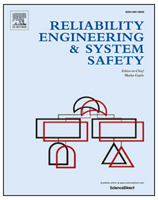
Contents lists available at ScienceDirect

Reliability Engineering and System Safety

journal homepage: www.elsevier.com/locate/ress



Probabilistic risk uncertainty assessment for driver over-trust and under-trust in Level 3 human-automated driving systems cooperative driving based on the drift-diffusion model

Song Ding [a], Lunhu Hu [b], Xing Pan [c,*], Jiacheng Liu [c], Fu Guo [a]

[a] School of Business Administration, Northeastern University, USA
[b] School of Mechanical Engineering, Inner Mongolia University of Technology, Mongolia
[c] School of Reliability & Systems Engineering, Beihang University, Beijing, China

ARTICLE INFO

Keywords:
Human-ADS cooperative driving
Over-trust
Under-trust
DDM
Time-varying risk uncertainty assessment

ABSTRACT

Over-trust in automated driving systems (ADS) can trigger severe accidents, whereas under-trust may reduce system acceptance and efficiency. Thus, assessing risk uncertainty is critical for ensuring driving safety and enhancing system performance. This study aims to develop a cognitive model–based framework for risk uncertainty assessment in human-ADS cooperative driving, enabling precise tracking of the evolving risks of over-trust and under-trust. We propose a drift-diffusion model (DDM)–based risk uncertainty assessment approach applicable across diverse driving task scenarios. A driving simulation experiment was conducted with three levels of ADS reliability and five levels of task difficulty, yielding 7200 behavioral observations for model fitting and validation. The hierarchical Bayesian DDM demonstrated strong predictive performance, with simulated distributions closely matching experimental data. Results reveal that higher ADS reliability significantly shortens trust decision time, while the impact of task difficulty is non-monotonic. More importantly, the model successfully quantifies the time-varying risk uncertainty of over-trust and under-trust. These findings highlight the proposed framework as an effective and interpretable tool for evaluating time-varying risk uncertainty in human-ADS cooperation, providing a crucial model foundation for the future development of real-time risk prediction and intervention systems.

1. Introduction

The advancement of automated driving systems (ADS) has enabled vehicles to perform (part of) the driving tasks traditionally executed by human drivers. This brings several benefits, including improved safety, enhanced comfort, reduced driver workload, and more efficient traffic flow [1–3]. However, due to technical limitations, highly automated driving (Level 4 and above) has not yet been widely deployed. Conditional automation (Level 3) remains the mainstream choice among car manufacturers [4], where human drivers must stay in the loop to continuously monitor the automated driving system and be ready to take over when necessary [5]. This constitutes a typical scenario of human-AI collaborative decision-making [6,7].

Safety is one of the most critical concerns in automated driving [8–10]. Designing risk warning mechanisms to mitigate potential accidents is essential for human-ADS cooperative driving [11–13]. Driver trust in ADS is a key factor influencing whether they choose to take over control from the ADS [14]. Trust has been defined as “the expectation and attitude of the trustor towards the trustee’s ability to achieve goals under conditions of uncertainty and vulnerability” [15,16]. In this definition, trust can be understood as a psychological variable referred to as trust level, while the behavioral outcome of whether to take over from the ADS is a decision, referred to in this study as a trust decision. Furthermore, this definition highlights that trust typically occurs in uncertain and risky environments. A driver’s expectations regarding uncertainty and vulnerability determine whether they continue to rely on ADS or take over to execute the upcoming driving task. Importantly, the goal of trust research is not to maximize trust but to align human trust with system capability, namely “trust calibration” [17–19]. When trust is not properly calibrated, two problematic states may occur: over-trust (trust level exceeds system capability) and under-trust (system capability exceeds trust level) [20,21].

* Corresponding author.
E-mail address: panxing@buaa.edu.cn (X. Pan).

https://doi.org/10.1016/j.ress.2026.112212
Received 19 September 2025; Received in revised form 5 January 2026; Accepted 6 January 2026
Available online 7 January 2026

Recent studies have shown that over-trust is a major cause of catastrophic accidents in automated driving [22]. In such cases, drivers hold overly optimistic expectations of ADS performance and underestimate the risks of potential system failures (e.g., failing to brake in time), which prevents timely takeover actions in critical moments. For example, on May 7, 2016, a Tesla Model S failed to distinguish a white truck from the sky. The driver's over-trust in the system resulted in the lack of takeover, ultimately leading to a fatal accident [8]. Conversely, under-trust can cause drivers to abandon the use of ADS, thereby limiting its potential long-term benefits [23]. Therefore, developing risk assessment methods for over-trust and under-trust and supporting ADS designers in creating risk warning mechanisms are of great significance for reducing potential accidents, improving system safety, and enhancing the overall effectiveness of human-ADS systems.

Risk is generally defined as a combination of event-related uncertainty and its consequences [24–27]. The task of risk assessment is to understand how risk events occur in the system, identify their potential consequences, and, most importantly, express and evaluate their uncertainties [28,29]. Among these, assessing the uncertainty of risk events is the core of risk assessment [30–32]. Traditional human-machine systems employ human reliability analysis (HRA) to evaluate human error risk uncertainty, which fundamentally establishes mapping relationships between performance shaping factors (PSFs) and human error probability (HEP) [33,34]. However, existing HRA methods cannot directly evaluate trust risks in human-ADS collaboration. On one hand, existing HRA methods do not incorporate PSFs specifically designed for over-trust and under-trust events; on the other hand, HRA essentially constitutes a direct evaluation of human macro-level performance, failing to accurately describe the cognitive processes underlying trust risk formation. Notably, a common feature of over-trust and under-trust events is that the driver first makes a trust decision, and the occurrence of the event depends on whether the ADS performs correctly. Therefore, analyzing the risks of over-trust and under-trust fundamentally requires evaluating the uncertainty of drivers' trust decision, that is, the uncertainty regarding whether drivers choose to trust or distrust the ADS within a limited time frame.

1.1. Approaches to trust modeling

Understanding how trust decisions are formed is a prerequisite for assessing their uncertainty. To explain this formation process, researchers in human–machine interaction have developed a variety of trust prediction models that describe both the dynamics of trust evolution and the mechanisms that guide trust-related decision behavior during human–automation interaction. Overall, existing studies can be grouped into two major methodological paradigms:

(1) Physiological signal-driven machine learning approaches

This approach utilizes physiological data such as EEG and eye-tracking to classify trust decisions into binary states [19,35,36]. For example, Ayoub et al. constructed an XGBoost model based on drivers' physiological indicators including galvanic skin response, heart rate, and eye movements, achieving real-time prediction of trust decisions in takeover scenarios with 89.1 % accuracy [19]. Tingru Zhang et al. demonstrated the feasibility of assessing drivers' trust in automated vehicles using electroencephalogram (EEG) signals, with a LightGBM-based model capable of distinguishing low, medium, and high trust states with 88.44 % accuracy [35]. However, these "black-box" models suffer from dual limitations: first, their predictions cannot reveal the cognitive mechanisms behind decision-making; second, they typically treat decisions as instantaneous events, neglecting the fundamental temporal dynamics of cognitive processing during human decision-making.

(2) Behavioral data-driven mathematical modeling approaches

This approach employs mathematical frameworks to describe the dynamic relationship between system reliability and trust levels by fitting experimental data [37–39]. Akash et al. pioneered a third-order linear time-invariant model based on gray-box system identification, successfully quantifying the interactions among three key state variables in driver trust dynamics: experience accumulation, accumulated trust, and expectation deviation [36]. Rabby et al. proposed a time-driven and performance-aware mathematical trust model that characterizes the nonlinear evolution of trust with human-machine performance differences using piecewise functions and hyperbolic tangent functions [38]. Seo and Kia developed a joint inference framework based on Bayesian online learning, which integrates target state estimation with human trust learning into a unified probabilistic graphical model through a Hidden Markov Model (HMM), enabling dynamic perception and updating of trust levels [39]. Although these models have made progress in characterizing long-term trust evolution, their core limitation lies in focusing solely on the continuous changes in trust levels while failing to model the essential feature of driving scenarios: completing trust decisions within bounded time windows.

Despite providing important foundations for understanding trust formation, neither approach addresses the core requirement of trust risk uncertainty assessment: quantifying the uncertainty of trust decision outcomes within specified time windows. This gap stems from two fundamental deficiencies: first, the lack of formalized descriptions of cognitive mechanisms; second, the inability to incorporate decision time distributions into risk uncertainty calculation frameworks. These elements are precisely the key factors for assessing the occurrence probability of over-trust/under-trust events.

1.2. The necessity of using drift-diffusion model to assess uncertainty of trust decision

To assess the uncertainty of trust decisions within a limited time window, it is necessary to model not only "what decision is made" but also "when the decision occurs", as well as the cognitive mechanisms mediating this temporal process. The drift diffusion model (DDM), originally developed in cognitive psychology to explain two-alternative forced-choice decision behavior, provides a mechanistic framework that describes how evidence accumulates over time until a threshold is reached and a decision is triggered [40,41]. Importantly, the DDM can simultaneously predict choice accuracy and reaction time, making it particularly suitable for quantifying the uncertainty of trust decisions within limited time windows, which is the core of trust risk uncertainty assessment.

Compared with existing trust modeling methods, using DDM to model trust decisions offers two core advantages. First, the DDM directly characterizes the cognitive process of trust decision-making through interpretable parameters, avoiding the explainability limitations of "black-box" machine learning models. Second, the response time distribution generated by the DDM enables us to calculate the probability of a driver making a trust decision within a limited time window. Coupling the DDM-predicted decision time distribution with ADS task reliability allows for a rigorous derivation of the probability of over-trust or under-trust risk events, which existing models are unable to achieve.

Therefore, this study pursues two objectives. First, we model the formation of driver trust decisions in automated driving using the DDM, which has been shown to effectively capture both accuracy and decision time in binary decision tasks. Second, building on this model, we assess the time-varying risk uncertainty of over-trust and under-trust events. Specifically, we designed a human-ADS cooperative driving experiment in a simulator, manipulating ADS reliability and task difficulty. Behavioral data were collected and analyzed using hierarchical Bayesian methods to estimate model parameters. We then evaluated the uncertainty of drivers' trust decisions within limited time frames and combined this with the ADS task reliability to compute the risk uncertainty of over-trust and under-trust events. To the best of our knowledge, this is the first study to systematically evaluate and model the time-varying risk uncertainties arising from over-trust and under-trust in human-ADS cooperative driving.

The main contributions of this work are as follows:

(1) We propose a risk uncertainty assessment framework for over-trust and under-trust in human–ADS cooperative driving.
(2) We characterize the cognitive process of driver trust decision using the DDM and quantify the uncertainty in typical driving tasks.
(3) We design a human-ADS cooperative driving experiment incorporating ADS reliability and task difficulty, and explore their impacts on the time-varying risk uncertainty of over-trust and under-trust within the proposed framework.

The remainder of this paper is organized as follows. Section 2 describes the proposed methodology, while Section 3 details the experimental design. Section 4 presents the results. Section 5 discusses the findings, contributions, limitations, and directions for future research. Finally, Section 6 concludes the study. Fig. 1

2. Method

2.1. Assessment of driver trust decision uncertainty using DDM

Conditional automated driving represents a typical scenario of human-AI collaborative decision-making. The ADS executes tasks (e.g., obstacle avoidance) based on perceived information, while the driver evaluates ADS reliability under the current task difficulty and decides whether to trust the system or intervene [42]. This process can be characterized as a binary decision. In psychology, evidence accumulation models provide a detailed cognitive account of decision-making by considering both choice and decision time, and have been widely used to explain decisions under uncertainty [43,44]. During human-ADS cooperative driving tasks, the driver is faced with a binary decision of whether to trust the ADS. In such uncertain environments, trust decisions are formed through evidence accumulation over time. The drift diffusion model, a classical evidence accumulation framework, has been shown to accurately capture both choice and reaction time in binary decision tasks [40,41]. The core assumption of the DDM is that a decision-maker begins at an initial bias point z, accumulates evidence at a constant drift rate v over time, and makes a decision once the accumulated evidence reaches a boundary defined by the threshold a. The accumulated evidence can be conceptualized as the trajectory of a diffusion particle, which triggers a response once it crosses a boundary.

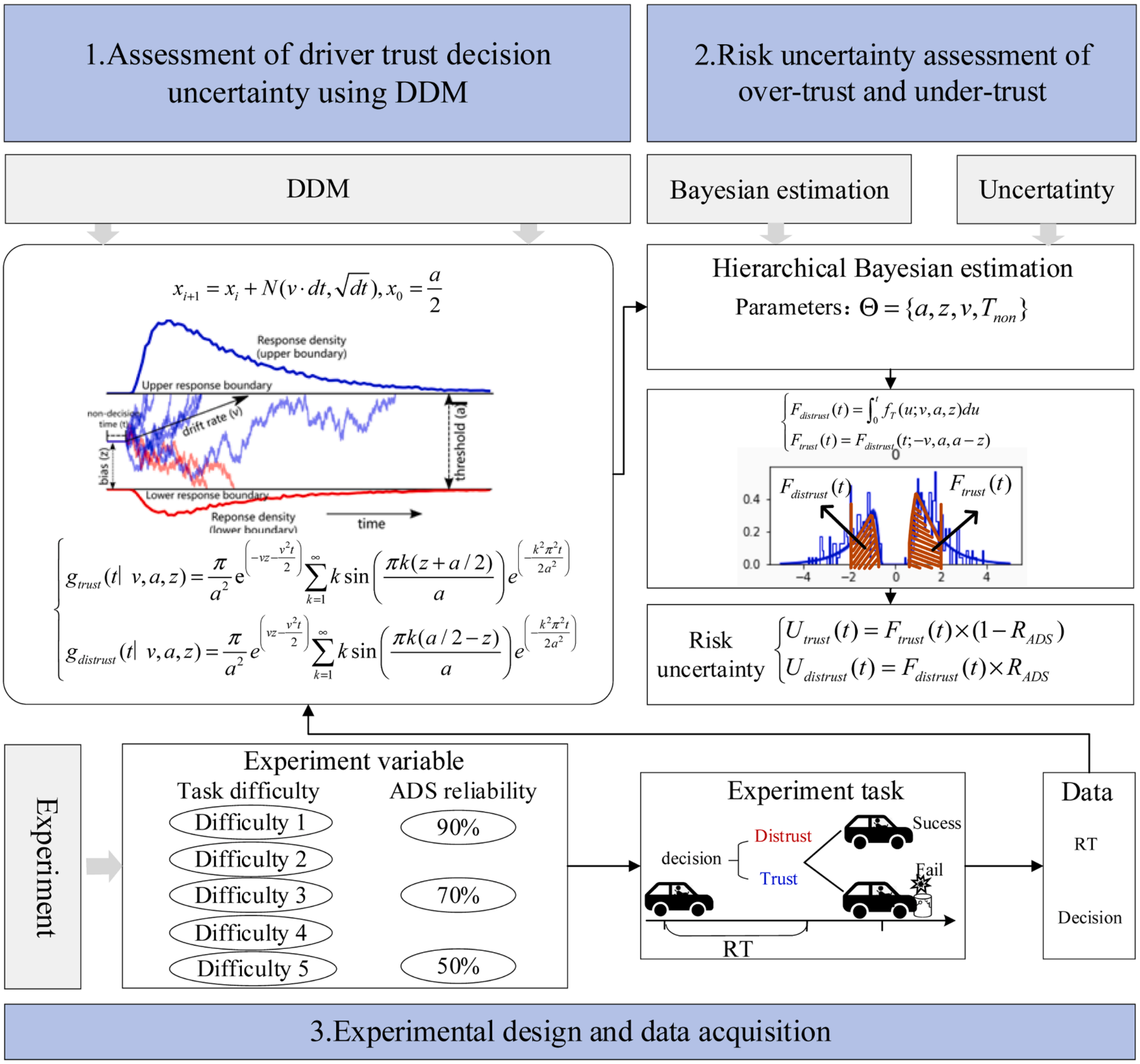

1.Assessment of driver trust decision uncertainty using DDM
2.Risk uncertainty assessment of over-trust and under-trust
DDM
Bayesian estimation
Uncertainty
$x_{i+1} = x_i + N(v \cdot dt, \sqrt{dt}), x_0 = \frac{a}{2}$
Response density (upper boundary)
Upper response boundary
non-decision time (t)
drift rate (v)
threshold (a)
bias (z)
Lower response boundary
Reponse density (lower boundary)
time
$g_{trust}(t|\ v,a,z) = \frac{\pi}{a^2} e^{\left(-vz-\frac{v^2 t}{2}\right)} \sum_{k=1}^{\infty} k \sin\left(\frac{\pi k(z+a/2)}{a}\right) e^{\left(-\frac{k^2\pi^2 t}{2a^2}\right)}$
$g_{distrust}(t|\ v,a,z) = \frac{\pi}{a^2} e^{\left(vz-\frac{v^2 t}{2}\right)} \sum_{k=1}^{\infty} k \sin\left(\frac{\pi k(a/2-z)}{a}\right) e^{\left(-\frac{k^2\pi^2 t}{2a^2}\right)}$
Hierarchical Bayesian estimation
Parameters: $\Theta = \{a, z, v, T_{non}\}$
$F_{distrust}(t) = \int_0^t f_T(u; v, a, z) du$
$F_{trust}(t) = F_{distrust}(t; -v, a, a-z)$
0
$F_{distrust}(t)$
$F_{trust}(t)$
0.4
0.2
0.0
−4
−2
0
2
4
Risk uncertainty
$U_{trust}(t) = F_{trust}(t) \times (1 - R_{ADS})$
$U_{distrust}(t) = F_{distrust}(t) \times R_{ADS}$
Experiment
Experiment variable
Task difficulty
ADS reliability
Difficulty 1
Difficulty 2
Difficulty 3
Difficulty 4
Difficulty 5
90%
70%
50%
Experiment task
decision
Distrust
Trust
Sucess
Fail
RT
Data
RT
Decision
3.Experimental design and data acquisition

Fig. 1. The framework of the methodology.

Key parameters of the model are:

(1) Drift rate (ν): average rate of evidence accumulation; reflects the strength of evidence favoring one decision. A higher drift rate indicates faster accumulation.
(2) Threshold (a): distance between upper and lower boundaries; smaller thresholds facilitate quicker decisions.
(3) Bias (z): starting point of evidence accumulation, typically between the boundaries; represents an a priori decision tendency. To more clearly capture the decision bias, the relative value$z_{obs} = z/a \in [0, 1]$ is often used.
(4) Non-decision time (T_{non}): accounts for processes outside decision-making, such as sensory encoding and motor execution.

The decision process is illustrated in Fig. 2. Multiple drift trajectories (blue and red) show evidence accumulation starting at bias z along the vertical axis, progressing at drift rate ν until reaching either boundary, with the separation determined by threshold a. The upper and lower areas represent the probability density functions of boundary crossing times, which align well with observed reaction time distributions.

Formally, evidence accumulation within a trial can be expressed as:

$$x_{i+1} = x_i + N\left(\nu \cdot dt, \sqrt{dt}\right), x_0 = z \tag{1}$$

where x_i is the accumulated evidence at iteration i, and a response is triggered when $x_i \geq a$ (trust) or $x_i \leq a$ (distrust). Here, $N(\cdot)$ denotes a normal distribution, t is decision time, and dt→0 ensures a continuous process.

Decision times follow the wiener first-passage time distribution, the probability density function is as following [46]:

$$\begin{cases} g_{trust}(t|v,a,z) = \dfrac{\pi}{a^2}\mathrm{e}^{\left(-vz-\frac{v^2 t}{2}\right)} \displaystyle\sum_{k=1}^{\infty} k\sin\left(\dfrac{\pi k(z+a/2)}{a}\right)e^{\left(-\frac{k^2\pi^2 t}{2a^2}\right)} \\ g_{distrust}(t|v,a,z) = \dfrac{\pi}{a^2}e^{\left(vz-\frac{v^2 t}{2}\right)} \displaystyle\sum_{k=1}^{\infty} k\sin\left(\dfrac{\pi k(a/2-z)}{a}\right)e^{\left(-\frac{k^2\pi^2 t}{2a^2}\right)} \end{cases} \tag{2}$$

As this distribution involves infinite series, it is often intractable in practice. Therefore, we adopted the approximation method proposed by Navarro and Fuss [47], which provides efficient probability density approximations for both small and large decision times:

$$f_T(t; v, a, z) = \frac{1}{a^2} e^{-vaw-\frac{1}{2}v^2 t} f_{\mathrm{std}}\left(\frac{t}{a^2}|w\right), \quad w = \frac{z}{a} \tag{3}$$

When t is a small-time:

$$f_{std\_small\_time}(t|w) = \frac{1}{\sqrt{2\pi}} \frac{1}{t^{3/2}} \sum_{k=-\infty}^{\infty} (w+2k)e^{\left(-\frac{(w+2k)^2}{2t}\right)} \tag{4}$$

When t is a large-time:

$$f_{std\_large\_time}(t|w) = \pi \sum_{k=1}^{\infty} k\sin(k\pi w)e^{\left(-\frac{k^2\pi^2}{2}t\right)} \tag{5}$$

where $k = 1, 2, 3 \ldots K,$ K is the number of truncation terms. The key quantity for determining whether t is considered “small” or “large” is given by$\widetilde{t} = t/a^2$. If $\widetilde{t} < 0.01$, the expression $f_{\mathrm{std\_small\_time}}(t|w)$ is used for computation; if $\widetilde{t} > 0.2$, $f_{\mathrm{std\_large\_time}}(t|w)$ is used. For values in the interval $0.01 < \widetilde{t} < 0.2$,the magnitude of the last term in both $f_{\mathrm{std\_small\_time}}(t|w)$ and $f_{\mathrm{std\_large\_time}}(t|w)$ is evaluated, and the expansion corresponding to the one with the smaller magnitude (indicating faster convergence) is selected. Series were truncated at $K = 20$, which yields numerical errors below 10^{-8} across our parameter ranges [47].

Integrating the density function gives the probability of trust and distrust decisions within time t:

$$\begin{cases} F_{distrust}(t) = \displaystyle\int_0^t f_T(u; v, a, z)du \\ F_{trust}(t) = F_{distrust}(t; -v, a, a-z) \end{cases} \tag{6}$$

Notably, at time t, the sum of the probabilities for making a trust decision and a distrust decision does not equal 1. Consequently, there exists a non-decision probability:

$$S(t) = 1 - F_{trust}(t) - F_{distrust}(t) \tag{7}$$

In our model, T_{non} assumed constant, so the observed reaction time is $t = T_{non} + T_{er}$, where T_{er} is the effective decision time.

As t→∞, the ultimate decision probability converges to:

$$\begin{cases} F_{trust}(t \to \infty) = \dfrac{1 - e^{-2vz}}{1 - e^{-2va}} \\ F_{distrust}(t \to \infty) = 1 - P_{trust}(t \to \infty) \end{cases} \tag{8}$$

The DDM thus provides a principled framework to model the cognitive mechanism of trust decision. By explicitly incorporating decision time, it enables the interpretation of complex behavioral patterns such as premature or delayed decisions.

2.2. Risk uncertainty assessment of over-trust and under-trust

The core of assessing the risk uncertainty of driver over-trust and under-trust lies in evaluating the uncertainty of trust and distrust decisions within a given time t, as expressed in Eq. (6). Therefore, once the parameters of the driver's DDM under typical task scenarios are obtained, the decision uncertainty can be computed directly.

In this study, the parameter set $\Theta = \{a, z, v, T_{non}\}$ is estimated using a hierarchical Bayesian model. Model fit and predictive accuracy were further evaluated using posterior predictive checks (PPCs) [48].

Given the estimated DDM parameters for trust decisions in typical driving tasks, the uncertainty of trust and distrust decisions at time t can be expressed as $F_{trust}(t)$ and $F_{distrust}(t)$, respectively.

Based on the definitions of over-trust and under-trust, two types of risk events are formalized:

1. Over-trust risk event: the driver decides to trust the ADS, but the ADS task fails.
2. Under-trust risk event: the driver decides not to trust the ADS, but the ADS task succeeds.

Here is a polished version with improved clarity and academic tone:

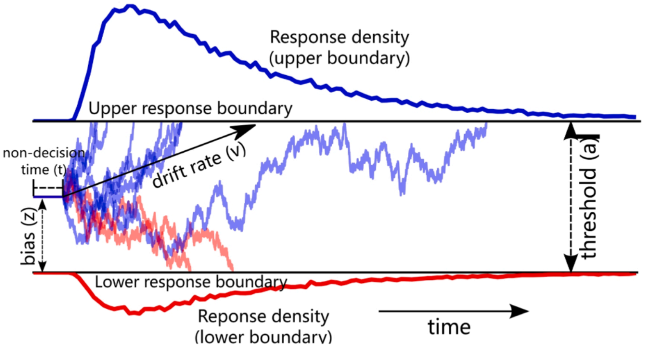


Fig. 2. Parameters of the DDM and their explanations [45].

Under the controlled experimental setting, where ADS outcomes are externally specified and temporally subsequent to the driver's trust decision, we assume conditional independence between the trust decision and the ADS task outcome. This assumption is justified because the trust decision is governed solely by the driver's internal evidence-accumulation process (as captured by the DDM), while the ADS outcome depends exclusively on the system's predefined reliability and is unaffected by the driver's cognitive state or choice. Under this assumption, the joint probabilities can be expressed as the product of these two independent components. Accordingly, the risk uncertainty at time t of over-trust and under-trust risk events at time t can be quantified as:

$$\begin{cases} U_{trust}(t) = F_{trust}(t) \times (1 - R_{ADS}) \\ U_{distrust}(t) = F_{distrust}(t) \times R_{ADS} \end{cases} \tag{9}$$

where R_{ADS} denote the task reliability of ADS, i.e., the probabilities of successful task completion under the given driving scenario.

3. Experiment

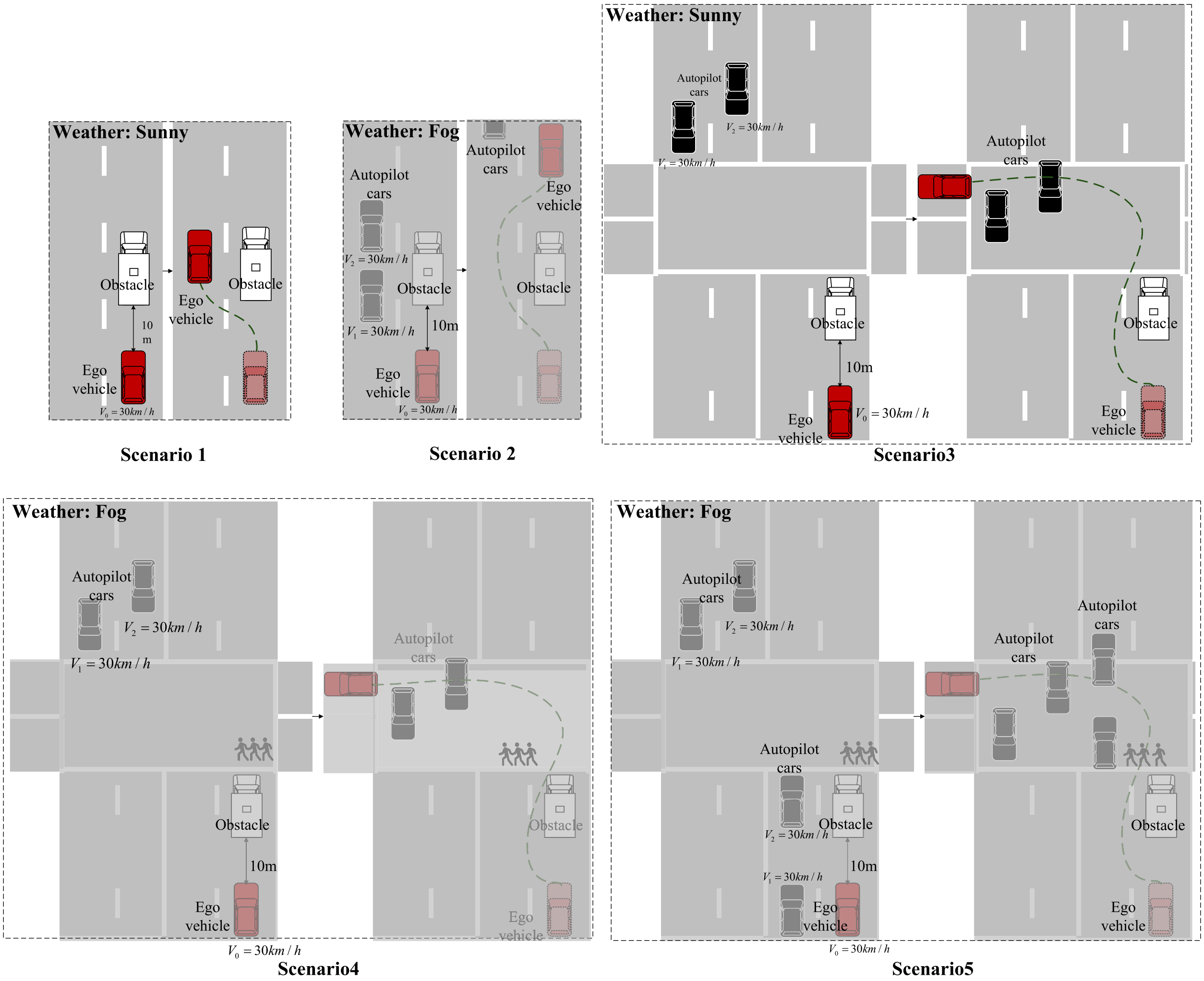
3.1. Experimental apparatus

The study was conducted in a laboratory setting. A virtual traffic environment was constructed and implemented using the CARLA simulation platform [49]. The driving simulator system consisted of a Logitech G923 racing peripheral (including a steering wheel and pedals), a GPU with 16 GB of VRAM, and a 34-inch display monitor.

3.2. Experimental scenario

Hoff and Bashir classified trust into dispositional trust, situational trust, and learned trust [50]. This study focuses on situational and learned trust. Situational trust is event-specific and influenced by the task context, while the accumulation of such events shapes learned trust. In automated driving, task difficulty is the primary factor affecting situational trust, and it is determined largely by road and weather conditions. Learned trust, on the other hand, is mainly influenced by ADS reliability. Therefore, task difficulty and ADS reliability were selected as the experimental variables in this study.

To operationalize these constructs, we designed five distinct driving scenarios that systematically vary in task difficulty, representing increasing levels of perceptual, cognitive, and control demands. Each scenario corresponds to a unique combination of static elements (e.g., weather and visibility), dynamic elements (e.g., traffic density and pedestrian activity), and maneuver complexity (e.g., straight driving vs. left turns across oncoming traffic) [51,52]. These five scenarios, illustrated in Fig. 3, were implemented in the CARLA autonomous driving simulator and serve as the basis for manipulating situational trust. As As summarized in Table 1, they form a graded difficulty scale, enabling us to examine how contextual complexity influences drivers' trust decisions under controlled conditions.



Scenario 1

Scenario 2

Scenario3

Scenario4

Scenario5

Fig. 3. Five experimental driving scenarios.

Table 1
Contextual elements across varying task difficulty levels.

Elements \Difficulty	Static elements	Dynamic elements	Control complexity
Scenario 1	Weather: Sunny Visibility: High Traffic light: None	Number of vehicles: 1. A disabled vehicle is ahead. Number of pedestrians: 0	Direction of travel: Straight ahead
Scenario 2	Weather: Fog Visibility: Low Traffic light: None	Number of vehicles: 3. A disabled vehicle is ahead, with two cars directly in front in the neighboring lane. Number of pedestrians: 0	Direction of travel: Straight ahead
Scenario 3	Weather: Fog Visibility: Low Traffic light: Yes	Number of vehicles: 3. A disabled vehicle is ahead, with two oncoming vehicles in the diagonal left-turn lane. Number of pedestrians: 3. And pedestrian crossing when red.	Direction of travel: Turn left
Scenario 4	Weather: Fog Visibility: Low Traffic light: Yes	Number of vehicles: 3. A disabled vehicle is ahead, with two oncoming vehicles in the diagonal left-turn lane. Number of pedestrians: 3. And pedestrian crossing when red.	Direction of travel: Turn left
Scenario 5	Weather: Fog Visibility: Low Traffic light: Yes	Number of vehicles: 5. A disabled vehicle is ahead, two oncoming vehicles in the diagonal left-turn lane, and two cars directly in front in the neighboring lane. Number of pedestrians: 3. And pedestrian crossing when red.	Direction of travel: Turn left

(1) Scenario 1

As shown in Fig. 4, the scenario is set under sunny weather with high visibility. A disabled vehicle is located 10 m ahead of the ego vehicle. The driving task requires the ego vehicle, starting at 30 km/h, to avoid the obstacle and proceed straight through the intersection.

(2) Scenario 2

As shown in Fig. 5, the scenario involves rainy and foggy weather with low visibility. Two ADS-driven vehicles are traveling straight ahead in the left lane at 30 km/h, while a stationary disabled vehicle is located 10 m in front of the ego vehicle. The ego vehicle, starting at 30 km/h, must avoid the obstacle, change lanes, follow the preceding vehicle, and continue straight to the intersection.

(3) Scenario 3

As shown in Fig. 6, the scenario is under sunny weather with high visibility. Two ADS-driven vehicles are approaching the intersection from the opposite lane at 30 km/h. A disabled vehicle is located 10 m ahead of the ego vehicle, and several pedestrians are jaywalking at the intersection. The ego vehicle, starting at 30 km/h, must avoid the obstacle and complete a left turn through the intersection.

(4) Scenario 4

As shown in Fig. 7, the scenario involves rainy and foggy weather with low visibility. Two ADS-driven vehicles are approaching the intersection from the opposite lane at 30 km/h. A disabled vehicle is located 10 m ahead, and multiple pedestrians are jaywalking at the intersection. The ego vehicle, starting at 30 km/h, must avoid the obstacle and complete a left turn through the intersection.

(5) Scenario 5

As shown in Fig. 8, the scenario is set under rainy and foggy weather with low visibility. Two ADS-driven vehicles are approaching the intersection from the opposite lane at 30 km/h. In addition, one ADS-driven vehicle is traveling ahead in the adjacent left lane, and another is following behind in the same lane. A disabled vehicle is located 10 m ahead of the ego vehicle, and several pedestrians are jaywalking at the intersection. The ego vehicle, starting at 30 km/h, must avoid the obstacle, change lanes to follow the vehicle in the left lane, and subsequently make a left turn through the intersection.

In addition to scenario-based task difficulty, ADS reliability was introduced as a second experimental factor to modulate learned trust. Three levels of system reliability, high (90 %), medium (70 %), and low

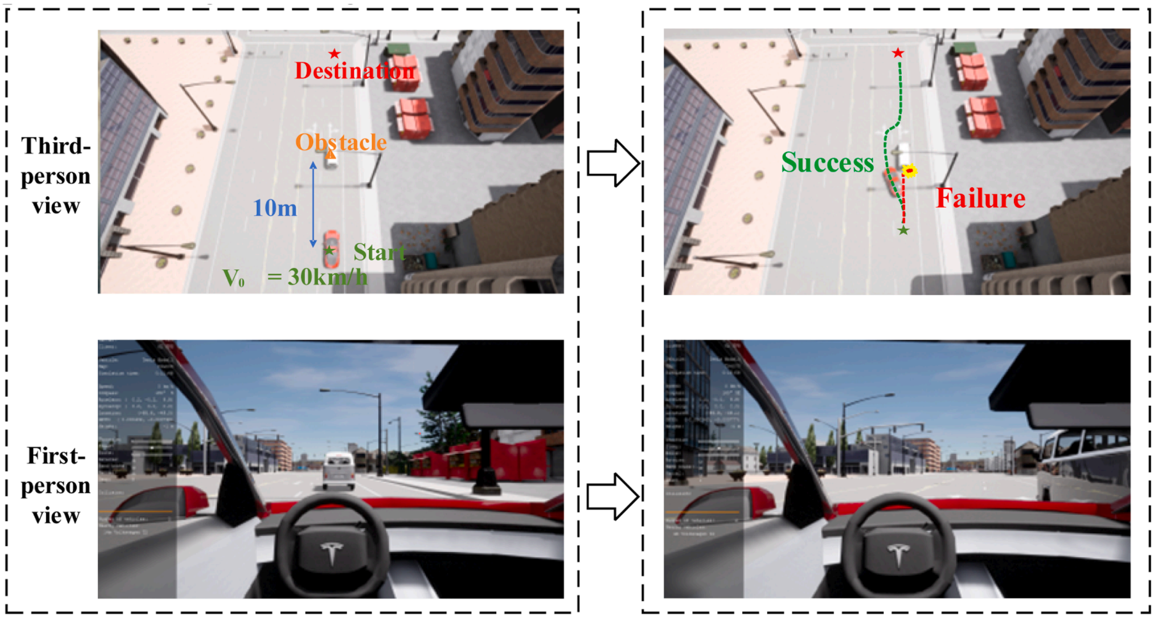


Fig. 4. Obstacle avoidance scenario setting for task difficulty 1.

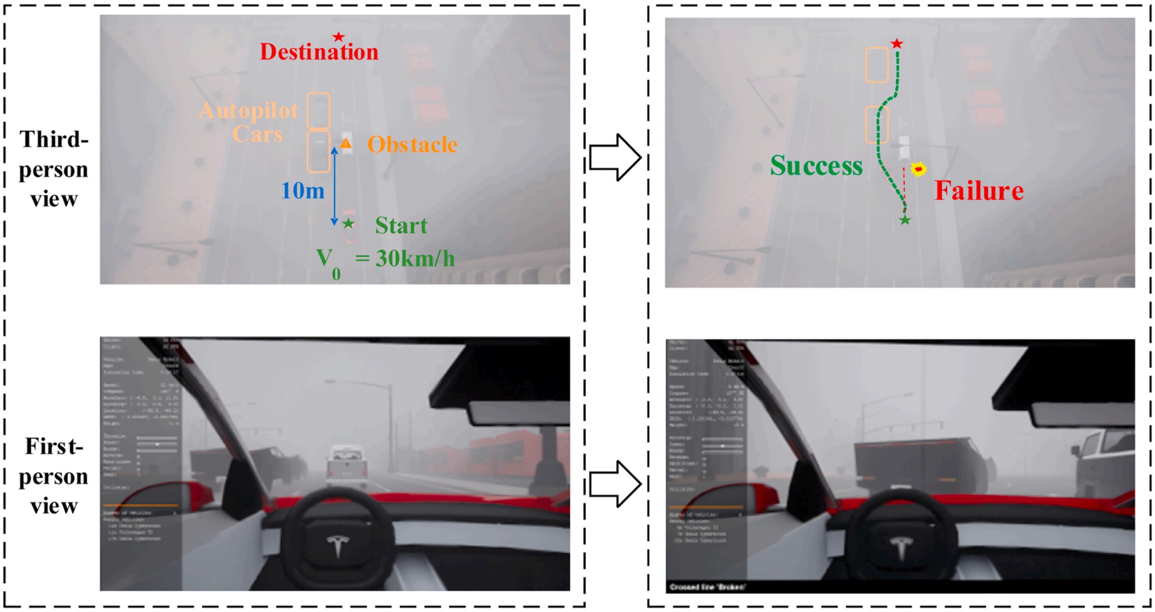


Fig. 5. Obstacle avoidance scenario setting for task difficulty 2.

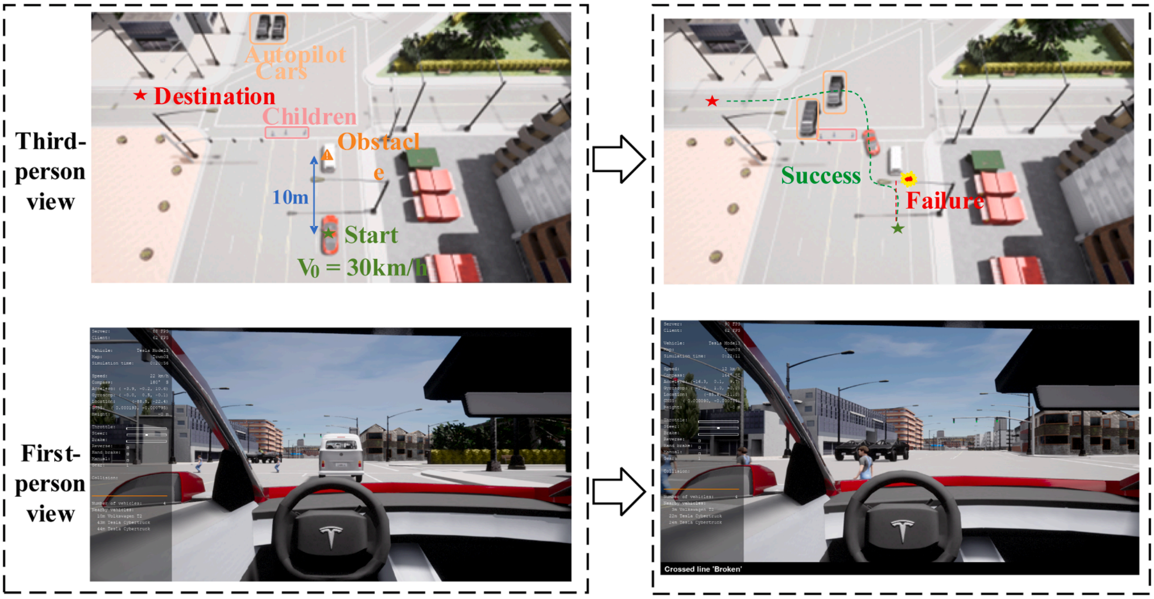


Fig. 6. Obstacle avoidance scenario setting for task difficulty 3.

(50 %), were implemented across the experiment. The combination of five driving scenarios and three reliability levels defined the full set of experimental conditions used in the study.

3.3. Experimental design

The primary objective of the experimental design was to establish a controlled paradigm that isolates and quantifies the core cognitive process underlying a driver's binary trust decision, rather than replicating the continuous monitoring demands of a real-world Level 3 driving scenario. By decomposing the complex, ongoing decision process, this approach enables precise examination of the critical cognitive unit of interest: the mechanism and timing through which a driver forms a "trust" or "distrust" judgment after receiving sufficient information about a driving situation. We argue that a foundational understanding of this discrete decision component is essential for modeling how it manifests within more dynamic, real-time human–automation interactions.

The experiment followed a 5 (task difficulty: Scenario 1–5) × 3 (ADS reliability: 90 %, 70 %, 50 %) within-subjects design. Each participant completed three experimental blocks, one for each reliability level. Within each block, all five scenarios were presented repeatedly in randomized order, resulting in 80 trials per block (16 trials per scenario on average).

Participants were informed that ADS performance would vary across blocks and that, within each block, the likelihood of system failure would increase with task difficulty. For example, under a block-level reliability of 70 %, the ADS might successfully complete 90 % of easy trials but only 50 % of the most difficult ones, while still maintaining an overall success rate of 70 % for that block. Importantly, the exact reliability level for each block was not revealed beforehand, and the order of blocks was counterbalanced across participants.

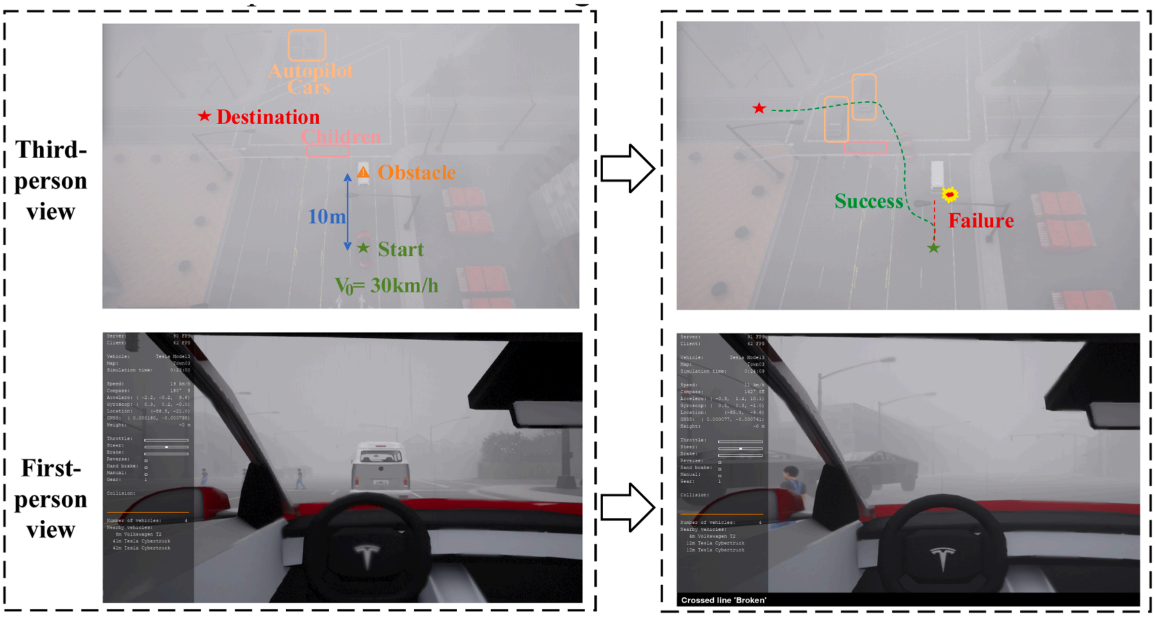


Fig. 7. Obstacle avoidance scenario setting for task difficulty 4.

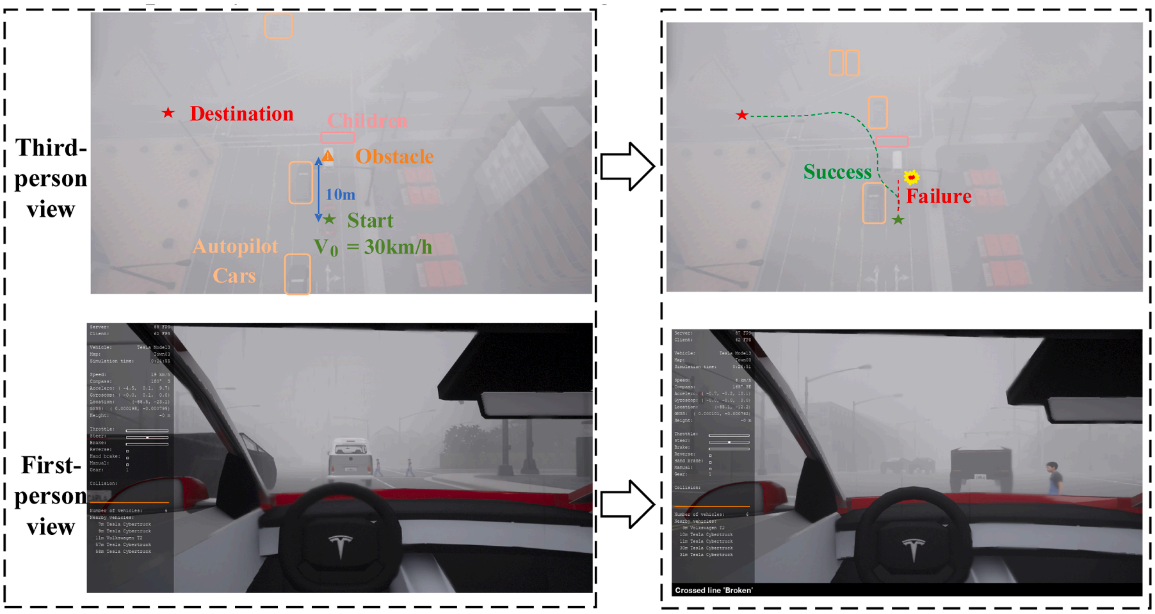


Fig. 8. Obstacle avoidance scenario setting for task difficulty 5.

In total, each participant performed 240 trials (80 × 3), yielding 7200 trials across all 30 participants. The experiment was implemented using the Expyriment package in Python [53], and the trial procedure is illustrated in Fig. 9.

(1) **Step 1**: Each trial began with a fixation cross (“+”) displayed at the center of the screen.
(2) **Step 2**: After pressing any button on the steering wheel device, participants viewed a video stimulus of an ADS task generated with the CARLA platform. The video ended once the ADS reached the designated location.
(3) **Step 3**: Participants were required to make a trust decision within 6 s. Trust was indicated by pressing the “Δ” button; distrust by pressing the “ × ” button, and decision times were recorded.
(4) **Step 4**: After the decision, a video showing whether the ADS successfully completed the task was presented, followed by outcome feedback.

The experiment was designed to conceptualize trust as a risk-sensitive decision process, rather than treating it as an abstract attitude. To achieve this, we implemented an asymmetric reward structure that captures the real-world costs and benefits associated with trusting an ADS:

- **Trusting a successful ADS (reward: +10):** reflects the advantage of effective human–ADS cooperation.
- **Trusting a failing ADS (penalty: –10):** represents the serious negative consequence of misplaced trust (e.g., potential accidents).

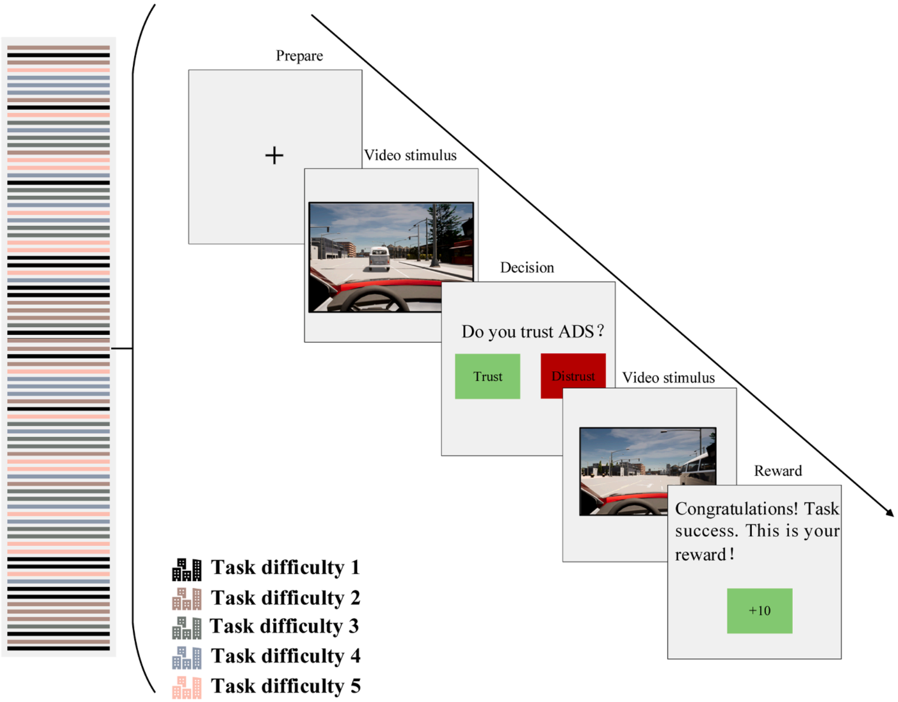


Fig. 9. Experimental design in each trial.

- **Distrusting a failing ADS (reward: +5):** rewards correct intervention, though less than seamless cooperation, to reflect the effort and disruption involved.
- **Distrusting a successful ADS (reward: 0):** imposes a minor opportunity cost for unnecessary caution.

This reward structure explicitly motivates participants to integrate their perception of ADS reliability with the potential outcomes of their choices, ensuring that "trust" responses reflect a realistic risk-management strategy. Participants were instructed to maximize cumulative rewards, mirroring a driver's objective of optimizing both safety and efficiency in real-world driving.

3.4. Participants

A total of 30 participants (15 males and 15 females) from Beihang University participated in this study. All participants were right-handed, had normal or corrected-to-normal vision, reported no history of neurological or psychiatric disorders, and were 24 ± 4 years old on average. Each participant possessed prior driving experience. Before the experiment, informed consent was obtained, and participants were notified that they would receive monetary compensation ranging from 60 to 80 RMB, depending on task performance.

4. Results

4.1. Influence of ADS reliability and task difficulty on trust decision time

A total of 7200 trust decision trials were collected, encompassing five task difficulty levels and three ADS reliability conditions. The decision time data were grouped by ADS reliability and task difficulty, and *t-tests* were conducted to examine between-group differences. Bar plots were generated to illustrate the results, with $p < 0.05$, $p < 0.01$, and $p < 0.001$ denoted by "*", "**", *and* "***", respectively (see Fig. 10 and Fig. 11).

Fig. 10 shows the effect of ADS reliability on trust decision time. Results indicate that decision time decreased as ADS reliability increased, with all pairwise comparisons reaching statistical significance ($p < 0.001$). By contrast, Fig. 11 shows the effect of task difficulty. No consistent monotonic trend was observed: decision time increased slightly at difficulty level 2, then gradually declined with higher difficulty, but differences across levels were not statistically significant. For example, no significant difference was found between levels 3 and 4, or between levels 4 and 5.

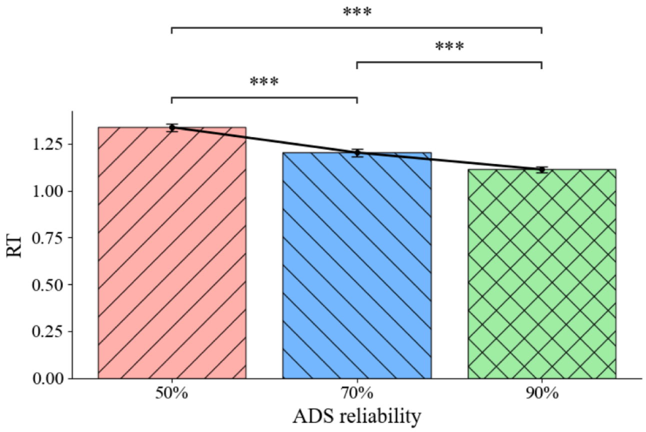


Fig. 10. Influence of ADS reliability on trust decision time.

4.2. Model fitting and parameter interpretation

The data were divided into 15 groups (3 ADS reliability levels × 5 task difficulty levels). A hierarchical Bayesian framework with MCMC inference was applied to fit the DDM parameters for each group. Model convergence was assessed using the Gelman–Rubin statistic ($\hat{R}$), where values close to 1 indicate convergence. Conventionally, $\hat{R} < 1.1$ is considered acceptable. For all 15 models, $\hat{R}$ values were below 1.1, confirming convergence. The parameter estimates are reported in Appendix A. Among the estimated parameters, the drift rate (υ) is the primary determinants of trust decisions and response times. Their mean values across experimental conditions are summarized in Fig. 12 and

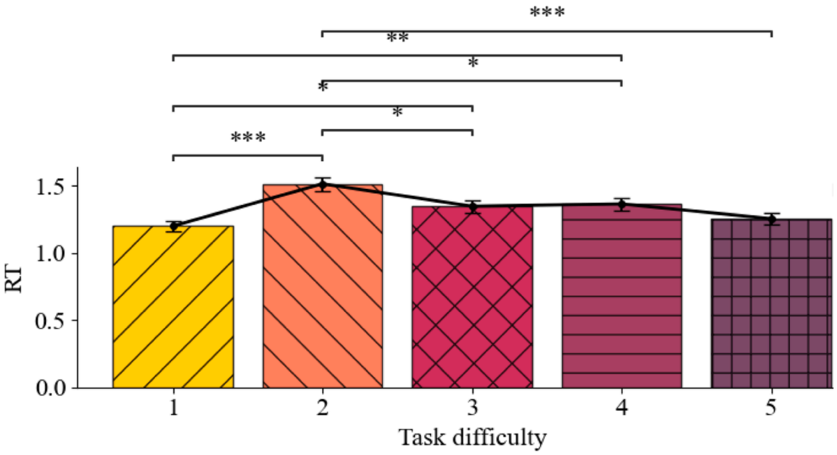


Fig. 11. Influence of task difficulty on trust decision time.

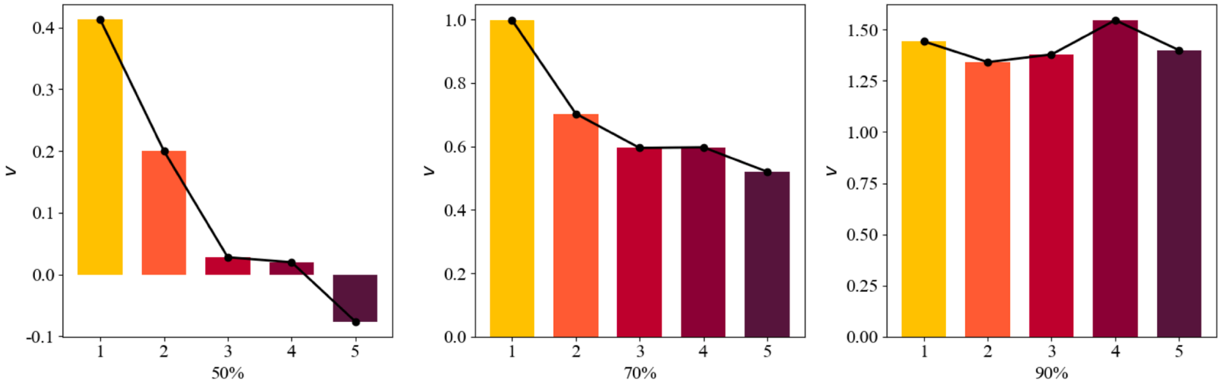


Fig. 12. Variation of parameter v with task difficulty under different ADS reliability.

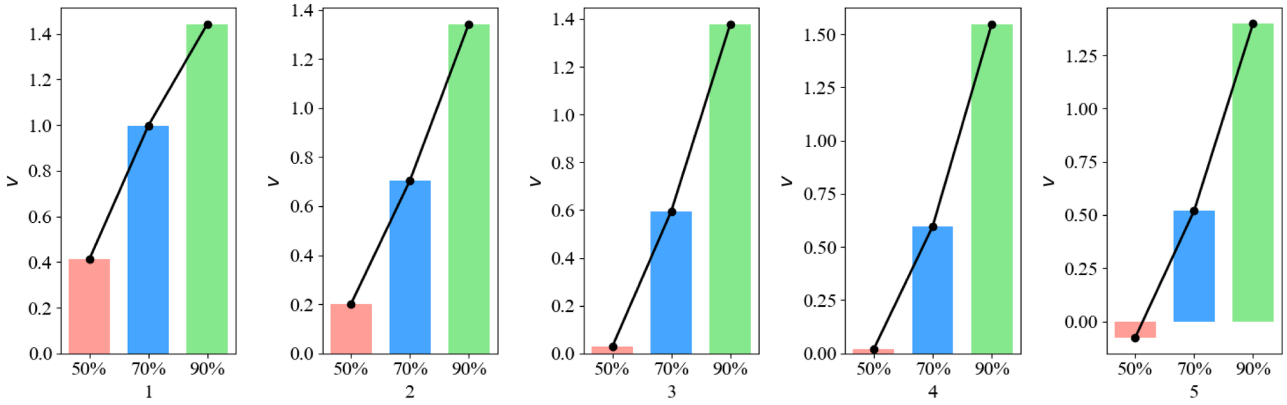


Fig. 13. Variation of parameter v with ADS reliability under different task difficulty.

Fig. 13.

Fig. 12 illustrates how v varied with task difficulty under three reliability levels (50 %, 70 %, and 90 %). When ADS reliability was low (50 %), v decreased markedly with increasing task difficulty, suggesting that participants were less likely to trust the system in more challenging scenarios. A similar but less pronounced trend was observed under medium reliability (70 %). In contrast, no such trend emerged under high reliability (90 %). A plausible explanation is that high reliability fosters habitual reliance on ADS, whereas under lower reliability, increasing task complexity further reduces participants' willingness to trust, given the reduced system accuracy. This discrepancy also explains why trust decision time (Fig. 11) did not exhibit a consistent relationship with task difficulty.

Fig. 13 shows the relationship between v and ADS reliability at each task difficulty level. Across all difficulty levels, v increased with higher ADS reliability, indicating a greater likelihood of trust decisions and shorter decision times. These results are consistent with the statistical analyses in Fig. 10, confirming that the DDM not only provides an excellent fit to observed trust decision times but also offers a mechanistic explanation through its parameters.

4.3. PPCs of trust decision time

Posterior predictive checks were conducted to evaluate the goodness-of-fit of the estimated DDMs. Fig. 14 compares the probability density of trust decision times generated from the PPCs with the empirical data across different task scenarios. The results demonstrate that the DDMs closely reproduced the observed decision time distributions, indicating strong model validity.

From a practical perspective, the fitted DDM parameters provide a direct means to evaluate the uncertainty of trust decisions. Specifically, given a task scenario defined by task difficulty and ADS reliability, the corresponding DDM parameters can be applied in Eq. (6) to compute the probability uncertainty of making either a trust or distrust decision within a given time window *t*. This approach enables efficient quantification of trust decision uncertainty under diverse operational conditions.

4.4. Risk uncertainty assessment results of over-trust and under-trust

Based on the DDM parameters estimated from experimental data, we

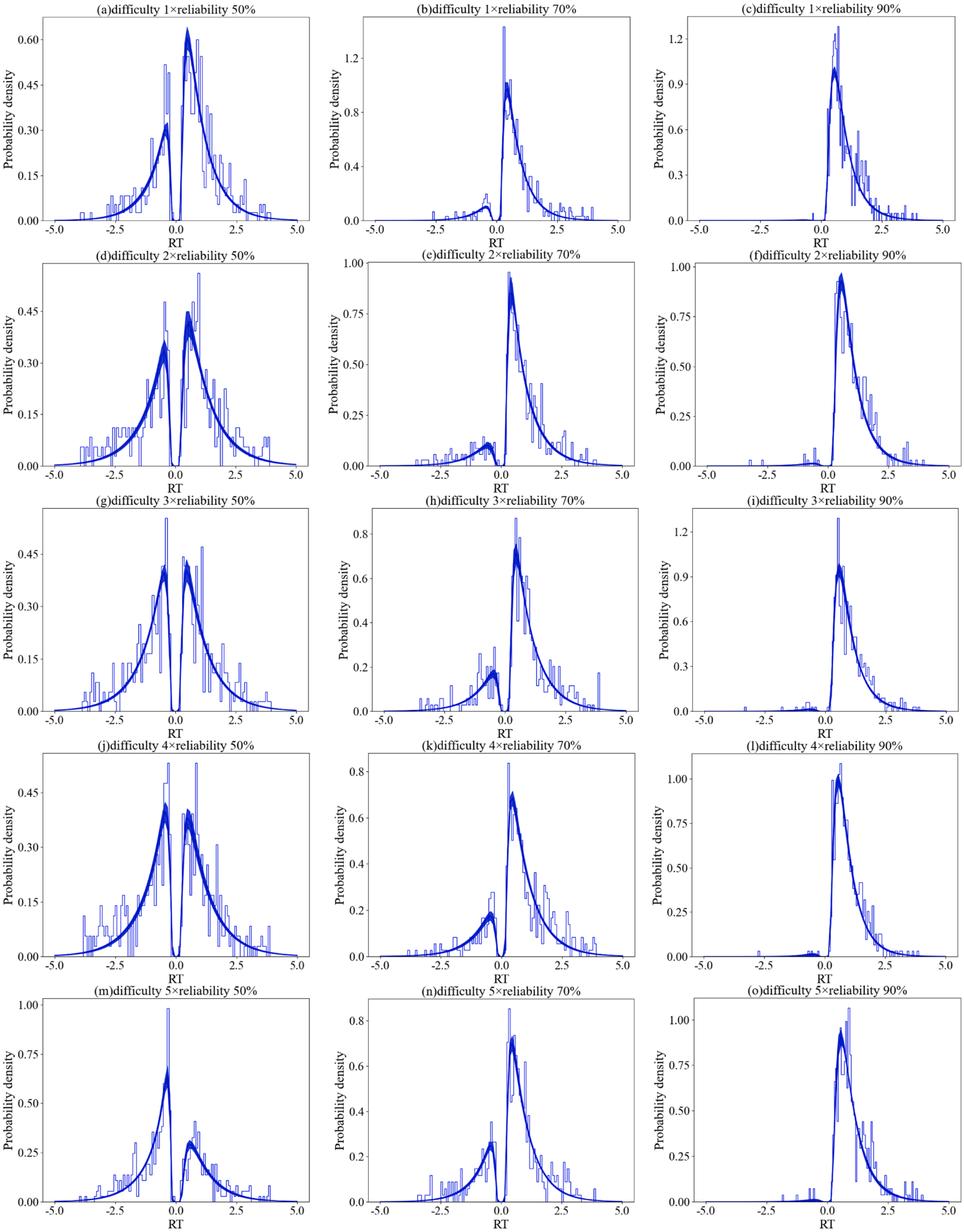


Fig. 14. Probability density of trust decision time across task scenarios fitted by the DDM.

first conducted a quantitative assessment of the time-varying uncertainty in trust and distrust decisions within typical human-ADS cooperative driving scenarios. As shown in Fig. 15, for a scenario with task difficulty level 3 and ADS reliability of 0.7, the model's dynamic predictions of decision uncertainty closely matched the cumulative frequencies observed in the experiment. This result demonstrates that the proposed model effectively captures the temporal evolution of decision uncertainty in realistic driving environments. Fig. 16

Furthermore, in this scenario, when $t > 5$ s, the decision probabilities converged to asymptotic values of $P_{trust}(t \to \infty) = 0.791$ and $P_{distrust}(t \to \infty) = 0.209$. Based on these outcomes, and using the predefined criteria for over-trust and under-trust events (Eq. (9)), we quantified the corresponding risk probabilities. The results, presented in Fig. 17, indicate convergence probabilities of 0.237 for over-trust events and 0.146 for under-trust events.

4.5. *Influence of task difficulty and ADS reliability on risk uncertainty of over-trust and under-trust*

To further examine the influence of different factors on the uncertainty of over-trust and under-trust risks, we generated heatmaps illustrating their time-varying evolution across 15 task scenarios (Fig. 17 and Fig. 18) and conducted both longitudinal and cross-sectional comparisons.

Longitudinal analysis shows that when ADS reliability is high (0.9), the probability of over-trust, $F_{trust}(t)$, remains largely unchanged with increasing task difficulty. In contrast, under medium and low reliability conditions (0.7 and 0.5), $F_{trust}(t)$ decreases as task difficulty rises. This can be explained by driver behavior: under high reliability, drivers exhibit strong reliance on ADS while the system itself rarely fails, maintaining a consistently high over-trust risk. Under lower reliability, however, drivers become more cautious as tasks grow more complex, increasingly opting for distrust decisions, thereby reducing over-trust risk uncertainty. For under-trust, $F_{distrust}(t)$ remains consistently low and relatively insensitive to task difficulty under high reliability (0.9). Under medium and low reliability, $F_{distrust}(t)$ also declines with task difficulty, reflecting drivers' adjustment of expectations and reduced reliance on limited-performance systems as task demands escalate.

Cross-sectional analysis reveals that under low task difficulty, $F_{trust}(t)$ increases as ADS reliability decreases. Yet, this trend weakens with higher task difficulty and even reverses at difficulty level 5, where $F_{trust}(t)$ first increases and then decreases. This suggests that in simple tasks, drivers readily develop reliance on automation, so declining reliability directly amplifies misjudgments and over-trust risk uncertainty. In more demanding tasks, however, drivers' vigilance and cognitive engagement increase, prompting greater awareness of system limitations and reliance on manual control, which suppresses or even reverses over-trust risk uncertainty. Conversely, $U_{distrust}(t)$ rises systematically as ADS reliability decreases, indicating that unstable performance and elevated error rates increase driver doubt, thereby elevating under-trust risk uncertainty.

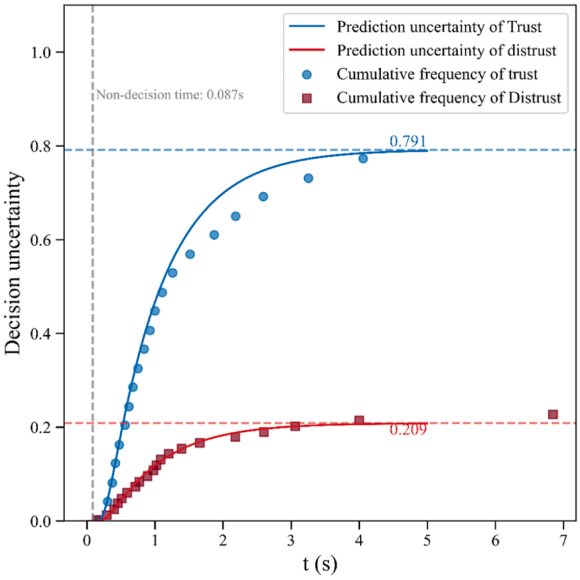


Fig. 15. Assessment of driver's decision-making uncertainty under task difficulty=3 and ADS reliability=0.7.

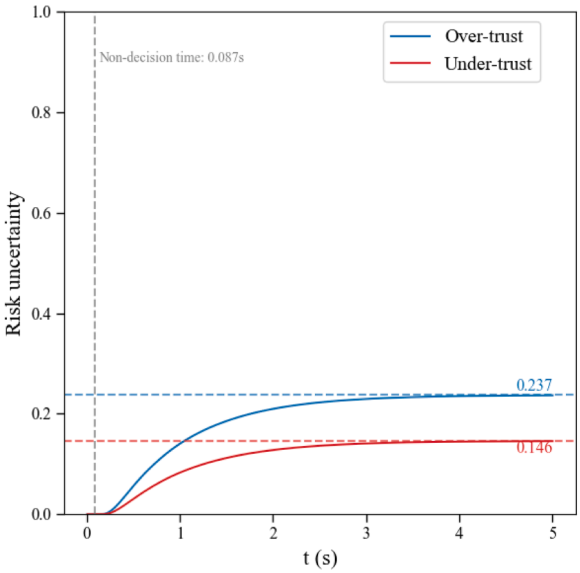


Fig. 16. Assessment of risk uncertainty of over-trust and under-trust under task difficulty=3 and ADS reliability=0.7.

5. Discussion

In this study, we propose a method for evaluating the risk uncertainty of over-trust and under-trust in human-ADS cooperative driving. The model was validated through human-in-the-loop experiments conducted in typical task scenarios, incorporating two variables: ADS reliability and task difficulty. The model demonstrated strong fitting performance and revealed several interesting phenomena. We will discuss the findings from the following aspects.

5.1. *Advantages of the proposed risk uncertainty assessment framework*

Compared with existing research [8,54,55], the proposed risk uncertainty assessment framework offers several notable advantages.

First, the model introduces a novel paradigm for evaluating risk uncertainty in human-ADS cooperative driving. Specifically, it assesses the risk uncertainty of over-trust and under-trust by jointly considering the variability in drivers' trust decisions and whether the ADS successfully completes the task. This approach more accurately reflects real-world conditions and aligns with the conceptual foundations established in our previous studies [31,32].

Second, traditional risk uncertainty assessment models are often static [8,55] or retrospective [54]. In contrast, the proposed model enables time-varying evaluation of risk uncertainty. This temporal feature makes it possible to detect risks in real time and implement timely interventions (e.g., system prompts or authority takeover), thereby enhancing the safety of human-ADS cooperation.

Third, the model is built on the DDM, which provides stronger explanatory power. A key advantage of the DDM lies in its ability to decompose the cognitive process underlying drivers' trust decisions into

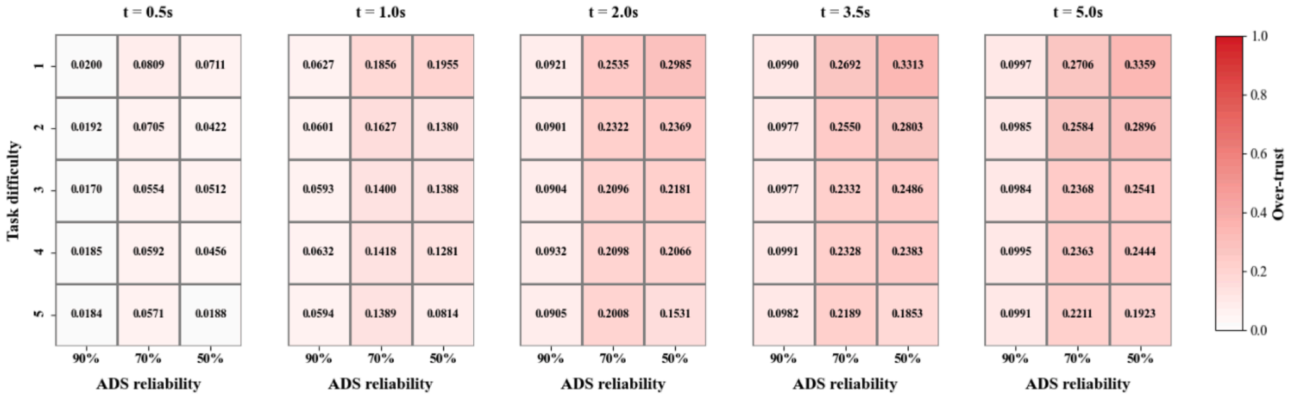


Fig. 17. Assessment of over-trust uncertainty over time across 15(5 × 3) task scenarios.

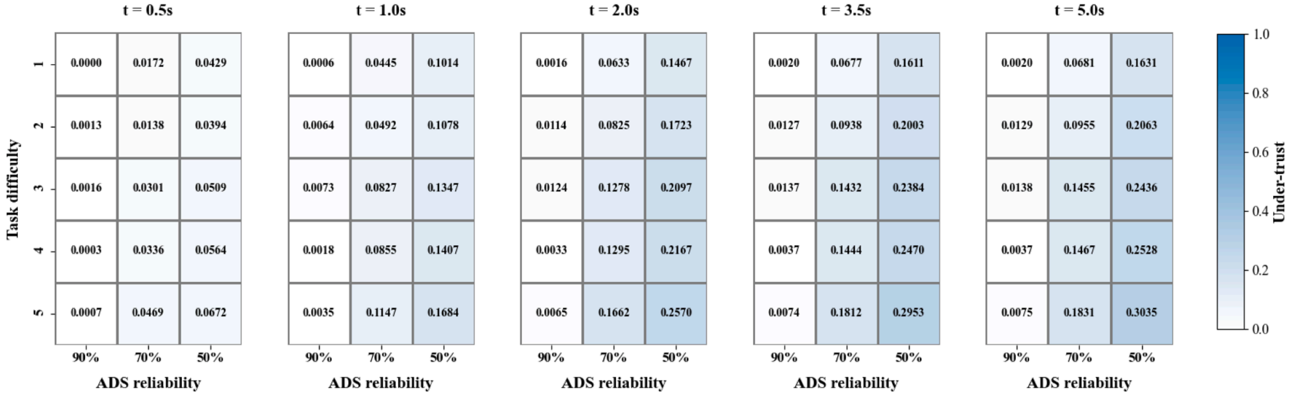


Fig. 18. Assessment of under-trust uncertainty over time across 15(5 × 3) task scenarios.

a set of interpretable parameters, thereby deepening our understanding of the cognitive mechanisms of trust in human-ADS interaction.

Finally, it is important to clarify that the “real-time” predictive capability of our framework refers to its ability to dynamically estimate the evolving probability of trust/distrust decisions within a given driving scenario based on pre-calibrated DDM parameters. This is distinct from predicting the behavior of an individual driver in a new context, which would require personalized model fitting. Our current work establishes the essential cognitive and statistical foundation for the former, which is a necessary precursor to the latter.

5.2. Model validation and predictive performance

This study established a hierarchical Bayesian DDM model based the behavioral data. All models passed convergence checks ($\hat{R} < 1.1$), and parameter estimation was stable. The results of PPCs (Fig. 14) showed that the model could accurately reproduce the distribution of decision times under different conditions. More importantly, the time-varying decision uncertainty generated by the model closely matched the observed cumulative frequencies (Fig. 15), demonstrating its effectiveness and accuracy in predicting trust decision uncertainty.

These results indicate that the proposed modeling framework has good generalizability and provides a reliable basis for the time-varying assessment of the trust risk uncertainty.

5.3. Interpretation of observed phenomena

The DDM modeling revealed the interactive influence of ADS reliability and task difficulty on drivers’ trust decisions.

When ADS reliability was low (50 % and 70 %), the drift rate v significantly decreased as task difficulty increased (Fig. 12). This suggests that under low-reliability system support, participants’ efficiency of information accumulation declined in more difficult tasks, making them more likely to make distrust decisions. This cognitive mechanism explains the non-monotonic pattern of decision times with difficulty observed in Fig. 11: under low reliability, high task difficulty led to more cautious decision strategies and thus longer decision times.

In contrast, under high reliability (90 %), the v parameter was not sensitive to changes in task difficulty, suggesting that participants had developed reliance on the system. Even when task difficulty increased, they maintained high efficiency of information accumulation. Although this cognitive dependence improved decision efficiency (Fig. 10), it may conceal potential risks.

Further analysis of time-varying risk uncertainty assessment (Fig. 17 and Fig. 18) revealed more complex mechanisms of risk formation. Longitudinal comparisons showed that under low reliability, over-trust risk uncertainty decreased as task difficulty increased, because drivers adjusted their reliance strategy and increased autonomous judgment when system performance was limited. Similarly, under low to medium reliability, under-trust risk uncertainty decreased with increasing difficulty, reflecting drivers’ adjustment of expectations toward limited system performance.

Cross-sectional comparisons further showed that in low-difficulty tasks, over-trust risk uncertainty increased as reliability decreased, indicating that drivers tended to form a false sense of security in simple tasks. In high-difficulty tasks (e.g., difficulty level 5), however, over-trust risk uncertainty first increased and then decreased as reliability decreased. This reflects drivers’ higher cognitive involvement under high workload, which promoted active recognition of system limitations and risk avoidance.

These findings collectively indicate that drivers adopt different cognitive strategies under different combinations of reliability and task difficulty: they tend to actively monitor and adjust under low reliability, form cognitive dependence under high reliability, are more prone to automation bias in low-difficulty tasks, and activate more active cognitive intervention in high-difficulty tasks.

5.4. Limitations and future research

While this study provided valuable insights, several limitations must be acknowledged. First, the ecological validity of the experimental paradigm is limited. The experiment was conducted in a controlled laboratory environment, where driving scenarios and ADS behaviors were necessarily simplified relative to the complexity of real-world conditions. Moreover, the design required participants to make trust judgments after viewing pre-recorded driving sequences, a post-hoc evaluation that differs from the dynamic, continuous monitoring and intervention required in Level 3 (L3) automated driving, where drivers must remain "in-the-loop" and respond to evolving situations in real time. This simplification was a deliberate methodological trade-off aligned with the study's core objective: isolating and quantifying the fundamental cognitive processes underlying trust decisions. By employing standardized stimuli and structured response intervals, we ensured consistent sensory input and system performance history across participants. Such experimental control is essential for reducing confounding variables and enabling the development of a clear and interpretable computational model.

Additionally, the model did not explicitly incorporate situational awareness (SA), a critical prerequisite for decision-making in dynamic environments. While the DDM effectively captures evidence accumulation and decision thresholds, it does not account for the preceding stages of SA (perception of environmental elements, comprehension of their meaning, and projection of future states). Finally, the proposed model operates at the population level and does not systematically examine how individual differences (e.g., driving experience, cognitive style, age) might influence trust decisions.

Future research should address these limitations by prioritizing three key directions. First, ecologically valid experimental paradigms should be developed using real-time driving simulators or on-road studies to better capture the dynamic interactions between drivers and ADS. Second, future work should explore integrating SA into the current framework. For example, a multi-stage modeling approach could link SA acquisition processes with subsequent trust decisions, building on our prior research [56]. Third, future work should integrate individual-difference factors, such as driving experience, cognitive style, and age, into the modeling framework to capture heterogeneous trust decision patterns.

6. Conclusion

We proposed a computational framework that quantifies over-trust and under-trust risk uncertainty in human-ADS cooperation. By integrating DDM with ADS reliability and task difficulty, the framework shifts risk uncertainty assessment from static evaluation to time-varying quantification.

Using 7200 behavioral observations data, we show that ADS reliability accelerates trust decisions while task difficulty produces non-monotonic effects. The hierarchical Bayesian DDM demonstrated excellent predictive validity, closely replicating observed behavior across all conditions. Beyond predictive accuracy, the model reveals how risk evolves in real time, exposing the mechanisms by which reliability and task complexity jointly shape drivers' trust.

These findings establish a quantitative foundation for real-time driver state monitoring and adaptive human-machine interaction. More broadly, the framework advances the research on trust in automation, providing a path toward safer, more intelligent, and more resilient human-AI collaboration.

CRediT authorship contribution statement

Song Ding: Writing – original draft, Methodology, Formal analysis, Conceptualization. **Lunhu Hu:** Methodology, Funding acquisition, Conceptualization. **Xing Pan:** Supervision, Resources, Project administration, Funding acquisition. **Jiacheng Liu:** Writing – review & editing, Software, Data curation. **Fu Guo:** Supervision, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This work is supported by the National Natural Science Foundation of China under Grants No.72071011/No.72501015/No.72561024, China Postdoctoral Science Foundation under Grant Number GZC20252705.

Appendix A. Parameters of the DDM for trust decision time

Reliability	90 %				70 %				50 %			
Difficulty												
1	a	v	T_{non}	z_{obs}	a	v	T_{non}	z_{obs}	a	v	T_{non}	z_{obs}
	3.450	1.443	0.069	0.609	2.123	0.998	0.089	0.520	2.006	0.413	0.104	0.475
2	a	v	T_{non}	z_{obs}	a	v	T_{non}	z_{obs}	a	v	T_{non}	z_{obs}
	2.900	1.342	0.069	0.562	2.198	0.703	0.101	0.542	2.210	0.200	0.098	0.475
3	a	v	T_{non}	z_{obs}	a	v	T_{non}	z_{obs}	a	v	T_{non}	z_{obs}
	2.851	1.379	0.090	0.528	2.143	0.595	0.087	0.513	2.078	0.028	0.096	0.496
4	a	v	T_{non}	z_{obs}	a	v	T_{non}	z_{obs}	a	v	T_{non}	z_{obs}
	3.185	1.547	0.072	0.556	2.142	0.597	0.066	0.510	2.115	0.020	0.083	0.481
5	a	v	T_{non}	z_{obs}	a	v	T_{non}	z_{obs}	a	v	T_{non}	z_{obs}
	3.120	1.400	0.047	0.546	1.984	0.520	0.099	0.501	2.200	−0.076	0.121	0.429

Data availability

Data will be made available on request.

References

[1] de Zwart R, Kamphuis K, Cleij D. Driver behavioural adaptations to simulated automated vehicles, potential implications for traffic microsimulation[J]. Transp Res F Traffic Psychol Behav 2023;92:255–65.

[2] Wang S, Li Z, Wang Y, et al. Evidence of automated vehicle safety's influence on people's acceptance of the automated driving technology[J]. Accid Anal Prev 2024;195:107381.

[3] Papadoulis A, Quddus M, Imprialou M. Evaluating the safety impact of connected and autonomous vehicles on motorways[J]. Accid Anal Prev 2019;124:12–22.

[4] Liao W, Qiao Y, Dong T, et al. A human reliability analysis method based on STPA-IDAC and BN-SLIM for driver take-over in level 3 automated driving[J]. Reliab Eng Syst Saf 2025;254:110577.

[5] Huang C, Yang B, Nakano K. Impact of duration of monitoring before takeover request on takeover time with insights into eye tracking data[J]. Accid Anal Prev 2023;185:107018.

[6] Ding S, Pan X, Hu L, et al. A new model for calculating human trust behavior during human-AI collaboration in multiple decision-making tasks: a bayesian approach[J]. Comput Ind Eng 2025;200(000).

[7] Cheng T, Wu B, U I B V A. Analysis of human errors in human-autonomy collaboration in autonomous ships operations through shore control experimental data[J]. Reliab Eng Syst Saf 2024;246:1 (Jun.).

[8] Zheng X, Liu Q, Li Y, et al. Safety risk assessment for connected and automated vehicles: integrating FTA and CM-improved AHP[J]. Reliab Eng Syst Saf 2025;257:110822.

[9] Liu P, Zhang Y, He Z. The effect of population age on the acceptable safety of self-driving vehicles[J]. Reliab Eng Syst Saf 2019;185:341–7.

[10] Liu Q, Li Y, Qin W, et al. Quantitative risk assessment for connected automated vehicles: integrating improved STPA-SafeSec and Bayesian network[J]. Reliab Eng Syst Saf 2025:253 (Jan.).

[11] Sohrabi S, Lord D, Dadashova B, et al. Assessing the collective safety of automated vehicle groups: a duration modeling approach of accumulated distances between crashes[J]. Accid anal prev 2024;198:107454.

[12] Qiu S, Rachedi N, Sallak M, et al. A quantitative model for the risk evaluation of driver-ADAS systems under uncertainty[J]. Reliab Eng Syst Saf 2017:167.

[13] Xia Y, Zhang Y, Xie J. If it cannot Be seen, can it Be quantified? Explicit and implicit risks in AV–HV mixed traffic[J]. Reliab Eng Syst Saf 2025:111717.

[14] Zhou S, Sun X, Liu B, et al. Factors affecting pedestrians' Trust in automated vehicles: literature review and theoretical model[J]. IEEE Trans Hum Mach Syst 2022;52(3):490–500.

[15] Lee JD, See KA. Trust in automation: designing for appropriate reliance.[J]. Hum Factors 2004;46(1):50.

[16] Gulati S, McDonagh J, Sousa S, et al. Trust models and theories in human–computer interaction: a systematic literature review[J]. Comput Hum Behav Rep 2024;16:100495.

[17] De Visser EJ, Monfort SS, Mckendrick R, et al. Almost Human: anthropomorphism increases trust resilience in cognitive agents[J]. J Exp Psychol Appl 2016;22(3):331.

[18] Hu WL, Akash K, Reid T, et al. Computational modeling of the dynamics of Human trust during Human–Machine interactions[J]. IEEE 2019;(6).

[19] Ayoub J, Avetisian L, Yang XJ, et al. Real-time trust prediction in conditionally automated driving using physiological measures[J]. IEEE Trans Intell Transp Syst 2023;(12):24.

[20] Hu C, Huang S, Zhou Y, et al. Dynamic and quantitative trust modeling and real-time estimation in human-machine co-driving process[J]. Transp Res F: Traffic Psychol Behav 2024;106:306–27.

[21] Lee M, Pitts BJ. The effects of automated vehicle reliability, self-estimated confidence, and repeated interactions on drivers' trust and takeover decisions[J]. Saf Sci 2026;195:107062.

[22] Merriman SE, Plant KL, Revell KMA, et al. What can we learn from automated vehicle collisions? A deductive thematic analysis of five automated vehicle collisions[J]. Saf Sci 2021;141:105320.

[23] Zhou S, Sun X, Wang Q, et al. Examining pedestrians' trust in automated vehicles based on attributes of trust: a qualitative study[J]. Appl Erg 2023.

[24] Zio E. The future of risk assessment[J]. Reliab Eng Syst Saf 2018:S0951832017306543.

[25] Aven T. On how to define, understand and describe risk[J]. Reliab Eng Syst Saf 2010;95(6):623–31.

[26] Vaddi PK, Diao X, Zhao Y, et al. Dynamic probabilistic risk assessment and game theory for cyber security risk analysis in nuclear power plants[J]. Reliab Eng Syst Saf 2025:111702.

[27] Teymoori S, Ghazaan MI, Malekitabar H. Developing a probabilistic risk assessment framework for construction projects based on Dynamic Bayesian Network[J]. Reliab Eng Syst Saf 2025:111897.

[28] Hu L, Kang R, Pan X, et al. Risk assessment of uncertain random system—Level-1 and level-2 joint propagation of uncertainty and probability in fault tree analysis [J]. Reliab Eng Syst Saf 2020;198:106874.

[29] Maidana RG, Parhizkar T, Gomola A, et al. Supervised dynamic probabilistic risk assessment: review and comparison of methods[J]. Reliab Eng Syst Saf 2023;230:108889.

[30] B NWA, B YGA, A CYL, et al. Integrated agent-based simulation and evacuation risk-assessment model for underground building fire: a case study[J]. J Build Eng 2021.

[31] Hu L, Pan X, Ding S, et al. A quantitative input for evaluating human error of visual neglection: prediction of operator's detection time spent on perceiving critical visual signal[J]. Reliab Eng Syst Saf 2022;225:108582.

[32] Hu L, Pan X, Kang R, et al. Dynamic risk assessment of Uncertain Random System considering operator's simple emergency-stop action in short time window[J]. Reliab Eng Syst Saf 2024:252.

[33] Levine CS, Al-Douri A, Paglioni VP, et al. Identifying human failure events for human reliability analysis: a review of gaps and research opportunities[J]. Reliab Eng Syst Saf 2024;245:109967.

[34] Patriarca R, Ramos M, Paltrinieri N, et al. Human reliability analysis: exploring the intellectual structure of a research field[J]. Reliab Eng Syst Saf 2020;203:107102.

[35] Zhang T, Yang J, Chen M, et al. EEG-based assessment of driver trust in automated vehicles[J]. Expert Syst Appl 2024;246:123196.

[36] Choo S, Nam CS. Detecting human trust calibration in automation: a convolutional neural network approach[J]. IEEE Trans Hum Mach Syst 2022;52(4):774–83.

[37] Akash K, Hu W, Reid T, et al. Dynamic modeling of trust in human-machine interactions. 2017 Am Control Conf (ACC)[C] 2017.

[38] Rabby MKM, Khan MA, Karimoddini A, et al. Modeling of trust within a human-robot collaboration framework. 2020 IEEE Int Conf Syst Man Cybern (SMC)[C] 2020.

[39] Seo M, Kia SS. Bayesian online learning for human-assisted target localization[J]. Automatica 2025;177:112288.

[40] Ratcliff R. A theory of memory retrieval.[J]. Psychol Rev 1978;85(2):59.

[41] Richter T, Ulrich R, Janczyk M. Diffusion models with time-dependent parameters: an analysis of computational effort and accuracy of different numerical methods [J]. J Math Psychol 2023;114:102756.

[42] Ren M, Chen N, Qiu H. Human-machine collaborative decision-making: an evolutionary roadmap based on cognitive intelligence[J]. Int J Soc Robot 2023;15 (7):1101–14.

[43] Donkin C, Brown SD. Response times and decision-making[J]. Stevens 19 Handb Exp Psychol Cogn Neurosci 2018;5:349–77.

[44] Forstmann BU, Ratcliff R, Wagenmakers E. Sequential sampling models in cognitive neuroscience: advantages, applications, and extensions[J]. Annu Rev Psychol 2016;67(1):641–66.

[45] Frank T V W M. HDDM Sequential Sampling Models[EB/OL].

[46] Feller W, Morse PM. An introduction to probability theory and its applications[Z]. American Institute of Physics; 1958.

[47] Navarro DJ, Fuss IG. Fast and accurate calculations for first-passage times in Wiener diffusion models[J]. J Math Psychol 2009;53(4):222–30.

[48] Kruschke JK. Bayesian data analysis[J]. Wiley Interdiscip Rev: Cogn Sci 2010;1(5):658–76.

[49] Dosovitskiy A, Ros G, Codevilla F, et al. CARLA: An open urban driving simulator: Conference on robot learning[C], 2017. PMLR.

[50] Hoff KA, Bashir M. Trust in automation: integrating empirical evidence on factors that influence trust[J]. Hum Factors 2015;57(3):407–34.

[51] Huang P, Ding H, Chen H. An entropy-based model for quantifying multi-dimensional traffic scenario complexity[J]. IET intell transp syst 2024;(7):18.

[52] Yu R, Zheng Y, Qu X. Dynamic driving environment complexity quantification method and its verification[J]. Transp Res C Emerg Technol 2021;(1):103051.

[53] Krause F, Lindemann O. Expyriment: a python library for cognitive and neuroscientific experiments[J]. Behav Res Methods 2014;46(2):416–28.

[54] Lee K, Lim JY. Development of an enhanced risk assessment model for Human–Robot collaboration and its application[J]. Saf Health Work 2024.

[55] Ding S, Pan X, Zuo D, et al. Uncertainty analysis of accident causality model using Credal Network with IDM method: a case study of hazardous material road transportation accidents[J]. Process Saf Environ Prot 2022;158:461–73.

[56] Ding S, Hu L, Pan X, et al. Assessing human situation awareness reliability considering fatigue and mood using EEG data: a bayesian neural network-bayesian network approach[J]. Reliab Eng Syst Saf 2025;260(000).