J. Shanghai Jiao Tong Univ. (Sci.), 2021, 26(5): 699-712
https://doi.org/10.1007/s12204-021-2360-y

Integrated Framework for Test and Evaluation of Autonomous Vehicles

SU Yimin (苏奕敏), *WANG Lin** (王　琳)
(Department of Automation; Key Laboratory of System Control and Information Processing of Ministry of Education, Shanghai Jiao Tong University, Shanghai 200240, China)

Abstract: Autonomous vehicles must pass effective standard tests to verify their reliability and safety. Accordingly, it is very important to establish a complete scientific test and evaluation system for autonomous vehicles. A comprehensive framework incorporating the design of test scenarios, selection of evaluation indexes, and establishment of an evaluation system is proposed in this paper. The aims of the system are to obtain an objective and quantitative score regarding the intelligence of autonomous vehicles, and to form an automated process in the future development. The proposed framework is built on a simulation platform to ensure the feasibility of the design and implementation of the test scenarios. The design principle for the test scenarios is also presented. To reduce subjective influences, the proposed framework selects objective indexes from four aspects: safety, comfort, driving performance, and standard regulations. The order relation analysis method is adopted to formulate the index weights, and fuzzy comprehensive evaluation is used to quantify the scores. Finally, a numerical example is provided to visually demonstrate the evaluation results for the autonomous vehicles scored by the proposed framework.

Key words: test and evaluation, scenario construction, evaluation index, autonomous vehicles
CLC number: U 467　　**Document code:** A

0 Introduction

With the development of artificial intelligence, research on autonomous driving technology has gradually increased in recent years. Autonomous vehicles have provided a new approach to solving the problems of traditional cars, such as traffic jams and frequent accidents. Research on the intelligence and networking of autonomous vehicles is constantly advancing. However, there are relatively few complete tests and standards for evaluating the performance of autonomous vehicles, such as in regard to their safety, comfort, and coordination. Autonomous vehicles must pass effective standard tests to verify their reliability and safety. It is estimated that autonomous vehicles have to be tested over 275 million miles to prove that they are as reliable as human-controlled vehicles[1]. This mileage is too high to be completed only using real vehicle testing. Therefore, an increasing number of companies and research institutions are testing autonomous driving technologies through simulation platforms. Simulation testing uses computer software to simulate real environments and natural traffic. It can compensate for the shortcomings of real vehicle testing, such as long cycles, high costs, high safety risks, and limited test scenarios.

Many competitions and projects of autonomous vehicle testing have been conducted, such as the European AdaptIVe project, China Future Challenge of Intelligent Vehicle, Defense Advanced Research Projects Agency (DARPA) Grand Challenge, and DARPA Urban Challenge in America[1]. The importance of establishing a test and evaluation system for autonomous vehicles has been recognized. However, there is no unified standard for evaluating the performance of autonomous vehicles. Researchers, companies, and related institutes have proposed their own evaluation indexes and methods from different perspectives and application ranges. For example, the "Autonomy Levels for Unmanned Systems" evaluation framework proposed by Huang et al.[2] divided the intelligence of autonomous vehicles into 10 levels from three dimensions: task complexity, environment complexity, and artificial independence. The DARPA Urban Challenge[3], held in 2007, evaluated participating vehicles according to the time and quality of their completed tasks. In 2010, the China Future Challenge of Intelligent Vehicle evaluated vehicles from

Received: 2021-01-31　**Accepted:** 2021-03-08
Foundation item: the National Natural Science Foundation of China (No. 61873167), and the Automotive Industry Science and Technology Development Foundation of Shanghai (No. 1904)
∗**E-mail:** wanglin@sjtu.edu.cn

two aspects: basic ability tests (traffic signal recognition, curve driving, and fixed-point parking) and complex environment comprehensive tests (recognition of traffic signals during driving, integrated control, lighting use, and road traffic situation perception). Since the fifth competition in 2013, the China Future Challenge of Intelligent Vehicle has evaluated vehicles from the perspective of the "4S" performance measures (i.e., safety, smartness, smoothness, and speed). References [4-7] proposed a hierarchical comprehensive evaluation system for unmanned ground vehicles. They used the expert evaluation method and fuzzy-extended analytic hierarchy process to evaluate the intelligence of autonomous vehicles from five aspects: basic vehicle control behavior, basic driving behavior, basic traffic behavior, advanced driving behavior, and advanced traffic behavior. Son et al.[8] proposed a method for specifying key performance indicators (KPIs) for different test scenarios to evaluate the performance of autonomous vehicles. For example, they selected KPIs such as time-to-collision (TTC), acceleration and deceleration, and fuel consumption for adaptive cruise control, and selected corresponding KPIs based on map generation, global planning, and vehicle control for automatic parking. Wang et al.[9] proposed a three-dimensional evaluation model based on perception, decision, and control layers. The weights of the indexes in each dimension were determined by the entropy method. The comprehensive score was obtained quantitatively through the fuzzy comprehensive evaluation and "Technique for Order Preference by Similarity to Ideal Solution" method. Weng et al.[10] proposed a model predictive instantaneous safety metric for certain situations in which autonomous vehicles were running in a multi-agent interactive high-dimensional continuous system. The proposed metric expanded the traditional TTC indicator, and improved the performance of the autonomous vehicles in terms of operational safety status maintenance. Li et al.[11] combined scenario-based testing and function-based testing by adding test tasks. They also proposed two evaluation methods, Boolean type and numerical type. The numerical type could evaluate from smoothness, safety, and smartness. In addition to the research on evaluation systems, some scholars have conducted research on test scenarios. Feng et al.[12-13] proposed a general framework for generating test scenario libraries. The proposed framework could generate effective test scenarios for operational design domains, autonomous vehicle models, and performance indicators, thereby enriching the current research in the field of test scenario generation.

It can be seen that there are several challenges in the existing research on autonomous vehicle tests and evaluation systems: ① Most existing evaluation systems rely on expert ratings, which are subjective and require human intervention. This is not beneficial for the development of automated processes. ② The scores for autonomous vehicles should ideally be quantitative, which can be more intuitive. ③ There is a lack of a complete system framework for the test and evaluation of autonomous vehicles. Because the performance of an autonomous vehicle is closely related to the surrounding environment, the test and evaluation system should also include a reasonable construction of test scenarios. ④ Many existing evaluation methods use a single index, which is only suitable for a certain function test or a certain scenario test. Such an approach lacks universality and expandability.

Therefore, this study proposes an integrated framework for the test and evaluation of autonomous vehicles. It includes the design of test scenarios, selection of evaluation indexes, and establishment of an evaluation system. The proposed framework is established on a simulation platform, ensuring the feasibility of the design, implementation of the test scenarios, and realization of comprehensive tests. The main contributions of this study are as follows: ① Principles are established for test scenario design. The entire testing process is based on a complete test scenario, rather than on a single scenario test. ② Appropriate objective evaluation indexes are selected based on four aspects: safety, comfort, driving performance, and standard regulations. ③ The priorities of functional safety and collision safety are considered, and the safety indexes adopt surrogate safety measures commonly used in traffic research. ④ The universality and expandability of the evaluation indexes and system are enhanced. More other evaluation indicators and dimensions can be added into the overall framework in future research.

The remainder of this paper is organized as follows. Section 1 introduces the overall framework and process of the proposed test and evaluation system for autonomous vehicles. Section 2 introduces the construction of test scenarios, selection of evaluation indexes, and establishment of the evaluation system. Section 3 provides an intuitive digital example of the proposed framework. Finally, Section 4 concludes the paper.

1 Comprehensive Framework

There are many standard definitions for the classification of autonomous vehicles. The autonomous vehicle classification system developed by the Society of Automotive Engineers[14] is a recognized standard. It divides autonomous vehicles into six levels, from L0 to L5. The autonomous vehicles discussed in this article refer to vehicles at L3 and above. This means that the testing vehicles should complete all driving tasks in a specific scenario, and human drivers basically do not need to be involved.

The test and evaluation system for autonomous vehicles considered in this study is mainly conducted on

a simulation test platform, and uses a continuously developing scenario-based simulation test. The scenario-based testing method tests the autonomous vehicle by allowing the vehicle to complete specific tasks in a preset scenario. On the premise of completing the tasks, the autonomous vehicle can independently choose a way to manage an encountered situation. It has a high degree of freedom and can test the comprehensive performance of autonomous vehicles, and is therefore suitable for testing high-level autonomous vehicles[15].

The test and evaluation system proposed in this study aims to carry out a comprehensive automated test on autonomous vehicles. The vehicles under testing can obtain a score by running in a selected test scenario. The scores are objective and quantitative. The entire evaluation process does not require humans to observe and evaluate the behavior of the test vehicles; rather, it scores according to the indexes formulated in the system.

Therefore, this simulation test and evaluation system contains two important parts: the construction of the test scenarios and establishment of the evaluation system, as shown in Fig. 1. In preliminary work, a scenario library is constructed by various atomic scenarios composed of a static environment and dynamic elements. These atomic scenarios are constructed through feature extraction and data analysis from naturalistic driving data, traffic accident databases, standard regulations, expert experience, and other data sources. The test scenarios used in the test and evaluation system are complete test scenarios which are composed of required atomic scenarios. These atomic scenarios are randomly selected from the scenario library. The test vehicle runs in the test scenarios, which can be model-in-loop (MIL), software-in-loop (SIL), or hardware-in-loop (HIL). The simulation platform records related data, and the tested vehicle is scored according to the evaluation indexes formulated in the evaluation system. Finally, a quantitative score is obtained, including the score for each index and overall total score.

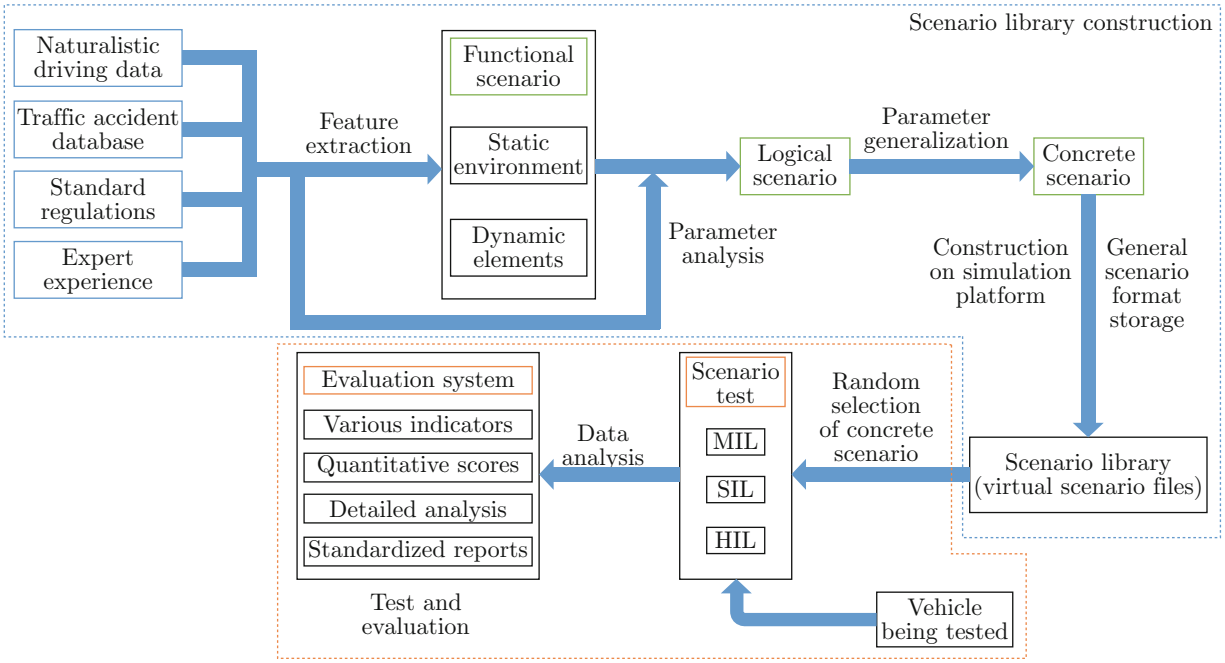


Fig. 1 Simulation test and evaluation system for autonomous vehicles

In contrast to the widely used expert evaluation method, this study adopts a joint index evaluation system composed of objective indexes. By scoring the objective measurement values in segments, the subjectivity is reduced, and automatic scoring can be achieved. The proposed evaluation system scores vehicles from four aspects: safety, comfort, driving performance, and standard regulations.

The overall process of the proposed test and evaluation system is shown in Fig. 2. First, determine p functional scenarios required to form the test scenario, and then randomly select q specific scenarios for each functional scenario from the scenario library. Thus, q^p test scenarios are formed. Then, the test vehicle attempts to pass each test scenario in turn on the simulation platform, and it is judged whether to pass each test scenario without collision. If the vehicle passes the entire test scenario set with no collision, the performance of the vehicle is then classified on each evaluation index. If the vehicle has a crash during driving, the number of collisions is recorded. If the vehicle's pass rate of scenarios (passing scenarios without collision) is found

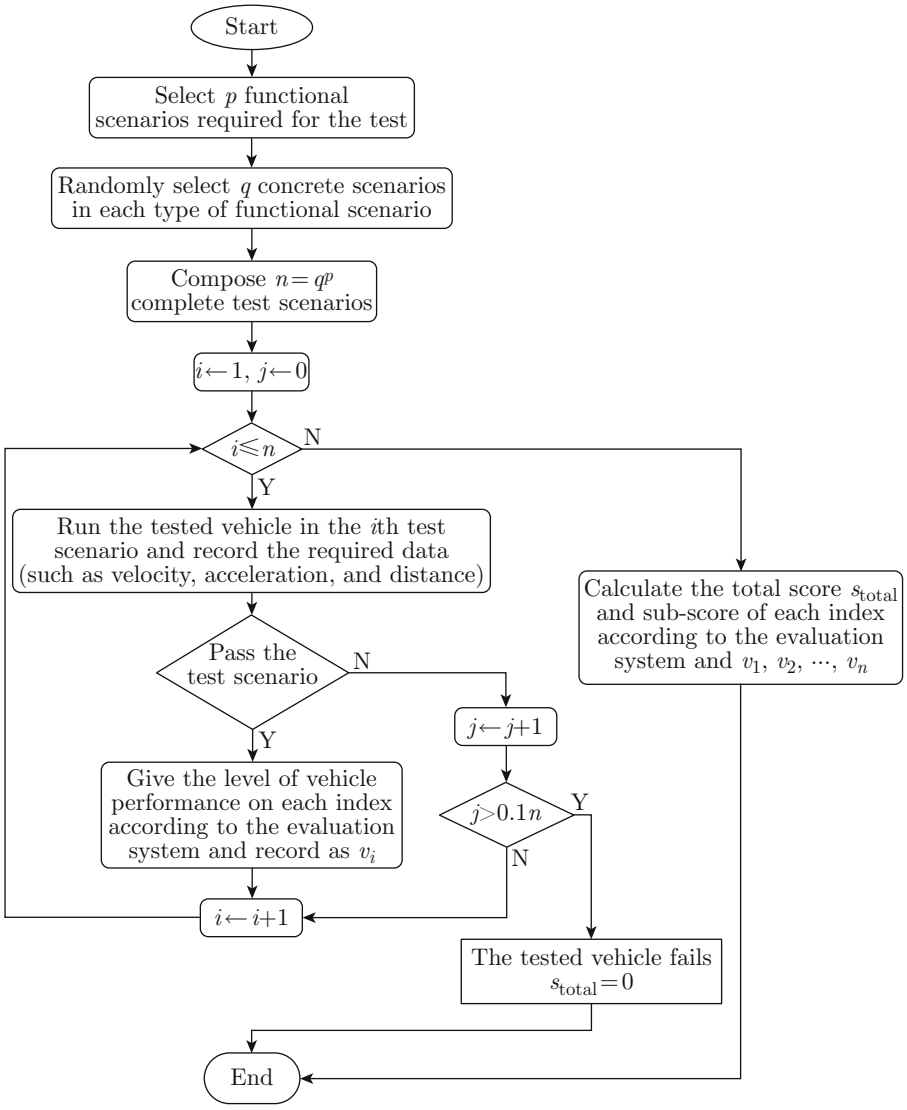


Fig. 2 Autonomous vehicle test evaluation flowchart

to be less than 90% during the entire test process, the vehicle is considered to be an unqualified vehicle, and there is no need to continue scoring. If the pass rate of scenarios is greater than or equal to 90%, the total score and sub-scores are calculated according to the evaluation system and vehicle performance on each evaluation index.

In the entire test process, the scenario library and the evaluation system are two important points, and are explained in further detail below.

2 Methodology

2.1 Construction of Scenario

In this scenario-driven autonomous driving test and evaluation system, the scientific and orderly construction of test scenarios can effectively support the subsequent test and evaluation. Therefore, the scenarios should be as complete and realistic as possible, and cover typical and critical scenarios in the real world.

In the PEGASUS project[16] carried out in Germany, scenarios are defined and classified as functional scenarios, logical scenarios, and concrete scenarios, owing to the different requirements for scenarios in different stages of autonomous driving product development[17]. Their abstraction degree goes from high to low, respectively, and the number of scenarios goes from fewer to greater, respectively. This standard can be referred to when designing a scenario library. The construction process of the scenario library is shown in the upper part of Fig. 1. First, features are extracted and summarized from naturalistic driving data, traffic accident databases, standard regulations, and expert experience. Then, common functional scenarios are selected, such as parking, lane changing on highways, and pedestrian avoidance at intersections. According to the definition

in the PEGASUS project, the construction of scenarios can be divided into six layers[16,18].

Layer 1: road attributes such as geometry, topology, and quality.

Layer 2: traffic infrastructure such as boundaries and traffic signs.

Layer 3: temporary manipulations of Layers 1 and 2.

Layer 4: static or dynamic objects and their maneuvers and interactions.

Layer 5: environmental conditions like weather and lighting.

Layer 6: digital information such as vehicle-to-everything information or digital maps.

Through data analysis of typical road scenarios (such as those from the highD dataset[19], or the Safety Pilot Model Deployment Program conducted by the University of Michigan[20-21]), traffic accident databases (such as the EU "ASSESS" data and China In-Depth Accident Study database), and standard regulations (such as advanced driver assistance systems (ADAS) regulations, International Organization for Standardization (ISO) 15622, and ISO 21201), the ranges of the related parameters (such as the type, location, target speed, and acceleration) of the traffic facilities and participants in each functional scenario (such as traffic signals, lanes, test vehicles, object vehicles, pedestrians, and obstacles) are determined. Thus, logical scenarios are formed. The specific value of each parameter in the logical scenarios can be determined based on expert experience and/or on probability distributions calculated from the above data sources. In this way, a large number of concrete scenarios are obtained, and can be stored in general file formats to construct a scenario library. Finally, through the relevant simulation software (such as Virtual Test Drive (VTD)), the required scenario file is run, and the test scenario is obtained. Figure 3 shows an example of a concrete scenario from the scenario library.

Fig. 3 Example of concrete scenario

In general, the scenarios in the library are concrete scenarios generalized from a single functional scene. To test autonomous vehicles at L3 and above, the test scenario should be a complete enclosed test field, covering natural driving scenarios, dangerous scenarios, and accident scenarios. Therefore, it is necessary to determine the functional scenarios required for the test according to the test objects (autonomous vehicles of different levels) and test purposes (driving performance on the highway, parking performance, etc.). Then, several concrete scenarios of each functional scenario are randomly selected from the library to form a complete test scenario. During the test, the tested vehicle passes through each test area and completes the corresponding tasks. The simulation platform records the data of the tested vehicle during the entire test process (such as its trajectory, speed, acceleration, distance to surrounding vehicles or pedestrians, and driving time). These data are used in the following evaluation system.

Figure 4 shows an example of a designed test scenario. The content of the test is autonomous parking in the test area. The test vehicle is required to start from the starting point and go through five scenarios in sequence to complete six tasks: obstacle detection, lane changing, car following, left turn at intersection, pedestrian avoidance, and parking.

The test scenario constructed above includes a variety of working conditions. Different evaluation indexes are set according to different working conditions to form a reasonable evaluation system for the test vehicle, as explained in Subsection 2.2.

2.2 Evaluation System

As shown in the schematic diagram and flowchart of the proposed test and evaluation system described in Section 1 (Figs. 1 and 2), the evaluation system is an important part of the entire framework. Although there is no unified evaluation standard for autonomous vehicles and the evaluation methods in different studies are not identical, most are based on the idea of a comprehensive evaluation. The first step of the comprehensive evaluation is to clarify the evaluation purpose(s) and object(s). The second step is to choose different evaluation dimensions and specific evaluation indexes. The third step is to determine the weights of the indexes, and to select the aggregation model. The final step is to integrate the results from all of the evaluation indexes to obtain a total result that can reflect the performance of the autonomous vehicle. This study also establishes an evaluation system based on this idea. Based on domestic and foreign evaluation experiences in autonomous driving and the test scenarios designed as described in the previous section, the evaluation system is chosen to score autonomous vehicles from four aspects: safety, comfort, driving performance, and standard regulations. The index weights are determined by the order relation analysis method, and comprehensive scores are calculated using the fuzzy comprehensive evaluation.

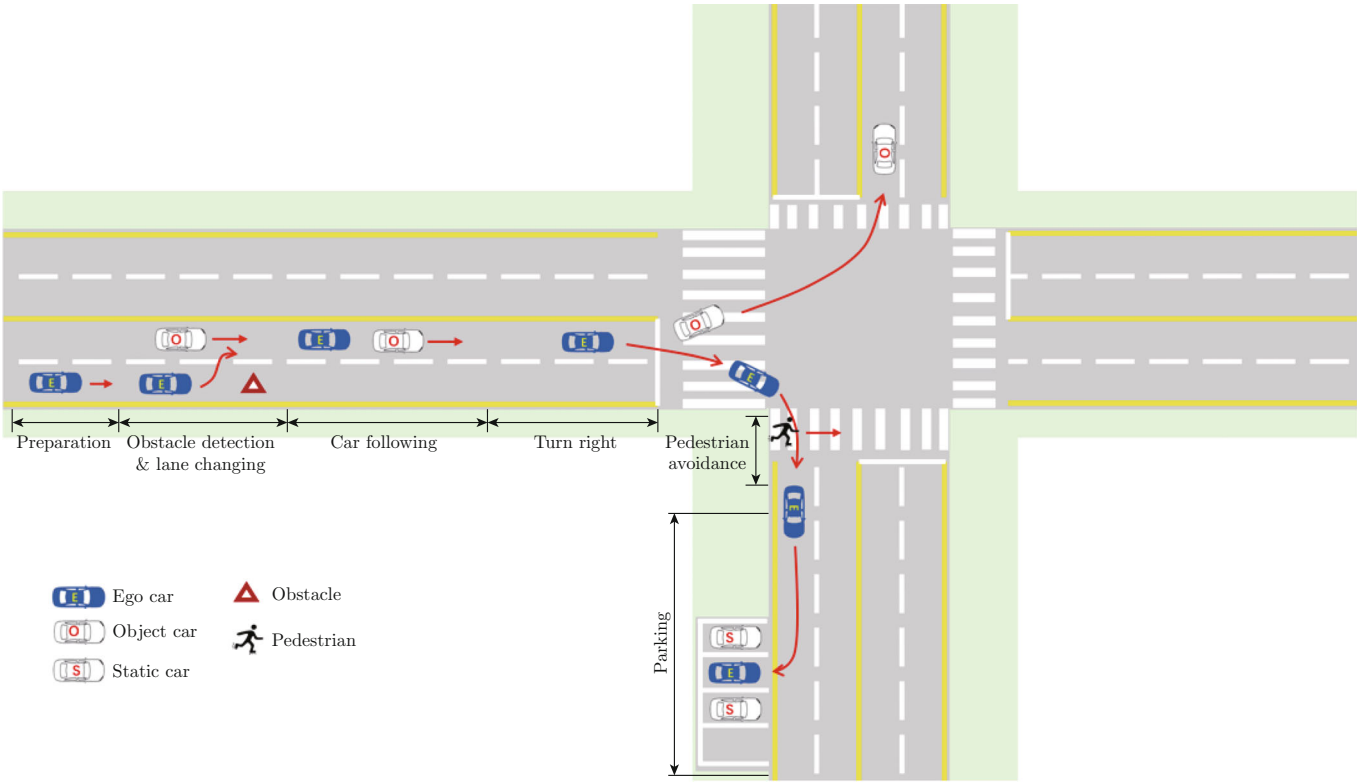


Fig. 4 Example of test scenario

2.2.1 *Evaluation Model*

The aim of the evaluation system proposed in this study is to comprehensively reflect the intelligence levels of autonomous vehicles, and to realize an automated process. Therefore, a joint index evaluation system composed of objective indexes is adopted. The proposed evaluation system scores vehicles from four aspects: safety, comfort, driving performance, and standard regulations.

Safety usually refers to functional safety and collision safety. Functional safety refers to whether the vehicle safety can be guaranteed in the event of a vehicle system failure. Collision safety refers to whether an autonomous vehicle will collide with surrounding vehicles, traffic facilities, or other traffic participants while driving. These two points are the most important, and should be first guaranteed. If these two points cannot be met, the vehicle should not be allowed to drive on the road. However, the current evaluation systems proposed by some scholars place functional safety and collision safety at the same levels as other indexes. In the framework proposed herein, we choose to prioritize functional safety and collision safety. According to the framework mentioned in Section 1, if the test vehicle cannot pass 90% of the test scenarios without collision, the vehicle is considered to be unqualified, and there is no need for further scoring. Under this premise, in the proposed evaluation system, the safety indexes are different from those commonly used in previous work (e.g., whether a collision occurs or whether it is safe when the vehicle has a system failure). A surrogate safety measure is adopted. Surrogate safety measures are often applied to evaluate the risks of crashes in road networks and traffic flows[22]. This is because crashes are rare events, and crash-free car-following algorithms are often used in traffic simulation software. This is similar to the evaluation system proposed in this study. Therefore, two commonly used safety indexes in surrogate safety measures are selected here: the time exposed time-to-collision (TET) and number of critical jerks (NCJ). The specific meanings of these two indexes are explained in Subsection 2.2.2.

Comfort mainly concerns the experience of the driver and passengers while driving. For autonomous vehicles, the main evaluation is whether the control algorithm can control the vehicle smoothly, and without frequent intense maneuvers. In the simulation platform, the maximum acceleration and deceleration, maximum jerk, maximum yaw rate, quickness, headway distance in seconds, and lane deviation during the driving process can be recorded to determine whether the vehicle is driving and turning smoothly. The specific meanings of the above indexes are explained in Subsection 2.2.3.

Driving performance mainly concerns the time and quality of the tasks completed by the test vehicle. The time required to complete the tasks reflects the

efficiency of the autonomous driving control algorithm, and the quality of the tasks reflects the control performance of the vehicle control algorithm.

Standard regulations mainly evaluate the degree of compliance of the tested vehicle with the national traffic laws.

The above four aspects form the evaluation system for autonomous vehicles through a hierarchical structure, as shown in Fig. 5.

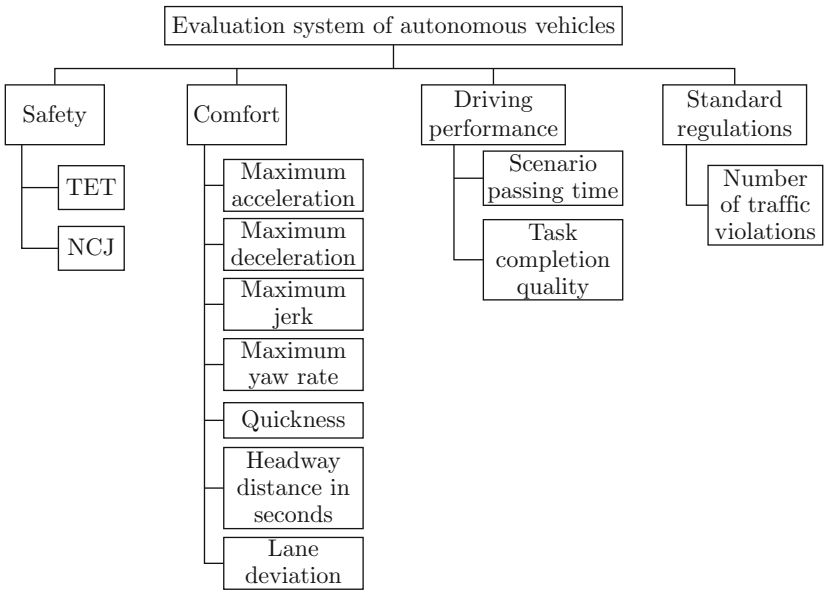


Fig. 5 Evaluation system for autonomous vehicles

Owing to the different test scenarios and purposes, the relative importance of each evaluation index is also different. Therefore, each evaluation index must be weighted. There are two main methods for assigning weights: objective and subjective[23]. The objective weighting method assigns a weight by analyzing the actual values of the indexes, and is not affected by human factors. For example, in the entropy method[24], the greater the degree of the difference between test vehicles on a certain index, the greater the weight of the index. However, the objective weighting method depends on the test results, and the index weight is not universal. In the subjective weighting method, experts first judge the importance of each index by experience, and then use certain mathematical methods to obtain concrete weight values. It is suitable for an autonomous driving evaluation, which has a wide range of evaluation objects and purposes, many test indexes, and uncertain test results. The analytic hierarchy process and order relation analysis method are two commonly used subjective weighting methods. The order relation analysis method has a small amount of calculation, and is suitable for evaluation systems with many indexes. Thus, the order relation analysis method is selected for the proposed evaluation system to assign weights to indexes. The specific principle of the order relation analysis method is explained in Subsection 2.2.6.

After determining the evaluation indexes and the corresponding weights, the next step is to select an appropriate aggregation model to integrate the values of the indexes into an overall evaluation result. Here, the fuzzy comprehensive evaluation is chosen. This method uses fuzzy set theory for evaluation, which can integrate qualitative and quantitative subjective and objective indexes, and can effectively solve the problems of ambiguity and uncertainty. It is easy to implement, and is suitable for multi-level index systems. Therefore, it is applicable to the evaluation system proposed above. The specific principle of the fuzzy comprehensive evaluation is explained in Subsection 2.2.7.

2.2.2 *Safety Indexes*

As mentioned in Subsection 2.2.1, the safety indexes in the proposed evaluation system adopt surrogate safety measures. There are two types of safety indexes in surrogate safety measures: time proximity-based indexes (such as the TTC) and evasive action-based indexes (such as the yaw rate)[22]. Here, two indexes are chosen, as mentioned above: the TET and NCJ.

(1) TET. The TTC is a commonly used surrogate safety measure. It represents the time required for two vehicles to collide when the following vehicle drives faster than the leading vehicle, and their speed difference is constant[22]. This can be expressed by[22]

$$\mathrm{TTC_f}(t)=\begin{cases}\dfrac{x_{\mathrm{l}}(t)-x_{\mathrm{f}}(t)-L_{\mathrm{l}}}{v_{\mathrm{f}}(t)-v_{\mathrm{l}}(t)}, & v_{\mathrm{f}}(t)>v_{\mathrm{l}}(t)\\ \infty, & v_{\mathrm{f}}(t)\leqslant v_{\mathrm{l}}(t)\end{cases}, \quad (1)$$

where $\mathrm{TTC_f}(t)$ is the TTC value of the following vehicle at time t, $x_1(t)$ and $v_1(t)$ are the longitudinal position and speed of the leading vehicle at time t, respectively, $x_f(t)$ and $v_f(t)$ are the longitudinal position and speed of the following vehicle at time t, respectively, and L_1 is the length of the leading vehicle.

A smaller TTC indicates a greater hazard to vehicles. When the TTC value is less than a certain value, it is considered dangerous. According to the standards ISO 15623 and GB/T 33577—2017[26], the TTC threshold can be selected as 2.4 s. The TET[27] is derived from this, and represents the total time when the vehicle is in a dangerous situation. It is determined based on the TTC value below the TTC threshold (TTC^*), and is expressed by

$$\left.\begin{aligned} \mathrm{TET} &= \sum_{t=1}^{T} \delta_t \Delta t \\ \delta_t &= \begin{cases} 1, & 0 < \mathrm{TTC_f}(t) \leqslant \mathrm{TTC}^* \\ 0, & \text{otherwise} \end{cases} \end{aligned}\right\}, \tag{2}$$

where δ is the switching variable, Δt is the time step, and T is the total simulation time.

(2) NCJ. Jerk is defined as the derivative of acceleration. It is an evasive action-based indicator that can be used to measure the severity of conflict, and can be calculated by

$$\mathrm{Jerk}(t) = \dot{a}(t) = \ddot{v}(t), \tag{3}$$

where $a(t)$ is the acceleration at time t, and $v(t)$ is the velocity at time t.

Bagdadi and Várhelyi[28] found that there is a proportional relationship between critical or dangerous jerks (i.e., jerk less than or equal to -9.9 m/s^3) and the number of crashes. Therefore, jerkiness during driving can reflect dangerous driving behaviors and a higher accident rate. Therefore, some scholars have used the number of jerks which are less than or equal to -9.9 m/s^3 as an index for evaluating safety-critical driving behaviors[22]. Here, we also consider the NCJ as a safety index.

2.2.3 *Comfort Indexes*

Comfort is another important aspect of vehicle evaluation. Traditional cars have mainly been evaluated based on vehicle vibrations, seat comfort, etc. In this study, the test and evaluation of autonomous vehicles are conducted on a simulation platform, and the test and evaluation system mainly focuses on the control algorithms; this is different from traditional vehicle evaluations. Therefore, the corresponding specific indexes are formulated according to the different working conditions in the designed test scenario, as shown in Table 1.

The human vestibular system is sensitive to acceleration and jerk, which affects the judgment of vehicle comfort. Therefore, the maximum acceleration, maximum deceleration, maximum jerk, and maximum yaw rate are the most important indexes for comfort evaluation. The jerk is calculated as shown in Eq. (3).

Table 1 Comfort indexes and corresponding working conditions

Comfort index	Corresponding conditions
Maximum acceleration	All working conditions
Maximum deceleration	All working conditions
Maximum jerk	All working conditions
Maximum yaw rate	All working conditions
Quickness	Car following, lane changing
Headway distance in seconds	Car following
Lane deviation	Cruising, car following

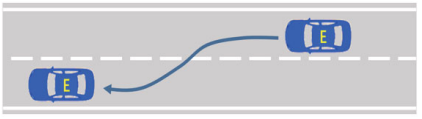
Quickness describes the swiftness of vehicle's certain maneuvers[29]. This index was originally used to evaluate the flight quality of aircraft. During the study of the objective indexes of comfort for autonomous vehicles, Bellem et al.[29] adjusted the calculation method of quickness and divided it into longitudinal quickness q_{long} and lateral quickness q_{lat}:

$$q_{\mathrm{long}} = \frac{\bar{a}}{\Delta v}, \tag{4}$$

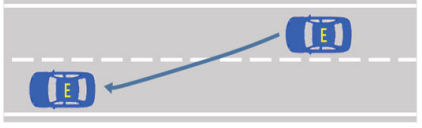
$$q_{\mathrm{lat}} = \frac{v_{\mathrm{lat}}}{d_{\mathrm{lat}}}, \tag{5}$$

where $\bar{a}$ is the mean longitudinal acceleration, Δv is the change in longitudinal velocity, v_{lat} is the lateral velocity, and d_{lat} is the lateral offset.

Longitudinal quickness can be used for car following, and lateral quickness can be used for lane changing. A higher quickness indicates that lane changes occur faster, as shown in Fig. 6.



(a) High quickness



(b) Low quickness

Fig. 6 Illustration of lateral quickness[29]

In car following and cruising, the distance from the leading vehicle and deviations from the lane line also affect the comfort of the driver and passengers. Driving too close to the leading vehicle or lane boundaries will cause people to feel nervous. Therefore, the headway distance in seconds and lane deviations are also considered. The lane deviation is defined as the distance

between the vehicle and the centerline of the lane, and can be obtained using sensors. The headway distance in seconds is calculated by

$$t_{\mathrm{HD}}(t)=\frac{x_{\mathrm{l}}(t)-x_{\mathrm{f}}(t)}{v_{\mathrm{f}}(t)}, \tag{6}$$

where t_{HD} is the headway distance in seconds, x_{l} is the longitudinal position of the leading vehicle, and x_{f} and v_{f} are the longitudinal position and speed of the following vehicle, respectively.

2.2.4 *Driving Performance Indexes*

The driving performance index mainly evaluates the time and quality of the tasks completed by the test vehicle. If the time to pass the entire scenario is shorter, it indicates that the efficiency of completing the test tasks is higher, and that the driving performance is better. The quality of task completion will have different evaluation focuses according to different working conditions. For example, in car following, the response time (the interval between the time the target vehicle changes velocity and the time the test vehicle starts to follow) and speed control accuracy (the maximum difference between the steady velocities of the test vehicle and target vehicle) are considered. During cruising, the evaluation should be based on the speed control accuracy. In parking conditions, the number of parking attempts, parking attitudes, and parking positions are considered. For example, the "CCRT (Smart Electric Vehicle) Management Rules (2020 Edition)" issued by the China Automotive Technology and Research Center and "Intelligent Parking Assist Rating Protocol" issued by the Intelligent Vehicle Integrated System Test Area have established specific indexes for related content, as can be seen in Table 2.

Table 2 Example of a score table for vertical parking with static vehicles on both sides

Evaluation content			Score
Successfully identify the target parking space			10
Parking times n		$n \leqslant 4$	20
		$4 < n \leqslant 6$	15
		$6 < n \leqslant 9$	10
		$n > 9$	0
Parking attitude α		$-1^\circ \leqslant \alpha \leqslant 1^\circ$	25
		$1^\circ < \lvert\alpha\rvert \leqslant 2^\circ$	20
		$2^\circ < \lvert\alpha\rvert \leqslant 3^\circ$	15
		$\lvert\alpha\rvert > 3^\circ$	0
Park accurately into the target area	Horizontal evaluation (distance from the two boundaries of the parking space)	$\Delta d \geqslant 0.2\,\mathrm{m}$	15
	Longitudinal evaluation (distance from the rear line of the parking space)	$\Delta d \leqslant 0.4\,\mathrm{m}$	10
Avoid crashing into the curb, etc.			20

2.2.5 *Standard Regulation Indexes*

The standard regulation index mainly evaluates the degree of compliance of the tested vehicle to national traffic laws, that is, whether it can accurately identify traffic lights, speed limit signals, lane lines, and other signals, and correctly comply with them. This part is evaluated based on the number of violations, according to the atomic results recorded on the simulation platform.

Using the definitions and concrete equations of the above indexes, the values of each index can be obtained. However, the relationship between the objective measurements and the intelligence level of an autonomous vehicle is not one-to-one, but rather is a many-to-one relationship. This means that the objective measurements within a range can correspond to the same intelligence level. Therefore, in the proposed evaluation system, the objective measurements are divided into five grades, according to the concrete data: very good, good, normal, poor, and very poor. This is also convenient for the fuzzy comprehensive evaluation used in the subsequent procedure to calculate the scores. The segmentation range for each index is different for different test scenarios. Thus, specific and reasonable standards should be formulated according to previous test situations.

2.2.6 *Order Relation Analysis Method (G1-Method)*

The performance values reflected by different indexes sometimes present contradictory situations. For example, reducing the acceleration for comfort increases the task completion time. In this situation, the comfort score increases, but the driving performance score decreases. The scores of the vehicles cannot be evaluated from a single perspective. Therefore, an overall evaluation must be achieved by adjusting the index weights. An index with greater importance (such as

the safety index) and difference usually has a higher weight. Here, the order relation analysis method[30] is used to decide the weight, and the detailed process is as follows.

(1) Experts or decision makers sort the indexes in the index set $\{u_1, u_2, \cdots, u_m\}$($m$ is the number of indexes) according to a certain criterion (such as the importance of the index or the complexity of the task reflected) to form an order relationship:

$$u_1^* > u_2^* > \cdots > u_m^*, \tag{7}$$

where $u_i^*\,(i = 1, 2, \cdots, m)$ indicates the index after sorting.

(2) The ratio of the importance between index u_{i-1}^* and index u_i^* is given as $\beta_i = w_{i-1}^*/w_i^*$, where w_i^* represents the weight coefficient corresponding to the sorted index. The assignment reference values for β_i are listed in Table 3.

Table 3 Assignment reference table of β_i

β_i	Explanation
1.0	u_{i-1}^* is as critical as u_i^*.
1.2	u_{i-1}^* is slightly more critical than u_i^*
1.4	u_{i-1}^* is evidently more critical than u_i^*
1.6	u_{i-1}^* is strongly more critical than u_i^*
1.8	u_{i-1}^* is extremely more critical than u_i^*

(3) The weight coefficient w_i^* is calculated by

$$w_m^* = \left(1 + \sum_{i=2}^{m} \prod_{k=i}^{m} \beta_k\right)^{-1}, \tag{8}$$

$$w_{i-1}^* = \beta_i w_i^*, \quad i = 2, 3, \cdots, m. \tag{9}$$

2.2.7 *Fuzzy Comprehensive Evaluation*

The fuzzy comprehensive evaluation is used in the proposed evaluation system to integrate the scores of each index into an overall evaluation result[4,31-32]. The process is as follows.

(1) After determining the evaluation index set $\boldsymbol{U} = (u_1, u_2, \cdots, u_m)$ and evaluation level set $\boldsymbol{V} = (v_1, v_2, \cdots, v_n)$ (n is the number of levels), judge the membership degree r_{ij} of a certain index $u_i\,(i = 1, 2, \cdots, m)$ at each evaluation level $v_j\,(j = 1, 2, \cdots, n)$. An evaluation matrix $\boldsymbol{R}$ is composed of the evaluation results of the m indexes, which reflects the fuzzy relationship between $\boldsymbol{U}$ and $\boldsymbol{V}$:

$$\boldsymbol{R} = (r_{ij})_{m\times n} = \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{bmatrix}. \tag{10}$$

Generally, $\boldsymbol{R}$ needs normalizing by row or column.

(2) The evaluation matrix $\boldsymbol{R}$ is combined with the weight set $\boldsymbol{A} = (w_1, w_2, \cdots, w_m)$ (m is the number of indexes) determined by the order relation analysis method to obtain the membership degree set $\boldsymbol{C}$ of the upper-level evaluation index:

$$\boldsymbol{C} = \boldsymbol{A}\boldsymbol{R} = \begin{bmatrix} w_1 & \cdots & w_m \end{bmatrix} \begin{bmatrix} r_{11} & \cdots & r_{1n} \\ \vdots & \ddots & \vdots \\ r_{m1} & \cdots & r_{mn} \end{bmatrix} = \begin{bmatrix} c_1 & \cdots & c_n \end{bmatrix}. \tag{11}$$

For a multi-level evaluation system, the calculation starts from the lowest level. The result calculated using Eq. (11) forms the evaluation matrix of the upper-level evaluation index and so on, until the membership degree set $\boldsymbol{C}_{\mathrm{t}}$ of the target level is obtained.

(3) The comprehensive evaluation score is calculated. By quantifying the evaluation levels (for example, very good, good, normal, poor, very poor) into the corresponding score set $\boldsymbol{\mu}$ (for example, $\boldsymbol{\mu}$ = [100 80 60 40 20]), a comprehensive score G under a hundred-mark system can be calculated by

$$G = \boldsymbol{C}\boldsymbol{\mu}^{\mathrm{T}} = \begin{bmatrix} c_1 & c_2 & \cdots & c_n \end{bmatrix} \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_n \end{bmatrix}^{\mathrm{T}}. \tag{12}$$

As discussed in Subsections 2.2.2—2.2.5, the performance values of vehicles on the different indexes are divided into five levels, according to the concrete values. As described in Section 1, the autonomous vehicle is tested in a number of test scenarios. The results from each test belong to one of five levels. Thus, the membership degree r_{ij} of each evaluation level is equal to the average times that this evaluation level is obtained, and the evaluation matrix $\boldsymbol{R}$ is thus normalized by row:

$$r_{ij} = \frac{e_{ij}}{N}, \tag{13}$$

where N is the total number of test scenarios, e_{ij} is the times that the index u_i is divided into the evaluation level v_j in the N tests, and $e_{i1} + e_{i2} + \cdots + e_{in} = N$.

3 Numerical Example

In this section, a numerical example is presented to show the evaluation results of an autonomous vehicle scored using the framework of the test and evaluation system proposed in this study.

The example was run on the VTD simulation platform. VTD is a complete modular simulation tool chain for ADAS and autonomous driving, and was developed by the German company VIRES. VTD runs on Linux. It can realize road simulations, traffic scenario modeling, weather and environment simulations, sensor simulations, scenario simulation management, and high-precision real-time screen rendering. It can also be co-simulated with third-party tools and plugins, such as

the joint simulation of VTD and CarSim. The simulation process of VTD mainly consists of three steps: road network construction, dynamic scenario configuration, and simulation operation. VTD uses a road network editor (ROD) to build static elements for test scenarios, including road networks, lanes, and surrounding environments. The ROD generates high-precision maps in the OpenDrive format. VTD uses Scenario Editor to build dynamic scenarios. It adds user-defined traffic entities or continuous traffic flows based on OpenDrive to form the final test scenarios. The scenario files are stored in the OpenScenario format. In addition to using the VTD custom language or graphical editor to define vehicle behaviors or using VTD's default vehicle controllers, VTD can be co-simulated with third-party tools, such as in the joint simulation of VTD and CarSim mentioned above. In this case, VTD provides the test scenarios, and CarSim, as a simulation software specifically for vehicle dynamics, provides controller algorithms for the test vehicles. In CarSim, the algorithms are written using MATLAB/Simulink.

The test scenario shown in Fig. 4 was constructed on the VTD simulation platform. In the test scenario, the tested vehicle was required to complete six test tasks: obstacle detection, lane changing, car following, left turn at intersection, pedestrian avoidance, and parking. The ranges of the parameters in the test scenario were set as follows: in the preparation part, the steady velocity of the test vehicle was 30—40 km/h; in the obstacle detection and lane changing part, the initial distance between the tested vehicle and obstacle was 40—70 m and the speed of the object car was 30—40 km/h; in the car-following part, the changes of the object car's speed were 40 km/h—60 km/h—30 km/h—40 km/h, and the acceleration was 3—6 m/s^2. In the pedestrian avoidance part, the speed of the pedestrian was 3—5 km/h, and he started to cross the intersection when he was 8—10 m away from the vehicle. Specific parameters within the ranges mentioned above were selected, and 10 concrete test scenarios were generated. The tested vehicle was run in these scenarios. The control algorithm of the tested vehicle was set as the default driver control algorithm in the VTD as a demonstration. The tested vehicle passed each test scenario without collisions. Then, the indexes were calculated according to the driving data of the vehicle. Each index was classified according to its value, and the probability of each grade was obtained. The statistics are listed in Table 4.

Table 4 Example of fuzzy evaluation table of a test vehicle

Index & corresponding weight		Sub-index & corresponding weight		Rank level				
				Very good	Good	Normal	Poor	Very poor
Safety	0.36	TET	0.58	0	0.6	0.4	0	0
		NCJ	0.42	0	0.6	0.2	0.2	0
Comfort	0.21	Maximum acceleration	0.24	0.1	0.6	0.3	0	0
		Maximum deceleration	0.20	0.1	0.5	0.4	0	0
		Maximum jerk	0.17	0	0.6	0.4	0	0
		Maximum yaw rate	0.14	0	0.6	0.3	0.1	0
		Quickness	0.09	0	0.5	0.4	0.1	0
		Headway distance in seconds	0.08	0.6	0.2	0.2	0	0
		Lane deviation	0.08	0.6	0.3	0.1	0	0
Driving performance	0.17	Scenario passing time	0.42	0	0.8	0.2	0	0
		Task completion quality	0.58	0.1	0.7	0.2	0	0
Standard regulations	0.26	Number of traffic violations	1	0.2	0.7	0.1	0	0

The index for task completion quality can be taken as an example to illustrate how the indexes were scored. In the built test scenario, there were two main parts related to the task completion quality, namely the performance (response time and speed control accuracy) in car following and the performance in parking. In 10 tests, the response time was approximately 3 s, and the speed control accuracy was ±1 km/h. There was not much difference. Therefore, the main impact on the task completion quality was in the parking performance. In one test, the test vehicle successfully identified the parking space, and after adjusting three times, it reversed into the parking space. However, the angle with the center line of the parking space was 2.45°, and it was too close to the right vehicle, at less than 0.2 m. In the longitudinal direction, the vehicle was within 0.4 m from the front and rear lines of the parking space. The tested vehicle did not crash into the surrounding vehicles and roadsides during the entire reversal process. According to Table 2, the parking performance of the test vehicle was scored as 75, which belongs to the second level, that is, the good level. Therefore, in this

test, the task completion quality of the test vehicle was considered as good. In each test, the rating was performed in this manner, and finally, 10 rating evaluations were obtained. After normalization, the probability of each level of task completion quality was obtained, as presented in Table 4. The remaining indexes were evaluated in the same way. Owing to space reasons, other processes are not repeated here.

The order relation analysis method was used to determine the weight of each index. The specific process is briefly described as follows.

The order relationship between safety, comfort, driving performance, and standard regulations was safety > standard regulations > comfort > driving performance. The importance ratios were 1.4, 1.2, and 1.2, respectively. Through Eqs. (8) and (9), the weight set for safety, comfort, driving performance, and standard regulations was determined as [0.36 0.21 0.17 0.26].

The order relationship between TET and NCJ was TET > NCJ. The importance ratio was 1.4. Thus, the weight set for TET and NCJ was [0.58 0.42].

The order relationship between the indexes of comfort was maximum acceleration > maximum deceleration > maximum jerk > maximum yaw rate > quickness > headway distance in seconds > lane deviation. The importance ratios were 1.2, 1.2, 1.2, 1.4, 1.2, and 1, respectively. Thus, the weight set for the maximum acceleration, maximum deceleration, maximum jerk, maximum yaw rate, quickness, headway distance in seconds, and lane deviation was [0.24 0.20 0.17 0.14 0.09 0.08 0.08].

The order relationship between scenario passing time and task completion quality was task completion quality > scenario passing time. The importance ratio was 1.4. Thus, the weight set for the scenario passing time and task completion quality was [0.42 0.58].

Table 4 lists the weights and grades of each index. The importance of each index varies according to the different test purposes and objects, so the weights are different in different tests.

The scores for each index and total score were obtained by fuzzy comprehensive evaluation, as shown in Table 5. Tables 4 and 5 show that the test vehicle may have low safety and comfort scores, owing to its excessive speed changes and yaw angle changes. The control algorithm can be optimized in this respect.

Table 5 Example of vehicle scores

Index	Score
Safety	70.32
Comfort	75.44
Driving performance	77.16
Standard regulations	82.00
Total score	75.60

The following is a brief description of the advantages of the framework proposed in this article relative to those from previous research.

In Ref. [31], a set of evaluation indicators and an evaluation system were constructed for L2 level autonomous vehicles. However, the indicators could not cover the advanced performance of L3 level autonomous vehicles and above. Expert scoring was used for scoring. Compared with the evaluation method in Ref. [31], one of the biggest advantages of the framework proposed herein is that the evaluation system is composed of objective indicators, rather than being scored by experts, and therefore facilitates the development of subsequent automated processes. In addition, it is more reasonable to prioritize the functional safety and collision safety.

Reference [1] defined four indexes: safety index, efficiency index, rationality index, and comfort index. The four indexes were calculated from an energy perspective, and were calculated through defined mathematical formulas. After comparison with human driving data, the indexes were normalized to form a score between 0 and 1, and were divided into intelligence levels. Compared with the evaluation method proposed in Ref. [1], the indicators selected by the framework proposed herein are more intuitive, and the final scores are also more intuitive, making it convenient for algorithm developers to find out which part of the vehicle performance needed to be improved. The evaluation index mentioned in Ref. [1] is too abstract; the final total score combines four evaluation indexes which only reflect the predefined vehicle intelligence, but do not provide directions for improvement. In addition, there is no scalability. Therefore, it is difficult to add new evaluation indicators.

The unmanned ground vehicle evaluation system proposed in Ref. [4] used an expert evaluation method and fuzzy-extended analytic hierarchy process to evaluate the intelligence of autonomous vehicles from five aspects. Compared with Ref. [4], the indicators and perspectives evaluated by the framework proposed herein are no longer basic vehicle function tests, but are considered from a more holistic perspective, and can better reflect the intelligence of autonomous vehicles. Reference [4] also used expert scoring and the extended analysis hierarchy method to determine weights, and therefore required a large amount of calculation. This study uses the order relation analysis method to determine the weights. The calculation amount is relatively small, and is suitable for a framework with many indicators, such as that proposed herein.

4 Conclusion

This paper proposes a complete and comprehensive test and evaluation framework for autonomous vehicles, including the design of test scenarios, the selection of

evaluation indexes, and the establishment of an evaluation system. It aims to form an objective and quantitative score for the intelligence of autonomous vehicles, and can be used to form an automated process in the future development. The proposed framework is built on a simulation platform to ensure the feasibility of the design and implementation of the test scenario. A numerical example is finally provided to visually demonstrate the evaluation results for the autonomous vehicle as scored by the proposed framework.

There are four main contributions of this study. ① Principles are established for test scenario design. The entire testing process is based on a complete test scenario, rather than on a single function test or a specific scenario test, making the testing process more reasonable. ② Appropriate evaluation indexes are selected. A comprehensive evaluation of autonomous vehicles is conducted based on four aspects: safety, comfort, driving performance, and standard regulations. The selected indexes are all objective, which can reduce subjective influences and provide automatic scoring without external human intervention. ③ The priorities of functional safety and collision safety are considered. Considering that the priority levels of these two indexes are significantly higher than those of other indexes, in the proposed framework, it is necessary to pass a scenario without failure and collision before performing subsequent evaluations. Therefore, the safety indexes adopt surrogate safety measures commonly used in traffic research; this is a different approach from previous research. ④ The universality and expandability of the evaluation indexes and system are enhanced. The final evaluation result is given in a quantitative form through fuzzy comprehensive evaluation, rather than using a grade classification. Thus, the final result is more intuitive. In future research, other evaluation indicators and dimensions can be added without affecting the overall framework.

For example, the coordination, learning abilities, and communication abilities of intelligent vehicles can also be evaluated. Coordination can be evaluated by designing special traffic scenarios and observing whether the vehicle performs a preset operation. Learning abilities can be evaluated by observing the increases in the scores of a vehicle over two tests. The communication ability can be evaluated based on the communication efficiency between the test vehicle and surrounding vehicles. Additional research is needed for the testing and selection of corresponding concrete indexes.

References

[1] HUANG H, ZHENG X, YANG Y, et al. An integrated architecture for intelligence evaluation of automated vehicles [J]. *Accident Analysis & Prevention*, 2020, **145**: 105681.

[2] HUANG H M, PAVEK K, NOVAK B, et al. A framework for autonomy levels for unmanned systems (ALFUS) [C]//*AUVSI's Unmanned Systems North America*. Baltimore: NIST, 2005: 1-9.

[3] DARPA. Urban challenge rules [DB/OL]. (2007-10-27) [2020-11-20]. https://www.grandchallenge.org/ grandchallenge/docs/Urban_Challenge_Rules_102707.pdf.

[4] SUN Y, CHEN H. Research on test and evaluation of unmanned ground vehicles [J]. *Acta Armamentarii*, 2015, **36**(6): 978-986 (in Chinese).

[5] SUN Y, XIONG G, CHEN H. Evaluation of the intelligent behaviors of unmanned ground vehicles based on fuzzy-EAHP scheme [J]. *Automotive Engineering*, 2014, **36**(1): 22-27 (in Chinese).

[6] SUN Y. Quantitative evaluation of intelligence levels for unmanned ground vehicles [D]. Beijing: Beijing Institute of Technology, 2014 (in Chinese).

[7] XIONG G M, GAO L, WU S B, et al. Intelligent behaviors and test and evaluation for unmanned ground vehicles [M]. Beijing: Beijing Institute of Technology Press, 2015 (in Chinese).

[8] SON T D, BHAVE A, VAN DER AUWERAER H. Simulation-based testing framework for autonomous driving development [C]//*2019 IEEE International Conference on Mechatronics*(*ICM*). Ilmenau: IEEE, 2019: 576-583.

[9] WANG G, DENG W, ZHANG S, et al. A comprehensive testing and evaluation approach for autonomous vehicles [J]. *SAE Technical Paper*, 2018: 2018-01-0124.

[10] WENG B, RAO S J, DEOSTHALE E, et al. Model predictive instantaneous safety metric for evaluation of automated driving systems [C]//*IEEE Intelligent Vehicles Symposium* (*IV*). Las Vegas: IEEE, 2020: 1899-1906.

[11] LI L, HUANG W, LIU Y, et al. Intelligence testing for autonomous vehicles: A new approach [J]. *IEEE Transactions on Intelligent Vehicles*, 2016, **1**(2): 158-166.

[12] FENG S, FENG Y, YU C, et al. Testing scenario library generation for connected and automated vehicles, part I: Methodology [J]. *IEEE Transactions on*

Intelligent Transportation Systems, 2021, **22**(3): 1573-1582.

[13] FENG S, FENG Y, SUN H, et al. Testing scenario library generation for connected and automated vehicles, part II: case studies [J]. *IEEE Transactions on Intelligent Transportation Systems*, 2021, **22**(9): 5635-5647.

[14] SAE International. Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles: J3016 [R]. Warrendale: SAE, 2018.

[15] YU Z, XING X, CHEN J. Review on automated vehicle testing technology and its application [J]. *Journal of Tongji University* (*Natural Science*), 2019, **47**(4): 540-547 (in Chinese).

[16] PEGASUS. PEGASUS joint project [DB/OL]. (2019-05-14) [2020-11-20]. http://www.pegasusprojekt.de/en/.

[17] MENZEL T, BAGSCHIK G, MAURER M. Scenarios for development, test and validation of automated vehicles [C]//*2018 IEEE Intelligent Vehicles Symposium*(*IV*). Changshu: IEEE, 2018: 1821-1827.

[18] WATANABE H, TOBISCH L, ROST J, et al. Scenario mining for development of predictive safety functions [C]//*2019 IEEE International Conference on Vehicular Electronics and Safety*(*ICVES*). Cairo: IEEE, 2019: 1-7.

[19] HighD Dataset. HighD dataset [DB/OL]. [2020-11-20]. https: //www.highd-dataset.com/.

[20] DATA.GOV. Safety pilot model deployment data [DB/OL]. (2020-08-21) [2020-11-20]. https://catalog.data.gov/dataset/safety-pilot-model-deployment-data.

[21] UMTRI. Safety pilot model deployment. [DB/OL]. [2020-11-20]. http://safetypilot.umtri.umich.edu/.

[22] RAHMAN M S, ABDEL-ATY M, LEE J, et al. Safety benefits of arterials' crash risk under connected and automated vehicles [J]. *Transportation Research Part C*: *Emerging Technologies*, 2019, 100: 354-371.

[23] CHEN J, LI R, XING X, et al. Survey on intelligence evaluation of autonomous vehicles [J]. *Journal of Tongji University* (*Natural Science*), 2019, **47**(12): 1785-1790 (in Chinese).

[24] ZHAO Y N, MENG K W, GAO L. The entropy-cost function evaluation method for unmanned ground vehicles [J]. *Mathematical Problems in Engineering*, 2015, **2015**: 1-6.

[25] HAYWARD J C. Near-miss determination through use of a scale of danger [C]//*51st Annual Meeting of the Highway Research Board*. Washington, DC: Highway Research Board, 1972: 24-34.

[26] Standardization Administration. Intelligent transportation systems: Forward vehicle collision warning systems: Performance requirements and test procedures: GB/T 33577—2017 [S]. Beijing: Standards Press of China, 2017 (in Chinese).

[27] MINDERHOUD M M, BOVY P H L. Extended time-to-collision measures for road traffic safety assessment [J]. *Accident Analysis & Prevention*, 2001, **33**(1): 89-97.

[28] BAGDADI O, VÁRHELYI A. Jerky driving: An indicator of accident proneness? [J]. *Accident Analysis & Prevention*, 2011, **43**(4): 1359-1363.

[29] BELLEM H, SCHOENENBERG T, KREMS J F, et al. Objective metrics of comfort: developing a driving style for highly automated vehicles [J]. *Transportation Research Part F*: *Traffic Psychology and Behaviour*, 2016, **41**: 45-54.

[30] GUO Y. Comprehensive evaluation theory and method [M]. Beijing: Science Press, 2002 (in Chinese).

[31] XIA Q, LIU K, HUANG L, et al. Study of comprehensive evaluation for L2 automated vehicles on field test [C]//*2019 3rd Conference on Vehicle Control and Intelligence* (*CVCI*). Hefei: IEEE, 2019: 1-6.

[32] WANG X H, LI T J, DING L L. Evaluation theory and technology of complex large-scale systems [M]. Jinan: Shandong University Press, 2010 (in Chinese).