall probability of a word can be expressed as:

$$\mathbb{P}(word_i) = \sum_{j=1}^n \mathbb{P}(word_i | C_j) \cdot \mathbb{P}(C_j)$$

where:

$$\mathbb{P}(word_i | C_j) = \frac{\text{Number of occurrences of } word_i \text{ in scenarios belonging to } C_j}{\text{Number of scenarios in } C_j}$$

And $\mathbb{P}(C_j) = \pi_j$ is the GMM mixing proposition. The results are presented in Figure 12 for the three approaches: belief theory (Figure 12a), weighted AHP (Figure 12b), and classical AHP (Figure 12c). The colour coding of the clusters is consistent with the Gaussian mixture distributions presented previously, allowing easier interpretation. For several attributes, similar trends to those observed in Figure 10 can be identified. For example, in the case of road signage, both the belief-based and classical approaches show little variation across clusters. This result is consistent with the relatively high confidence weight associated with this attribute, meaning that it contributes positively across most scenarios regardless of the cluster. A similar pattern is observed for other attributes, confirming the consistency between attribute-level weights and their distribution across scenario clusters.

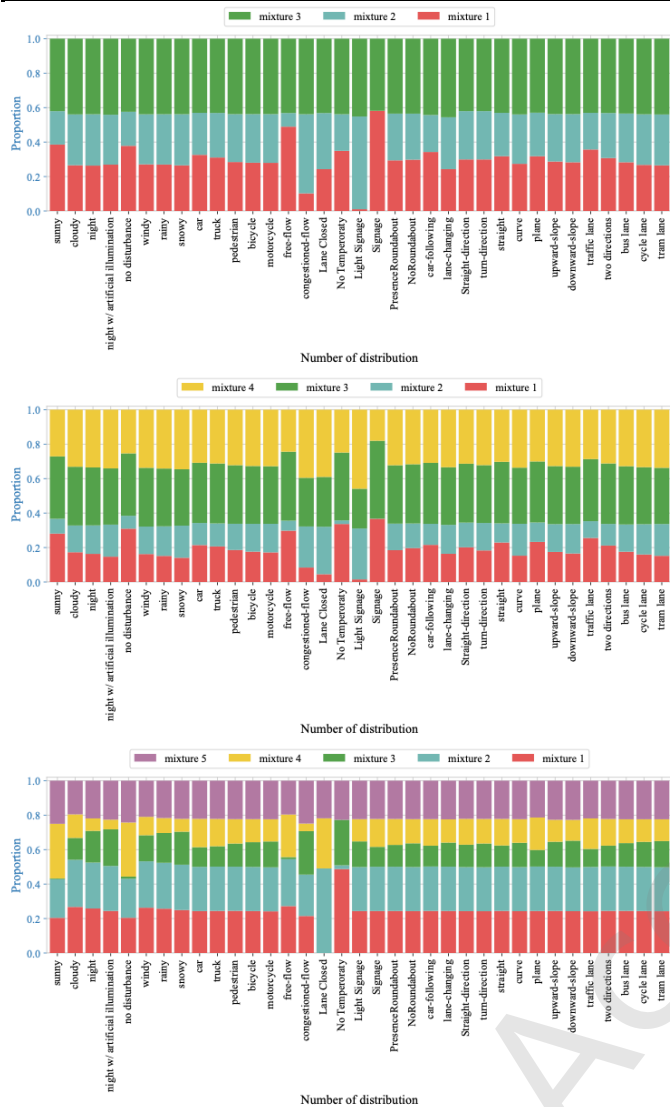


Figure 12. Word distribution across clusters for each approach
(a) Belief-based AHP
(b) Weighted AHP
(c) Classical AHP

4.5. Example of high- and low-confidence Scenarios

Tables 4 and 5 presents, respectively, the scenarios with the highest confidence scores (i.e., score equal to 1) and the lowest confidence scores (i.e., score equal to 0) for the three approaches. For the most high-confidence scenarios, the belief-based and weighted AHP approaches produce identical results. In contrast, the classical AHP approach differs slightly, particularly in Layer 3 and Layer 1, where it includes a *lane closure* and the *presence of a roundabout*. This difference suggests that the classical approach may be less reliable in identifying truly high-confidence scenarios, as it does not account for respondent expertise. For the least high-confidence scenarios, the belief and weighted approaches again show very similar structures. The main difference lies in the illuminance conditions: the belief-based approach identifies *nighttime* as the least high-confidence condition, whereas the weighted approach considers *nighttime with artificial lighting* as less reliable. This distinction highlights subtle differences in how uncertainty and expertise are incorporated into the evaluation process.

Table 4. Most high-confidence scenarios for each approach

Layer / Objective	Belief Approach	Weighted AHP	Classical AHP
Layer 5	Sunny; no disturbance	Sunny; no disturbance	Sunny; no disturbance
Layer 4	Car; free-flow	Car; free-flow	Car; free-flow
Layer 3	No temporary changes	No temporary changes	Lane closed
Layer 2	Signage	Signage	Signage
Layer 1	No roundabout; simple junction; straight; plain; traffic lane	Same as belief	Presence of roundabout; simple junction; straight; plain; traffic lane
Objective	Car-following	Car-following	Car-following

Table 5. Low-confidence scenarios for each approach

Layer / Objective	Belief Approach	Weighted AHP	Classical AHP
Layer 5	Night; snowy	Artificial lighting; snowy	Artificial lighting; snowy
Layer 4	Motorcycle; congested	Motorcycle; congested	Motorcycle; congested
Layer 3	Lane closed	Lane closed	No temporary changes
Layer 2	Light signage	Light signage	Light signage
Layer 1	Roundabout; intersection; curved; downward slope; tram lane	Same as belief	No roundabout; intersection; curved; downward slope; tram lane
Objective	Turning	Turning	Turning

Overall, these results confirm that incorporating expertise (weighted AHP) or uncertainty (belief theory) leads to more consistent and interpretable scenario rankings, whereas the classical AHP approach may produce less expert-consistent outcomes in certain cases. W-AHP and B-AHP yield similar rankings (both use $\alpha_k, \rho = 0.9$), differing mainly in uncertainty propagation. Classical AHP shows essentially no rank agreement with either method ($\rho = -0.10$ and 0.16), confirming that expertise substantially reshapes scenario rankings.

5 Conclusion and Outlook

This article addresses the challenge of assessing the confidence that can be assigned to simulation tools for reproducing safety indicators across various road situations. This work: (1) formulates simulation confidence, rather than scenario complexity or risk, as an AHP objective for CAV acceptance testing over a large ontology-based scenario space, (2) integrates respondent expertise heterogeneity through differential weighting and compares classical, weighted, and belief-theory aggregation on the same expert survey, (3) propagates epistemic uncertainty via Dempster-Shafer discounting and pignistic transformation (B-AHP), and (4) provides a decision-support confidence classification of scenarios

together with open replication materials. The results highlight the critical role of expertise in producing confidence levels, showing that the influence of scenario layers is highly sensitive to the level of expert knowledge considered. The proposed methodology enables the identification of scenarios with high confidence in simulation outputs, which is particularly relevant for the validation of CAV systems. Nevertheless, several limitations must be acknowledged. In particular, some questionnaire branches include only a small number of respondents (3/7 branches with $n = 2-4$), the associated local weights and scenario rankings should therefore be interpreted as exploratory and may be less stable than those derived from larger branches. The confidence scores reported here reflect expert judgments on microscopic traffic simulation in general, at the scenario-attribute level. The reported scores reflect experts' confidence in simulation for given scenario attributes; they do not constitute an empirical validation of simulation performance against real-world observations. Transferability to specific simulation platforms (e.g., SUMO, VISSIM, CARLA) is outside the scope of this paper and is addressed in separate ongoing work. The assumption of independence among scenario attributes underlying the computation of scenario confidence scores. While this assumption simplifies score propagation within the proposed framework, it does not fully reflect real-world traffic situations, where attributes are often coupled. Adverse weather, for example, is frequently accompanied by reduced visibility, while congested traffic tends to be closely associated with more frequent lane changes and greater interaction complexity. Neglecting such dependencies can lead to under- or overestimation of simulation confidence for certain scenario combinations. Future work should therefore explore dependency-aware approaches, such as Bayesian networks, to enhance the robustness of confidence assessment. In addition, the survey design and the use of visual representations may have introduced biases, while the assumption of equal similarity weights was a simplifying choice that could be refined in future work. Furthermore, self-assessment of expertise may lead to inaccuracies, notably due to cognitive biases such as the Dunning-Kruger effect (Dunning 2011). Future research should focus on integrating these findings into decision-support tools and on validating the proposed confidence measures by comparing simulation outputs with real-world criticality indicators. Ultimately, this work contributes to defining a clearer boundary between scenarios that can be reliably assessed in simulation and those that require real-world experimentation.

Author contributions

Hugues Blache: Writing – original draft, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Pierre-Antoine Laharotte:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Nour-Eddin El Faouzi:** Writing – review & editing, Supervision, Methodology, Conceptualization.

Replication and data sharing

The materials required to reproduce the results of this study are publicly available. The implementation of the Analytic Hierarchy Process (AHP), including the weighted and belief theory-based approaches, is accessible via the following GitHub repository: https://github.com/HuguesBlache/AHP_analyse. The replication package includes the scripts used for data processing, weighting, and aggregation of expert judgments. The survey data and pairwise comparison matrices used in this study can be provided by the corresponding author upon reasonable request, subject to any confidentiality constraints. Detailed instructions for reproducing the main results, including the execution of the code and the reconstruction of the confidence levels, are provided within the repository and accompanying documentation.

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the author(s) used ChatGPT to correct grammar. After utilizing this tool/service, the author(s) thoroughly reviewed and edited the content as necessary and take full responsibility for the final content of the published article.

References

- Aruldos, M., Lakshmi, T. M., Venkatesan, V. P., 2013. A survey on multi-criteria decision making methods and its applications. *American Journal of Information Systems*, 1(1), 31–43.
- BLACHE, Hugues, LAHAROTTE, Pierre-Antoine, EL FAOUZI, Nour-Eddin, et al, 2025 Unifying expert knowledge and field data toward an enhanced scenario description for CAV certification: a comprehensive scenario-based approach. *Journal of Intelligent Transportation Systems* p. 1-27.
- Çelikbilek, Y., Duleba, S., Moslem, S. et al. Ranking the criteria for autonomous vehicle adoption from different stakeholders' viewpoints with a grey AHP model. *Eur. Transp. Res. Rev.* 18, 17 (2026). <https://doi.org/10.1186/s12544-026-00781-w>
- M. A. Del Cacho Estil-Les, A. Marcello Mangini and M. P. Fanti, "Optimizing AHP for Autonomous Vehicle Risk Evaluation in Mixed Traffic," 2026 12th International Conference on Control, Decision and Information Technologies (CoDIT), Bari, Italy, 2026, pp. 2072-2077, doi: 10.1109/CoDIT70676.2026.11630708.
- Dunning, D., 2011. The Dunning–Kruger effect: On being ignorant of one's own ignorance. In: *Advances in Experimental Social Psychology*, 44, 247–296.
- El Faouzi, N.-E., Klein, L. A., De Mouzon, O., 2009. Improving travel time estimates from inductive loop and toll collection data with Dempster–Shafer data fusion. *Transportation Research Record*, 2129(1), 73–80.
- El Faouzi, N.-E., Leung, H., Kurian, A., 2011. Data fusion in intelligent transportation systems: Progress and challenges—A survey. *Information Fusion*, 12(1), 4–10.
- Gao, F., Duan, J., He, Y., Wang, Z., 2019. A test scenario automatic generation strategy for intelligent driving systems. *Mathematical Problems in Engineering*, 2019.
- Gebre, S. L., Cattrysse, D., Van Orshoven, J., 2021. Multi-criteria decision-making methods to address water allocation problems: A systematic review. *Water*, 13(2), 125.
- Hill, R., 1998. What sample size is “enough” in internet survey research. *Interpersonal Computing and Technology*, 6(3–4), 1–12.
- Klir, G., Yuan, B., 1995. *Fuzzy Sets and Fuzzy Logic*. Prentice Hall.
- Kotrlík, J., Higgins, C., 2001. Organizational research: Determining appropriate sample size in survey research. *Information Technology, Learning, and Performance Journal*, 19(1), 43.
- Li, S., Zhou, R. & Huang, H. Multidimensional Evaluation of Autonomous Driving Test Scenarios Based on AHP-EWN-TOPSIS Models. *Automot. Innov.* 8, 237–251 (2025). <https://doi.org/10.1007/s42154-024-00344-6>
- Ni, D., Leonard, J. D., Guin, A., Williams, B. M., 2004. Systematic approach for validating traffic simulation models. *Transportation Research Record*, 1876(1), 20–31.
- Riedmaier, S., Ponn, T., Ludwig, D., Schick, B., Diemeyer, F., 2020. Survey on scenario-based safety assessment of automated vehicles. *IEEE Access*, 8, 87456–87477.
- Saaty, R. W., 1987. The analytic hierarFichy process—what it is and how it is used. *Mathematical Modelling*, 9(3–5), 161–176.
- Saaty, T. L., 2004. Decision making—the analytic hierarchy and network processes (AHP/ANP). *Journal of Systems Science and Systems Engineering*, 13, 1–35.
- Saaty, T. L., 2008. Decision making with the analytic hierarchy process. *International Journal of Services Sciences*, 1(1), 83–98.
- Sadiq, R., Tesfamariam, S., 2009. Environmental decision-making under uncertainty using intuitionistic fuzzy analytic hierarchy process (IF-AHP). *Stochastic Environmental Research and Risk Assessment*, 23(1), 75–91.
- Schwartz, B., 2015. The paradox of choice. In: *Positive Psychology in Practice*, 121–138.
- Sharma, H., 2022. How short or long should be a questionnaire for any research? Researchers' dilemma in deciding the appropriate questionnaire length. *Saudi Journal of Anaesthesia*, 16(1), 65–68.
- Shafer, G., 1992. Dempster-Shafer theory. *Encyclopedia of Artificial Intelligence*, 1, 330–331.
- Taherdoost, H., Madanchian, M., 2023. Multi-criteria decision making (MCDM) methods and concepts. *Encyclopedia*, 3(1), 77–87.

Vaidya, O. S., Kumar, S., 2006. Analytic hierarchy process: An overview of applications. *European Journal of Operational Research*, 169(1), 1–29.

Vargas, L. G., 1987. Priority theory and utility theory. *Mathematical Modelling*, 9(3–5), 381–385.

Zhang, S., Tak, T., 2024. Risk analysis of autonomous vehicle test scenarios using a novel analytic hierarchy process method. *IET Intelligent Transport Systems*, 18(5), 794–807.

Zhang, S., Sun, X., Li, G., Pan, Y., & Tak, T. (2025). Critical test scenario generation for autonomous vehicles using reinforcement learning. *Transportmetrica A: Transport Science*, 2476604.

Author biography



Hugues Blache received the M.S. and Ph.D. degrees in 2021 and 2024, respectively, from ENTPE, and the M.A.Sc. degree from Polytechnique Montréal in 2021. During the Ph.D. study, his work focused on scenario sampling for testing new VAC systems during validation phases. He is currently a Research Associate at UNSW in the

Research Centre for Integrated Transport Innovation (rCITI), with the XXXX. His areas of interest include Machine Learning, Ontology, Embedding Approaches, Transport Innovation, Traffic Modeling and Simulation, and Transport Data Analysis.



Pierre-Antoine Laharotte is a Researcher at the LICIT-Eco7 (French laboratory dedicated to transportation systems modelling and data analysis) supported by the Université Gustave Eiffel and ENTPE (Université de Lyon), Lyon, France. He received the M.S. and Ph.D. degrees respectively in 2013 and 2016 from the University of Lyon. He mainly works on the (Connected-) Intelligent Transport System ((C)-ITS) and especially the application of machine learning processes to deal with ITS issues from traffic state prediction to the impact assessment of connected vehicles.



Nour-Eddin El Faouzi is Professor at Université Gustave Eiffel, Lyon, France, Research Director in Sustainable and Smart Mobility and Affiliate Professor at Mohammed VI Polytechnic University (UM6P), Morocco. He holds the Ph.D. degree in Applied Mathematics from Montpellier University and the HDR (Accreditation

to Supervise Research) in Traffic Modelling from the University of Lyon. With over 30 years of international experience, his research covers traffic modelling, Intelligent Transportation Systems (ITS/C-ITS), Network Resilience, Connected and Automated Mobility, and Data Analytics. He has coordinated and contributed to numerous European and international projects and serves on editorial boards and scientific committees in the ITS field.