Engineering 33 (2024) 77–89

Contents lists available at ScienceDirect

Engineering

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

Research
Safety for Intelligent and Connected Vehicles—Article

# Toward Trustworthy Decision-Making for Autonomous Vehicles: A Robust Reinforcement Learning Approach with Safety Guarantees

Xiangkun He, Wenhui Huang, Chen Lv *

*School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore 639798, Singapore*

ARTICLE INFO

*Article history:*
Received 15 October 2022
Revised 22 March 2023
Accepted 18 October 2023
Available online 27 November 2023

*Keywords:*
Autonomous vehicle
Decision-making
Reinforcement learning
Adversarial attack
Safety guarantee

ABSTRACT

While autonomous vehicles are vital components of intelligent transportation systems, ensuring the trustworthiness of decision-making remains a substantial challenge in realizing autonomous driving. Therefore, we present a novel robust reinforcement learning approach with safety guarantees to attain trustworthy decision-making for autonomous vehicles. The proposed technique ensures decision trustworthiness in terms of policy robustness and collision safety. Specifically, an adversary model is learned online to simulate the worst-case uncertainty by approximating the optimal adversarial perturbations on the observed states and environmental dynamics. In addition, an adversarial robust actor-critic algorithm is developed to enable the agent to learn robust policies against perturbations in observations and dynamics. Moreover, we devise a safety mask to guarantee the collision safety of the autonomous driving agent during both the training and testing processes using an interpretable knowledge model known as the Responsibility-Sensitive Safety Model. Finally, the proposed approach is evaluated through both simulations and experiments. These results indicate that the autonomous driving agent can make trustworthy decisions and drastically reduce the number of collisions through robust safety policies.



## 1. Introduction

In recent years, autonomous vehicles have gained momentum with the rapid development of emerging technologies such as advanced mobile communication [1] and artificial intelligence (AI) [2], and are expected to revolutionize human mobility and transportation systems [3–5]. However, real-world traffic scenarios involve unpredictable noise or uncertainties, making it challenging to ensure the robustness and safety of driving policies. Hence, the trustworthiness of autonomous driving raises major concerns for various institutions and the general public [6–8]. Given these intricate challenges, meeting the rigorous requirements and high expectations pertaining to autonomous driving remains a significant concern [9–11].

The decision-making system can be likened to the brain of an autonomous vehicle, primarily responsible for determining the optimal driving mode or policy based on perception information [12–14]. Numerous studies have reported advances in decision-making methods for autonomous driving [15–17]. The finite-state machine (FSM), a rule-based technique, is the most popular approach for developing decision-making systems [18,19]. Although such a scheme is simple to implement and interpret, it relies heavily on the prior knowledge of specialists, thus making it difficult to design driving rules for complex traffic scenarios.

As a vital component of modern AI technologies, reinforcement learning (RL) provides a feasible and effective paradigm for solving complex sequential decision-making tasks via interactions with an environment [20–22]. Consequently, several studies have attempted various RL methods to address the sequence of autonomous driving tasks [23–25]. Researchers have leveraged RL algorithms to learn lane-change policies for autonomous driving [26,27]. For instance, a lane-change decision-making framework for autonomous vehicles was developed using a risk-awareness-prioritized replay deep Q-network (RA-PRDQN) method [28]. A safe lane-change decision scheme for autonomous driving was developed using an RL approach with a rule-based safety verification [29]. Some studies have employed RL algorithms to learn optimal target speeds or speed patterns (e.g., acceleration, deceleration, and maintenance) of autonomous vehicles [30,31]. For example, a cooperation-aware on-ramp merging decision-making scheme for autonomous vehicles was developed using

* Corresponding author.
*E-mail address:* lyuchen@ntu.edu.sg (C. Lv).

https://doi.org/10.1016/j.eng.2023.10.005

the belief-state RL method [32]. The subgoal-based speed patterns of autonomous vehicles were determined using a state-attention-model-based hierarchical RL approach [33]. To ensure the robustness of the on-ramp merging policies against environmental uncertainties, a robust decision-making solution for autonomous driving was proposed using a constrained adversarial RL technique [34]. Many researchers have leveraged RL algorithms to simultaneously learn optimal lane-change policies and speed patterns of autonomous vehicles [35–37]. For instance, longitudinal and lateral decision-making behaviors for autonomous driving can be learned via a double deep Q-network (DDQN) with a short-horizon safety checker [38], while target speeds and lane-change policies of autonomous vehicles can be determined using a hierarchical program-triggered RL technique based on multiple agents [39]. In another study, a trustworthy improvement RL scheme with a rule-based policy was developed to enable an autonomous driving agent to learn safe longitudinal and lateral driving velocities [40].

Although existing research on driving decisions has achieved numerous compelling results that can enhance the performance of autonomous vehicles, there is still room for improvement and perfection in terms of trustworthiness. Moreover, most studies assume that traffic scenarios are devoid of environmental uncertainty or involve only one specified type of uncertainty. Unfortunately, real-world scenarios involve substantial and inevitable uncertainties that can cause autonomous driving agents to make undesired or even unsafe decisions. In real-world traffic scenarios, multiple sources of uncertainty, such as observational noise and environmental changes, may coexist, leading to complex and challenging driving situations. Hence, policy robustness against multiple uncertainties should be considered in the autonomous driving domain. However, few studies have addressed the challenge of guaranteeing the safety of RL-based autonomous driving agents during training and testing in stochastic dynamic traffic flows with adversarial environmental uncertainties.

Consequently, all the above insights motivated us to explore a new technique to ensure the trustworthiness of autonomous driving decisions, including policy robustness and collision safety. In this study, we introduce a novel robust RL approach with safety guarantees (RRL-SG) aimed at achieving trustworthy decision-making for autonomous vehicles. The main contributions of this study are summarized as follows:

(1) An adversarial agent is trained online to model the worst-case multiple uncertainties by approximating the optimal adversarial perturbations for both observed states and environmental dynamics. An adversarial robust actor (ARAC) algorithm is developed to enable the agent to learn robust policies against observational noises and environmental changes.

(2) Using an interpretable knowledge model proposed by Intel, Responsibility-Sensitive Safety (RSS) [41,42], a safety mask is developed to guarantee the collision safety of the autonomous driving agent during both the training and testing processes, which can transform the probability corresponding to an unsafe decision into zero (i.e., a safe action space is formed by shielding risky actions).

(3) Numerical simulation results with Simulation of Urban Mobility (SUMO) [43] indicate that the proposed RRL-SG approach guarantees the trustworthiness of autonomous vehicles in stochastic dynamic traffic flows with adversarial environmental perturbations. Experiments using a real autonomous vehicle further confirm the effectiveness of the proposed technique.

The remainder of this paper is organized as follows. Section 2 describes the proposed RRL-SG solution. Section 3 presents details of the technical implementation. Section 4 details the simulations and experiments, and analyzes the resulting performance. Finally, Section 5 concludes the study.

## 2. Methodology

### 2.1. Overview

In this section, we provide an overview of the proposed technique. Fig. 1 illustrates a block diagram of our RRL-SG framework designed to realize trustworthy decision-making for autonomous vehicles. $\Delta_o^*$ and $\Delta_d^*$ represent the optimal adversarial perturbations on observed states and environmental dynamics, respectively. $M_s$, $s$, $a$, $r$, and $\pi$ denote the safety mask, state, action, reward, and policy of the agent, respectively. $\pi_s$ represents a safe policy. $t$ is the time step and $T$ is the last time step. $\Delta$, $\gamma$, $\beta$, and $Q^\pi$ denote the environmental uncertainty, discount factor, weight, and action–value function in our optimization objectives, respectively.

The input of the adversary model is the state $s$ of the agent, and its output contains adversarial perturbations $\Delta_o^*$ and $\Delta_d^*$. $\Delta_o^*$ simulates the worst-case observational noise, which aims to maximize the average variation distance on perturbed policies. Moreover, $\Delta_d^*$ models the worst-case environmental dynamics uncertainty, which seeks to minimize the expected return of the agent.

The input to the RSS-based safety mask is state $s$ of the agent. A safety mask can create a safe action space by shielding it against risky actions. Hence, the autonomous driving agent interacts with the environment through actions sampled from safety policy $\pi_s$. The ARAC algorithm enables an agent to learn robust policies against perturbations in observations and dynamics.

Our autonomous driving agent was an intelligent vehicle colored gold, as shown in Fig. 1. Furthermore, the surrounding vehicles of other colors were controlled using an intelligent driving model (IDM) based on SUMO. The action space of our autonomous driving agent is discrete, encompassing five distinct decision-making behaviors: maintaining the current state, accelerating, decelerating, and changing lanes to either the left or the right.

### 2.2. Adversary model

The adversary model aims to generate optimal adversarial perturbations in the observed states and environmental dynamics.

To measure the variations in the policy caused by adversarial perturbations on observations, we leverage the Jensen–Shannon (JS) divergence, which can be considered a symmetrized and smoothed Kullback–Leibler (KL) divergence [44,45]. One of its key characteristics is that the JS divergence binds the distance between the two probability distributions to within 1.0. Thus, the objective function related to the perturbations in the observations, $J_o$, can be defined as follows:

$$\begin{aligned} J_o(s,\pi,\Delta_o) &= D_{JS}[\pi(a|s)\|\pi(\tilde{a}\,|\,\tilde{s})] \\ &= D_{JS}[\pi(a|s)\|\pi(\tilde{a}\,|s+\Delta_o)] \\ &= \frac{1}{2}D_{KL}[\pi(a|s)\|M] + \frac{1}{2}D_{KL}[\pi(\tilde{a}\,|s+\Delta_o)\|M] \end{aligned} \tag{1}$$

$$M = \frac{1}{2}[\pi(a|s) + \pi(\tilde{a}\,|s+\Delta_o)] \tag{2}$$

where $D_{JS}$ represents the distance based on the JS divergence, $D_{KL}$ denotes the distance based on the KL divergence, $\Delta_o$ represents the perturbation on observations, $M$ is an expression regarding the agent policy and perturbed policy, and $\tilde{s}$ and $\tilde{a}$ are the state and action perturbed by $\Delta_o$, respectively.

In this study, the adversarial perturbation of the dynamics attempted to minimize the expected return of the agent. We leverage an action–value function $Q^\pi(s)$ to estimate the expected return based on a pair of the state $s$ and the action $a$ when the agent

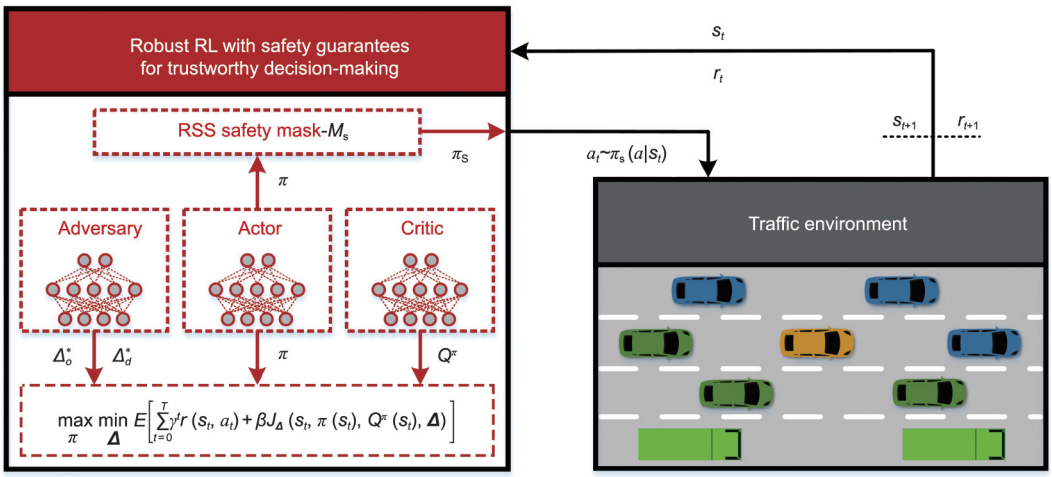


**Fig. 1.** Schematic of the proposed RRL-SG framework for trustworthy decision-making of autonomous vehicles. $\Delta_o^*$: the optimal adversarial perturbations on observed states; $\Delta_d^*$: the optimal adversarial perturbations on environmental dynamics; $M_s$: the safety mask of the agent; $s$: the state of the agent; $a$: the action of the agent; $r$: the reward of the agent; $\pi$: the policy of the agent; $\pi_s$: a safe policy; $t$: the time step; $T$: the last time step; $\Delta$: the environmental uncertainty; $\gamma$: discount factor; $\beta$: weight; $Q^\pi$: action–value function; $E$: mathematic expectation; $J_\Delta$: the objective function of the adversary.

follows the policy $\pi$. As the action space of our agent is discrete, the input to the action–value function $Q^\pi(s)$ does not include the action $a$. Hence, the objective function related to the perturbations on dynamics, $J_d$, can be designed as

$$J_d(s, Q^\pi, \Delta_d) = \Delta_d Q^\pi(s) \tag{3}$$

where $\Delta_d$ represents the perturbation of dynamics in the form of a probability distribution. Furthermore, the objective function of the adversary, $J_\Delta$, can be defined as

$$J_\Delta(s, \pi, Q^\pi, \Delta) = (\alpha - 1)J_o(s, \pi, \Delta_o) + \alpha J_d(s, Q^\pi, \Delta_d) \tag{4}$$

where $\alpha \in (0, 1)$ denotes a weight, $\Delta = [\Delta_o, \Delta_d]$ represents the environmental uncertainty.

The optimization problem with regard to the adversary model can be formulated as

$$\begin{aligned} &\Delta^* \in \arg\min_{\Delta} E[J_\Delta(s, \pi, Q^\pi, \Delta)], \\ &\text{subject to } |\Delta_o| \le \eta_1, |\Delta_d| \le \eta_2 \end{aligned} \tag{5}$$

where $\Delta^*$ represents the optimal environmental uncertainty, the notion of "argmin", which stands for argument of the minimum, and $\eta_1$ and $\eta_2$ denote the bounds of the perturbations on observations and dynamics, respectively. Hence, the adversarial agent aims to maximize $J_o$ and minimize $J_d$.

To simplify the aforementioned constrained optimization problem, we constrained the magnitude of the perturbations using the hyperbolic tangent and softmax functions. Specifically, the perturbations on observations and dynamics can be represented as $\Delta_o = \eta\tanh[x(s;\bar{\theta})]$, $\Delta_d = \text{softmax}[x(s;\bar{\theta})]$, respectively. In addition, $\eta$ represents the scale factor, $x$ denotes the output of the hidden layer of the adversary network, and $\bar{\theta}$ represents the adversary model parameter.

Consequently, to determine the optimal adversarial perturbation, Eq. (5) can be converted into

$$\bar{\theta}^* \in \arg\min_{\bar{\theta}} E[J_\Delta(s, \pi, Q^\pi; \bar{\theta})] \tag{6}$$

where $\bar{\theta}^*$ represents the parameters of the optimal adversary model. Clearly, the optimal adversarial perturbations on observations and dynamics can be expressed as $\Delta_o^* = \eta\tanh[x(s;\bar{\theta}^*)]$, $\Delta_d^* = \text{softmax}[x(s;\bar{\theta}^*)]$, respectively.

### 2.3. RSS-based safety mask

In this section, a safety mask is developed using an interpretable RSS model to guarantee the collision safety of autonomous vehicles.

To consider driving comfort, we leveraged the jerk-bounded RSS model [42] proposed by Intel to design a safety mask. This model describes the following braking processes: a vehicle starts decreasing its acceleration with a maximum jerk $j_{max}$ until it reaches a minimum deceleration $a_{min,r}$, and then the vehicle continues to brake with the deceleration $a_{min,r}$ until reaching a full stop. The jerk-bounded RSS model, $D_{min}^{RSS}$, yields the following expression for the minimum safe distance between front and rear vehicles:

$$D_{min}^{RSS} = \left| v_r \bar{T} + \frac{1}{2} a_r \bar{T}^2 - \frac{1}{6} j_{max} \bar{T}^3 + \frac{\left(v_r + a_r \bar{T} - \frac{1}{2} j_{max} \bar{T}^2\right)^2}{2|a_{min,r}|} - \frac{v_f^2}{2|a_{max,f}|} \right| \tag{7}$$

where $a_r$ is the initial acceleration of the rear vehicle; $v_f$ and $v_r$ denote the initial speeds of the front and rear vehicles; $a_{max,f}$ denotes the maximum deceleration of the front vehicle; $\bar{T}$ represents the time from the beginning until the rear vehicle's deceleration first equals $a_{min,r}$ or its speed decreases to zero.

We illustrate the proposed safety mask technique using the two cases shown in Fig. 2. As shown in Fig. 2(a), if the distance from the front vehicle in the same lane (denoted $D_f$) is less than or equal to $D_{min}^{RSS}$, the mask will transform the probability corresponding to the acceleration decision-making (denoted $a^4$) to zero (i.e., the safe action space including $a^1, a^2, a^3$, and $a^5$) is formed by shielding the risky action $a^4$. Although only the minimum longitudinal safe distance model is provided in Ref. [42], we can still employ this model to evaluate the lane-change risk if we assume that the vehicle can move laterally to the target lane instantaneously. Such an assessment is risky because the distance between the two vehicles may be further shortened during lane changing. Here, we designed a simple minimum lateral safety distance model based on $D_{min}^{RSS}$ as follows:

$$\bar{D}_{min}^{RSS} = \xi D_{min}^{RSS} \tag{8}$$

where $\xi$ represents a scale coefficient greater than 1.0, and $\bar{D}_{min}^{RSS}$ denotes the minimum lateral safety distance model.

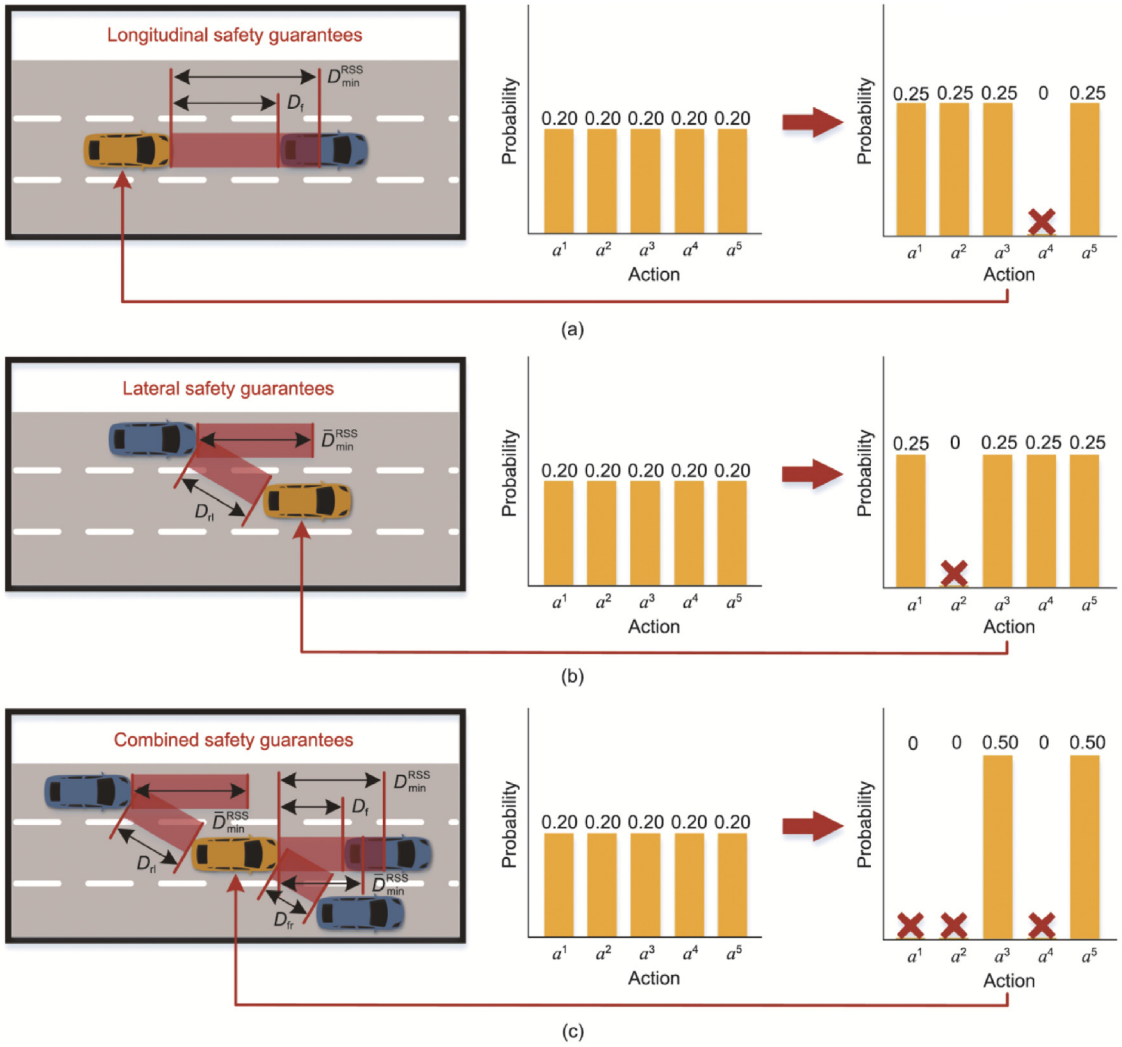


**Fig. 2.** Illustration of the RSS-based safety mask for trustworthy driving decisions. (a) Schematic diagram of longitudinal safety guarantees. (b) Schematic diagram of lateral safety guarantees. (c) Schematic diagram of combined safety guarantees. $a^1$, $a^2$, $a^3$, $a^4$, and $a^5$ represent changing lanes to the right, changing lanes to the left, keeping the current state, accelerating, and decelerating, respectively. $D_f$: the distance from the front vehicle in the same lane; $\bar{D}_{\min}^{RSS}$: the minimum lateral safety distance model; $D_{rl}$: the distance from the rear vehicle in the left lane; $D_{fr}$: the distance from the front vehicle in the right lane.

In Fig. 2(b), if the distance from the rear vehicle in the left lane (denoted $D_{rl}$) is less than or equal to $\bar{D}_{\min}^{RSS}$, the mask transforms the probability corresponding to the left lane-changing decision (denoted $a^2$) to zero.

In Fig. 2(c), when the distance from the rear vehicle in the left lane (denoted $D_{rl}$), distance from the front vehicle in the right lane (denoted $D_{fr}$), and distance from the front vehicle in the same lane (denoted $D_f$) are less than or equal to their corresponding minimum safety distances, the mask transforms the probability corresponding to the left lane-changing (denoted $a^2$), right lane-changing (denoted $a^1$), and accelerating (denoted $a^4$) decisions to zero.

Algorithm 1 provides an overview of the design of our RSS-based safety-mask module, where $D_r, D_{fl}, D_{fr}$, and $D_{rr}$ represent the distances from the rear, front-left, front-right, and rear-right vehicles, respectively; $D_{\min,f}^{RSS}, D_{\min,r}^{RSS}, \bar{D}_{\min,fl}^{RSS}, \bar{D}_{\min,rl}^{RSS}, \bar{D}_{\min,fr}^{RSS}$, and $\bar{D}_{\min,rr}^{RSS}$ denote the minimum safe distances from the front, rear, front-left, rear-left, front-right, and rear-right vehicles, respectively. Moreover, $M_s[m]$ denotes the $m$-th element in safety mask $M_s$. The mask element associated with the hazardous action is assigned a negative infinity value.

**Algorithm 1.** RSS-based safety mask.

**Input:** State of the autonomous driving agent
  Initialize a mask $M_s$ = [0, 0, 0, 0, 0]
  **if** $D_f \le D_{\min,f}^{RSS}$ **then**
    $M_s[4] = -\infty$ *Mask accelerating decision-making
  **end if**
  **if** $D_r \le D_{\min,r}^{RSS}$ **then**
    $M_s[5] = -\infty$ *Mask decelerating decision-making
  **end if**
  **if** $D_{fl} \le \bar{D}_{\min,fl}^{RSS}$ **or** $D_{rl} \le \bar{D}_{\min,rl}^{RSS}$ **then**
    $M_s[2] = -\infty$ *Mask left lane-changing decision-making
  **end if**
  **if** $D_{fl} \le \bar{D}_{\min,fr}^{RSS}$ **or** $D_{rr} \le \bar{D}_{\min,rl}^{RSS}$ **then**
    $M_s[1] = -\infty$ *Mask right lane-changing decision-making
  **end if**
**Output:** $M_s$

### 2.4. ARAC-critic

#### 2.4.1. Safe robust Markov decision process (MDP)

A MDP provides a mathematical paradigm for RL problems, aiming to find optimal policies [46]. In this section, the existing standard MDP mathematical formalism is extended to explicitly model the behavior of an autonomous driving agent under adversarial perturbation and a safety mask. Here, we introduce a safe robust MDP (SR-MDP) defined as follows:

A SR-MDP can be defined via a seven-tuple $[S, A, p, r, \Delta, M_s, \gamma]$ with state space $S$, action space $A$, state transition probability $p$, reward function $r$, safety mask $M_s$, environmental uncertainty $\Delta$, and discount factor $\gamma \in (0, 1)$.

In our study, SR-MDP attempts to solve the following problem:

$$\max_{\pi} \min_{\Delta} E\left[\sum_{t=0}^{T}\gamma^t r(s_t, a_t) + \beta J_{\Delta}(s_t, \pi(s_t), Q^{\pi}(s_t), \Delta)\right] \tag{9}$$

where $T$ is the last time step, and $\beta > 0$ is a trade-off coefficient.

We employ a novel policy iteration (PI) algorithm—known as the safe robust PI (SR-PI)—to solve the SR-MDP. The SR-PI method comprises two critical phases: safe–robust policy evaluation and robust policy improvement. Furthermore, both phases were updated iteratively until convergence was achieved.

#### 2.4.2. Safe robust policy evaluation

In the safe robust policy evaluation stage, we aim to estimate the expected return of the policy $\pi$ under environmental uncertainty $\Delta$. For a fixed policy, the action–value function $Q^{\pi}(\cdot)$ can be approximated iteratively by employing the following Bellman backup operator $\mathscr{T}^{\pi,\Delta}$:

$$\mathscr{T}^{\pi,\Delta}Q^{\pi}(s_t) = r(s_t, a_t) + \gamma E[V^{\pi,\Delta}(s_{t+1})] \tag{10}$$

where

$$V^{\pi,\Delta}(s_{t+1}) = \pi(s_{t+1})Q^{\pi}(s_{t+1}) + \beta J_{\Delta}(s_{t+1}, \pi(s_{t+1}), Q^{\pi}(s_{t+1}), \Delta) \tag{11}$$

denotes the value function of the agent based on $\pi$ under the adversarial perturbations.

Here, we can rewrite Eq. (10) as:

$$\mathscr{T}^{\pi,\Delta}Q^{\pi}(s_t) = r_a(s_t, a_t) + \gamma\pi(s_{t+1})Q^{\pi}(s_{t+1}) \tag{12}$$

where $r_a(s_t, a_t) = r(s_t, a_t) + \gamma\beta J_{\Delta}(\cdot)$ is the augmented reward. Hence, the convergence of our policy evaluation can be guaranteed by drawing upon findings related to policy evaluation convergence in standard RL algorithms.

To enhance the efficiency of model training, we employ two parameterized action–value functions with parameters $\phi^p$, $p \in \{1, 2\}$. The parameters of the two action–value functions can be optimized by minimizing the following objective function concerning the critic network:

$$J_Q(\phi^p) = \underset{T_s\sim\mathscr{B}}{E}[(y_t^{\Delta} - Q^{\pi}(s_t; \phi^p))^2] \tag{13}$$

where $T_s$ represents state transitions sampled from the replay buffer $\mathscr{B}$, and $y_t^{\Delta}$ denotes the target value of the action–value function with the uncertainty at the time step $t$, $J_Q$ is the function for optimizing the critic network. A smaller value is used for both action–value functions to mitigate the overestimation of the value function during the training of the critic network. As a result, $y_t^{\Delta}$ can be defined as:

$$\begin{aligned} y_t^{\Delta} = {} & r(s_t, a_t) + \gamma\pi(s_{t+1})\widehat{Q}^{\pi}_{\min}(s_{t+1}; \bar{\phi}^p) \\ & + \beta J_{\Delta}(s_{t+1}, \pi(s_{t+1}), \widehat{Q}^{\pi}_{\min}(s_{t+1}; \bar{\phi}^p), \Delta) \end{aligned} \tag{14}$$

where $\widehat{Q}^{\pi}(s; \bar{\phi}^p)$ is the target action–value function with the parameter $\bar{\phi}^p$, $\widehat{Q}^{\pi}_{\min}(s; \bar{\phi}^p)$ represents the smaller value of both the target action–value functions, for example, $\widehat{Q}^{\pi}_{\min}(s; \bar{\phi}^p) = \min_{p\in\{1,2\}} \widehat{Q}^{\pi}(s; \bar{\phi}^p)$.

Here, the gradient of Eq. (13) can be derived as:

$$\begin{aligned} \nabla_{\phi^p} J_Q(\phi^p) &= \nabla_{\phi^p} \underset{T_s\sim\mathscr{B}}{E}[((y_t^{\Delta} - Q^{\pi}(s_t; \phi^p))^2] \\ &= -2 \underset{T_s\sim\mathscr{B}}{E}[(y_t^{\Delta} - Q^{\pi}(s_t; \phi^p))\nabla_{\phi^p} Q^{\pi}(s_t; \phi^p)] \end{aligned} \tag{15}$$

Furthermore, we can update $\bar{\phi}^p$ via Polyak averaging:

$$\bar{\phi}^p \leftarrow \mu\bar{\phi}^p + (1 - \mu)\phi^p \tag{16}$$

where $\mu \in (0, 1)$ denotes a scale coefficient.

#### 2.4.3. Safe robust policy improvement

In the safe robust policy improvement stage, we attempt to optimize the policy given the action–value function $Q^{\pi}(\cdot)$ under the adversarial perturbations. Since the action–value function $Q^{\pi}(s)$ is employed to estimate the expected return based on a pair of the state $s$ and the action $a$ when the agent follows the policy $\pi$, the optimization problem Eq. (9) can be rewritten as:

$$\max_{\pi} \min_{\Delta} E[J(\pi, \Delta)] \tag{17}$$

where $J(\cdot)$ represents the objective function of the proposed SR-MDP, and $J(\pi, \Delta) = \pi(s)Q^{\pi}(s) + \beta J_{\Delta}(s, \pi, Q^{\pi}, \Delta)$.

Consequently, the optimal policy $\pi^*$ and optimal adversarial perturbation $\Delta^*$ for the observed states and environmental dynamics can be approximated using the following alternating procedure: Firstly, fix a policy $\pi$, then solve the optimal adversarial perturbation $\Delta^*$ through minimizing $J(\pi, \Delta)$. Secondly, with $\Delta^*$, learn the optimal policy $\pi^*$ through maximizing $J(\pi, \Delta^*)$. According to Eq. (17), the following relational expression is derived:

$$\Delta^* = \arg\min_{\Delta} E[J(\pi, \Delta)] \tag{18}$$

$$\pi^* = \arg\max_{\pi} E[J(\pi, \Delta^*)] \tag{19}$$

We observe that Eq. (17) represents a zero-sum game. In addition, the theoretical results [47–49] were established to guarantee the convergence of solutions for a zero-sum game, which can also ensure the convergence of our policy improvement.

To decrease the learning error of the policy $\pi$, we utilize the double $Q^{\pi}(\cdot)$ trick in Ref. [50]. Consequently, the policy model parameter $\theta$ can be learned by maximizing the following objective function concerning the actor network:

$$\begin{aligned} J_{\pi}(\theta) = {} & \underset{T_s\sim\mathscr{B}}{E}[\pi(s_t; \theta)Q^{\pi}_{\min}(s_t; \phi^p) \\ & + \beta J_{\Delta}(s_t, \pi(s_t; \theta), Q^{\pi}_{\min}(s; \phi^p), \Delta)] \end{aligned} \tag{20}$$

where $Q^{\pi}_{\min}(s; \phi^p)$ represents the smaller value of both the action–value functions, for example, $Q^{\pi}_{\min}(s; \phi^p) = \min_{p\in\{1,2\}} Q^{\pi}(s_t; \phi^p)$, $J_{\pi}$ is the function for optimizing the actor network.

We are able to derive the gradient of Eq. (20), as follows:

$$\begin{aligned} \nabla_{\theta} J_{\pi}(\theta) &= \nabla_{\theta} \underset{T_s\sim\mathscr{B}}{E}[\pi(s_t; \theta)Q^{\pi}_{\min}(s_t; \phi^p) \\ &\quad + \beta J_{\Delta}(s_t, \pi(s_t; \theta), Q^{\pi}_{\min}(s; \phi^p), \Delta)] \\ &= \underset{T_s\sim\mathscr{B}}{E}[\nabla_{\theta}\pi(s_t; \theta)Q^{\pi}_{\min}(s_t; \phi^p) + (\alpha - 1)\beta\nabla_{\theta} J_{\mathrm{o}}(s_t, \pi(s_t; \theta))] \\ &= \underset{T_s\sim\mathscr{B}}{E}[\nabla_{\theta}\pi(s_t; \theta)Q^{\pi}_{\min}(s_t; \phi^p) \\ &\quad + \frac{1}{2}(\alpha - 1)\beta(\nabla_{\theta} D_{\mathrm{KL}}(\pi(a|s; \theta)\|M(s; \theta)) \\ &\quad + \nabla_{\theta} D_{\mathrm{KL}}(\pi(\tilde{a}\,|s + \Delta_{\mathrm{o}}; \theta)\|M(s; \theta))] \end{aligned} \tag{21}$$

In addition, according to Eqs. (4) and (5), the adversary's model can be optimized by minimizing the following objective function:

$$\begin{aligned} J_{\bar{\pi}}(\bar{\theta}) &= \underset{T_s\sim\mathscr{B}}{E}[J_{\Delta}(s_t, \pi(s_t; \theta), Q^{\pi}_{\min}(s_t; \phi^p); \bar{\theta})] \\ &= \underset{T_s\sim\mathscr{B}}{E}[(\alpha - 1)J_{\mathrm{o}}(s_t, \pi(s_t); \bar{\theta}) + \alpha J_{\mathrm{d}}(s_t, Q^{\pi}_{\min}(s_t; \phi^p); \bar{\theta})] \end{aligned} \tag{22}$$

where $\bar{\theta}$ represents the adversary model parameter, $J_{\bar{\pi}}$ is the function for optimizing the adversary network.

Here, the gradient of Eq. (22) can be derived as:

$$\begin{aligned}\nabla_{\bar{\theta}} J_{\bar{\pi}}(\bar{\theta}) &= \nabla_{\bar{\theta}} \underset{T_s\sim\mathscr{B}}{E}[J_A(s_t, \pi(s_t;\theta), Q^{\pi}_{\min}(s_t;\phi^p);\bar{\theta})] \\ &= \nabla_{\bar{\theta}} \underset{T_s\sim\mathscr{B}}{E}[(\alpha-1)J_o(s_t,\pi(s_t);\bar{\theta}) + \alpha J_d(s_t, Q^{\pi}_{\min}(s_t;\phi^p);\bar{\theta})] \\ &= \nabla_{\bar{\theta}} \underset{T_s\sim\mathscr{B}}{E}[\tfrac{1}{2}(\alpha-1)(D_{KL}(\pi(a|s)\|M(s;\bar{\theta})) \\ &+ D_{KL}(\pi(\tilde{a}|s+\Delta_o(s;\bar{\theta}))\|M(s;\bar{\theta})) + \alpha\Delta_d(s;\bar{\theta})Q^{\pi}(s)] \\ &= \underset{T_s\sim\mathscr{B}}{E}[\tfrac{1}{2}(\alpha-1)(\nabla_{\bar{\theta}} D_{KL}(\pi(a|s)\|M(s;\bar{\theta})) \\ &+ \nabla_{\bar{\theta}} D_{KL}(\pi(\tilde{a}|s+\Delta_o(s;\bar{\theta}))\|M(s;\bar{\theta})) + \alpha\nabla_{\bar{\theta}}\Delta_d(s;\bar{\theta})Q^{\pi}(s)]\end{aligned} \tag{23}$$

## 3. Technical implementation

### 3.1. Algorithm

Here, we provide a detailed introduction to the implementation specifications of the proposed technique. Algorithm 2 outlines the RRL-SG approach for trustworthy autonomous driving decision-making. The initial model parameters for the actor, adversary, and critic were set using a random distribution. In terms of interaction with the environment, our agent interacts with the environment based on actions sampled from the safety policy $\pi_s$. In terms of policy learning, an agent policy can be optimized by combining Eqs. (13), (16), (20), and (22). $d_t$ represents a completed signal, implying that the ego vehicle encounters a collision at time step $t$. The details of the neural networks and hyperparameters are provided in Table S1 in Appendix A.

**Algorithm 2.** Robust RL with safety guarantees.

Initialize actor model parameters $\theta$, adversary model parameter $\bar{\theta}$, critic model parameters $\varphi^1$ and $\varphi^2$, target action–value function parameters $\bar{\varphi}^1 \leftarrow \varphi^1$ and $\bar{\varphi}^2 \leftarrow \varphi^2$, and an empty replay buffer $B$
**for** episode step $e$ = 1, 2,…, $E$ **do**
  Reset state $s_0$
  **for** time step in the environment $t$ = 1, 2,…, $T$ **do**
    Determine a safe policy $\pi_s(s_t;\theta)$ via Algorithm 1:
      $\pi_s(s_t;\theta)$ = softmax($\pi(s_t;\theta)$ + $M_s$)
    Select an action via the safe policy $\pi_s(s_t;\theta)$:
      $a_t \sim \pi_s(s_t;\theta)$
    Execute $a_t$ in the environment and receive a transition:
      $s_{t+1}$, $r_t$, $d_t \sim p(s_{t+1}|s_t, a_t)$
    Store the transition in the replay buffer $B$:
      $B \leftarrow B \cup \{(s_t, a_t, r_t, s_{t+1}, d_t)\}$
  **end if**
  **for** gradient step $g$ = 1, 2,…, $G$ **do**
    Sample a batch of transitions from the replay buffer $B$
    Update the actor model parameters via Eq. (21):
      $\theta \leftarrow \nabla_{\theta} J_{\pi}(\theta)$
    Update the critic model parameters via Eq. (15):
      $\varphi^1 \leftarrow \nabla_{\varphi 1} J_Q(\varphi^1)$, $\varphi^2 \leftarrow \nabla_{\varphi 2} J_Q(\varphi^2)$
    Update the target action–value function parameters via Eq. (16):
      $\bar{\varphi}^1 \leftarrow \mu\bar{\varphi}^1 + (1-\mu)\varphi^1$, $\bar{\varphi}^2 \leftarrow \mu\bar{\varphi}^2 + (1-\mu)\varphi^2$
    **if** $g$ mod $\delta$ **then**
      Update the adversary model parameters via Eq. (23):
        $\bar{\theta} \leftarrow \nabla_{\bar{\theta}} J_{\bar{\pi}}(\bar{\theta})$
    **end if**
  **end for**
**end for**

### 3.2. State space and action space

Designing the state, action, and reward functions of the autonomous driving agent was essential to implement the proposed scheme. In this study, we consider the relevant states of the six nearest social vehicles in the ego vehicle lane and adjacent lanes as observations for the autonomous driving agent (i.e., the ego vehicle). The state space of the autonomous driving agent has 15 dimensions, including the relative distance and velocity of the surrounding social vehicles and the velocity, acceleration, and lane index of the ego vehicle. The lane index is the index of the lane where the ego vehicle is located.

The action space of our autonomous driving agent is discrete and contains five decision-making behaviors: changing lanes to the right, changing lanes to the left, maintaining the current state, accelerating, and decelerating. According to the research results in Ref. [51], typically, the acceleration of the vehicle operated by a normal driver does not exceed 1.47 m·s$^{-2}$, and the deceleration is not less than −2 m·s$^{-2}$. Consequently, when our autonomous driving agent executes acceleration decision-making, the ego vehicle will accelerate at a fixed acceleration of 1.47 m·s$^{-2}$. Moreover, if the agent performs the decelerating decision-making, the ego vehicle decelerates at a fixed deceleration of −2.00 m·s$^{-2}$.

### 3.3. Reward function

The reward function plays a pivotal role in the performance of the RL agents. Our reward function was designed by considering factors related to travel efficiency, driving safety, and passenger comfort. Specifically, we encouraged autonomous driving agents to operate at high speeds. In addition, we penalized the agent if its driving policy caused a collision. An autonomous driving agent is subject to penalties if it performs high-speed lane-change maneuvers. Eq. (24) is the designed reward function $r(\cdot)$, where e denotes the natural logarithm, and $v_0$ is the ego vehicle speed. Moreover, $A = \{\text{vehicle changes lane}\}$, $B = \{v_0 > 30\}$, and $C = \{\text{collision}\}$ are the event sets. Here, collision refers to the collision between an ego vehicle and the surrounding social vehicles.

$$r(\cdot) = \begin{cases} e^{v_0/35-1} - v_0/350 & A\wedge B\wedge\neg C = 1 \\ e^{v_0/35-1} - 0.5 - v_0/100 & \neg(A\wedge B)\wedge C = 1 \\ e^{v_0/35-1} - v_0/350 - 0.5 - v_0/100 & A\wedge B\wedge C = 1 \\ e^{v_0/35-1} & \text{otherwise} \end{cases} \tag{24}$$

## 4. Simulations and experiments

### 4.1. Baseline

We set up comparisons with state-of-the-art RL agents in both simulations and experiments to benchmark the RRL-SG approach for trustworthy autonomous driving decision-making.

As the dueling DDQN (D3QN) is a state-of-the-art Q-learning algorithm [52,53], D3QN was adopted as one of the baselines in this study. Moreover, we leverage the proximal policy optimization (PPO) [54], soft actor-critic (SAC) [50,55], and observation adversarial RL (OARL) [56] algorithms as competitive baselines, representing state-of-the-art on-policy, off-policy, and robust RL technologies, respectively.

### 4.2. Metric

We employed the expected return to assess the comprehensive performance of the autonomous driving agents. The average running speed and number of collisions were utilized to evaluate the

travel efficiency and traffic safety of autonomous vehicles. In addition, Eq. (1) is used to measure policy robustness against adversarial perturbations, implying that the smaller the policy change attacked by the adversary, the stronger the robustness of the policy.

In the on-ramp merging scenario, in addition to the above metrics, we assessed the vehicle performance using the merging success rate. In this study, a successful on-ramp merging was defined as a vehicle entering the main lane completely from the ramp without experiencing any collisions within a test episode.

### 4.3. Simulations with SUMO

To assess the performance of the proposed decision-making technique for autonomous vehicles, we implemented model training and testing using the SUMO simulator. We leveraged SUMO to create stochastic dynamic traffic flows with different densities in highway and on-ramp merging scenarios. In addition, we trained five different runs of each approach with different random seeds and 400 episodes in a highway scenario with a normal-density traffic flow ($P = 0.12$). $P$ denotes the probability of starting the vehicle in seconds. The maximum time step for each episode is 200 s. The maximum traffic speed for all the lanes was set to 35.0 m·s$^{-1}$.

Unlike the highway scenario, the on-ramp merging scenario was utilized only for model testing.

#### 4.3.1. Highway scenario

Fig. 3 illustrates our evaluation scheme for the highway scenario. An ego vehicle is the golden RL-driven autonomous vehicle. $P$ is set as 0.06, 0.12, and 0.24 to produce the traffic flows with low, normal, and high densities, respectively. Autonomous driving agents were trained only in traffic flows with normal density. In the model-testing phase, traffic flows with low, normal, and high densities were leveraged for assessment. Each trained agent (including different random seeds) was evaluated for over 100 episodes. Each evaluation calculated the average metrics for ten episodes in testing. As we employ stochastic dynamic traffic flows, the environmental dynamics are continuously changing. To further verify policy robustness, each autonomous driving agent was attacked by optimal adversarial observational perturbations from the trained adversary during model testing. In other words, unlike the model training phase, in the test case with adversarial attacks, the autonomous driving agent receives states $\tilde{s}$ perturbed by the adversary model.

Fig. 4 shows the learning curves of the proposed RRL-SG method and the baselines for normal-density stochastic dynamic traffic flows. Overall, the results indicate that the proposed scheme outperforms the baselines in terms of return and safety. Clearly, our autonomous driving agent drastically reduces the number of collisions and enhances the learning efficiency during model training compared with the baselines because the proposed RSS-based safety mask forms a safe action subspace by shielding it against risky actions. Thus, sampling actions from the safe action subspace ensures decision safety and avoids redundant exploration.

During model testing, the final policy models based on five random seeds were evaluated for each method. Qualitatively, we report the average metrics in Table 1 for the model evaluation results. Bold numbers indicate the best values for each metric. In general, the results indicate that the RRL-SG agent surpasses the baseline by a large margin for all tasks in terms of robustness and safety. In contrast to the baselines, the JS divergence is approximately zero for changes in the RRL-SG policy attacked by the adversary model in the three stochastic dynamic traffic flows with different densities, implying that the RRL-SG policies were hardly affected by adversarial attacks. Moreover, unlike the D3QN, PPO, SAC, and OARL autonomous driving agents, the RRL-SG agent did not cause collisions in any of the test cases.

More specifically, in low-density traffic flows with and without adversarial attacks, the OARL autonomous driving agent performs comparably to the OARL agent and outperforms the D3QN, PPO, and SAC agents by a large margin in terms of return. In normal-density traffic flows without adversarial attacks, compared to the D3QN, PPO, SAC, and OARL agents, the RRL-SG agent gains approximately 22.31%, 7.22%, 10.34%, and 1.97% improvements with respect to the return, respectively. In high-density traffic flows without adversarial attacks, compared to the D3QN, PPO, SAC, and OARL agents, the returns of the RRL-SG agent improved by approximately 78.63%, 47.41%, 25.45%, and 13.84%, respectively. Additionally, in high-density traffic flows with adversarial attacks, compared with the D3QN, PPO, SAC, and OARL agents, the return of the RRL-SG agent was enhanced by approximately 7669.57%, 2666.25%, 511.57%, and 8.99%, respectively.

Fig. 5 illustrates the performance of the D3QN, PPO, SAC, OARL, and RRL-SG autonomous driving agents in stochastic dynamic traffic flows with different densities and attack situations. As shown in Fig. 5, adversarial attacks based on trained adversary models distinctly impact the comprehensive performance, travel efficiency, and safety of autonomous vehicles driven by baseline agents. For instance, in normal-density traffic flows, compared to the case without adversarial attacks, the number of collisions of the attacked D3QN, PPO, SAC, and OARL autonomous driving agents increased by approximately 358.82%, 583.33%, 1378.57%, and 5.71%, respectively. In contrast, the proposed RRL-SG autonomous driving agent performed consistently across all test cases, with zero collision accidents recorded.

Here, we empirically assessed policy robustness against perturbations in environmental dynamics by calculating the mean square deviation of returns for each method across all testing scenarios, including various traffic densities and attack scenarios. According to Table 1, the mean square deviations of the returns for the D3QN, PPO, SAC, OARL, and RRL-SG agents across all testing cases are 31.23, 16.11, 39.60, 12.73, and 7.50, respectively, indicating that the RRL-SG agent was the least affected by environmental changes compared to the baselines. In other words, the RRL-SG policy is robust and safe and exhibits stability, thus highlighting the

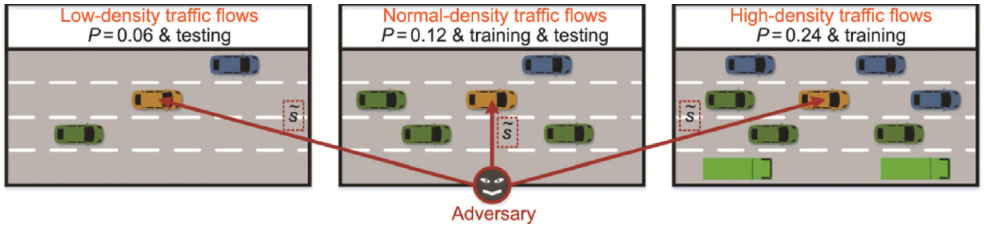


**Fig. 3.** Evaluation scheme based on highway scenarios with stochastic dynamic traffic flows and adversarial attacks. $\tilde{s}$: states perturbed by the adversary model.

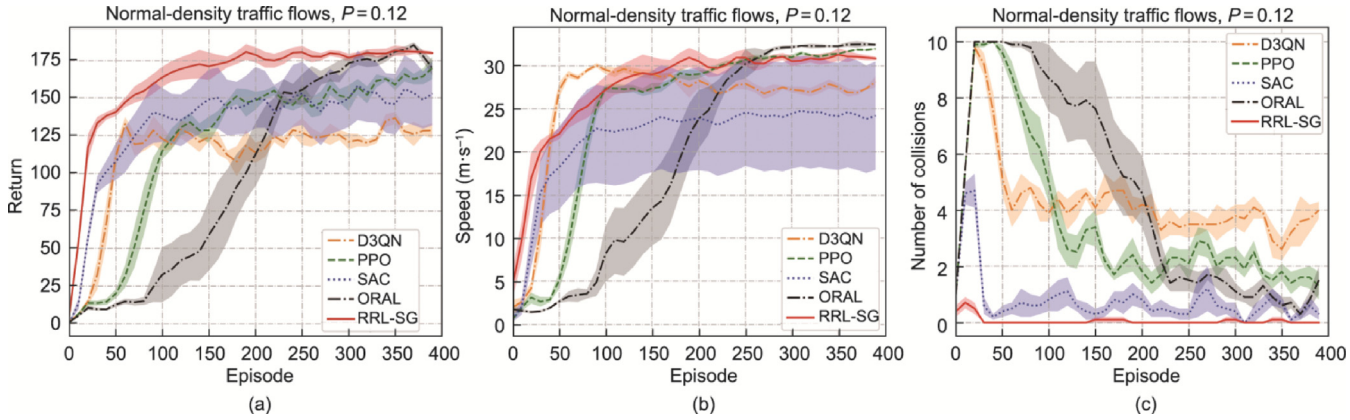


**Fig. 4.** Learning curves of autonomous driving agents in the normal-density stochastic dynamic traffic flows. (a) Return; (b) speed; (c) number of collisions.

**Table 1**
Statistical results of autonomous driving agents in the highway scenario with stochastic dynamic traffic flows under different densities and attack situations.

| Method | Metric | Low-density traffic flows | | Normal-density traffic flows | | High-density traffic flows | |
|---|---|---|---|---|---|---|---|
| | | Without attacks | With attacks | Without attacks | With attacks | Without attacks | With attacks |
| D3QN | Return | 151.64 ± 25.67 | 34.89 ± 42.83 | 148.72 ± 32.69 | 17.21 ± 32.97 | 100.82 ± 38.53 | 2.30 ± 14.71 |
| | Speed | 27.78 ± 4.95 | 16.33 ± 9.17 | 27.08 ± 6.15 | 13.09 ± 8.04 | 21.82 ± 5.90 | 10.08 ± 6.04 |
| | Robustness | – | 0.17 ± 0.07 | – | 0.21 ± 0.07 | – | 0.22 ± 0.07 |
| | Number of collisions | 0.68 ± 1.17 | 3.70 ± 3.29 | 1.02 ± 1.45 | 4.68 ± 3.27 | 2.84 ± 2.28 | 6.24 ± 2.51 |
| PPO | Return | 178.99 ± 11.10 | 59.91 ± 21.00 | 169.65 ± 16.15 | 31.24 ± 16.70 | 122.17 ± 27.04 | 6.46 ± 4.64 |
| | Speed | 32.65 ± 0.46 | 23.06 ± 2.76 | 32.24 ± 0.56 | 17.45 ± 4.93 | 30.10 ± 1.29 | 8.01 ± 4.43 |
| | Robustness | – | 0.21 ± 0.03 | – | 0.23 ± 0.03 | – | 0.25 ± 0.03 |
| | Number of collisions | 0.88 ± 0.89 | 7.86 ± 1.22 | 1.32 ± 1.14 | 9.02 ± 1.10 | 4.24 ± 1.96 | 9.92 ± 0.27 |
| SAC | Return | 174.41 ± 23.44 | 128.64 ± 66.03 | 164.85 ± 23.63 | 81.20 ± 58.03 | 143.55 ± 22.25 | 29.22 ± 44.26 |
| | Speed | 30.56 ± 3.51 | 26.40 ± 7.61 | 29.09 ± 3.75 | 21.61 ± 9.28 | 25.83 ± 3.73 | 9.97 ± 7.87 |
| | Robustness | – | 0.21 ± 0.21 | – | 0.26 ± 0.22 | – | 0.41 ± 0.23 |
| | Number of collisions | 0.06 ± 0.24 | 1.30 ± 2.02 | 0.28 ± 0.57 | 4.14 ± 3.35 | 0.72 ± 1.11 | 6.42 ± 3.90 |
| OARL | Return | **190.53 ± 1.88** | **187.06 ± 5.44** | 178.38 ± 10.33 | **179.49 ± 14.28** | 158.20 ± 21.91 | 163.96 ± 22.56 |
| | Speed | **33.04 ± 0.26** | **32.83 ± 0.39** | **32.30 ± 0.73** | **32.72 ± 0.42** | **31.65 ± 0.95** | **31.98 ± 0.76** |
| | Robustness | – | $(5.06 \pm 3.16) \times 10^{-4}$ | – | $(7.05 \pm 4.27) \times 10^{-4}$ | – | $(1.29 \pm 0.73) \times 10^{-3}$ |
| | Number of collisions | 0.02 ± 0.14 | 0.22 ± 0.41 | 0.70 ± 0.78 | 0.74 ± 1.00 | 2.18 ± 1.65 | 1.58 ± 1.44 |
| RRL-SG | Return | 189.91 ± 1.66 | 185.98 ± 8.38 | **181.90 ± 5.03** | 175.27 ± 17.30 | **180.09 ± 5.07** | **178.70 ± 7.56** |
| | Speed | 32.88 ± 0.36 | 32.02 ± 1.82 | 31.23 ± 1.07 | 29.87 ± 3.78 | 30.90 ± 1.08 | 30.59 ± 1.64 |
| | Robustness | – | **$(3.91 \pm 6.83) \times 10^{-13}$** | – | **$(3.94 \pm 10.92) \times 10^{-12}$** | – | **$(1.96 \pm 3.69) \times 10^{-12}$** |
| | Number of collisions | **0 ± 0** | **0 ± 0** | **0 ± 0** | **0 ± 0** | **0 ± 0** | **0 ± 0** |

Bold numbers indicate the best values for each metric.

primary contribution of this study toward achieving trustworthy decision-making for autonomous vehicles.

*4.3.2. On-ramp merging scenario*

To further evaluate the trustworthiness of the decisions of the autonomous driving agents, on-ramp merging was added as an additional testing scenario. Inappropriate merging behaviors can lead to typical outcomes, including congestion, collisions, and increased travel time.

The proposed evaluation scheme—based on an on-ramp merging scenario—is illustrated in Fig. 6(a). We directly deployed the model trained in the highway scenario to the on-ramp merging scenario for testing purposes. All autonomous driving agents were assessed in stochastic dynamic traffic flows with high density (i.e., $P = 0.24$) under different attack situations across a total of 100 episodes. Similar to the highway scenario, each model evaluation computed the average metrics over ten testing episodes with a maximum of 200 time steps in each episode.

As seen in Figs. 6(b) and (c), the RRL-SG autonomous driving agent outperforms the baselines by a significant margin, with or without adversarial attacks, in terms of both travel efficiency and merging success rate.

Table 2 presents the average metrics for the results of the model evaluation in the on-ramp merging scenario. Bold numbers represent the best in each column. For instance, without adversarial attacks, compared to the D3QN, PPO, SAC, and OARL agents, the RRL-SG agent gains approximately 16.97%, 21.91%, 29.63%, and 21.18% improvements with respect to return, respectively. Without adversarial attacks, compared with D3QN, PPO, SAC, and OARL, the speed of the RRL-SG agent increased by approximately 31.84%, 42.60%, 62.62%, and 40.69%, respectively. As shown in Table 2, the robustness of the RRL-SG policy was significantly better than that of the baseline policies.

As shown in Fig. 6(c) and Table 2, our RRL-SG agent can complete the on-ramp merging task with a probability of 100.00%, regardless of the presence or absence of adversarial attacks. In other words, adversarial attacks on observations have almost no impact on the RRL-SG policy. In addition, with adversarial attacks, compared to the D3QN, PPO, SAC, and OARL agents, the RRL-SG agent gains approximately 13.00%, 9.00%, 6.00%, and 1.00% improvements in the merging success rate, respectively.

The environmental dynamics associated with the on-ramp merging scenario were notably distinct from those of the highway scenario. Because we utilize stochastic dynamic traffic flows, the environmental dynamics are subject to continuous changes. Here,

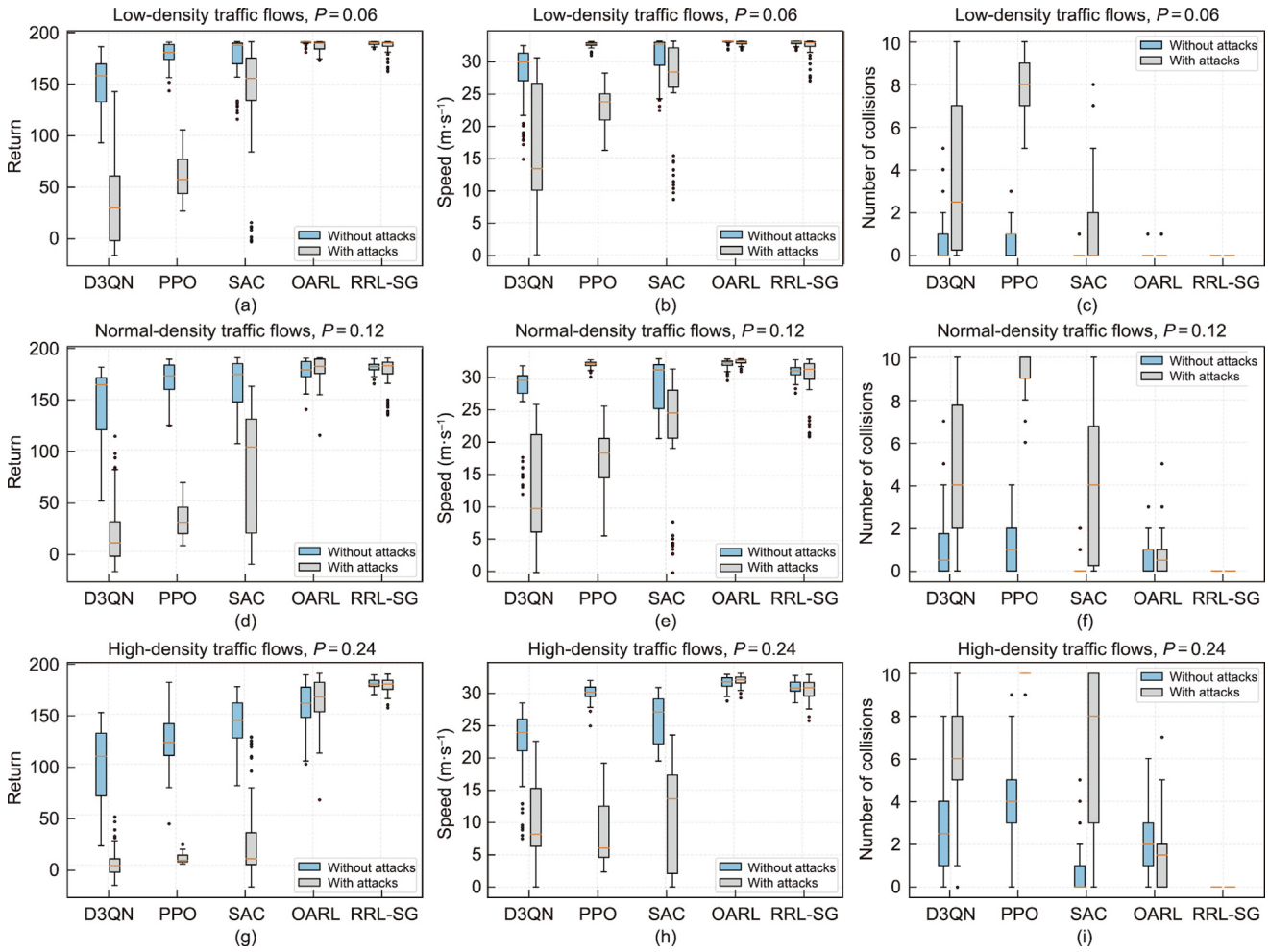


**Fig. 5.** Performance of autonomous driving agents in the highway scenario under different traffic densities and attack situations. (a–i) Return, speed, and number of collisions of autonomous driving agents in the low-density, normal-density, and high-density stochastic dynamic traffic flows with different attack situations.

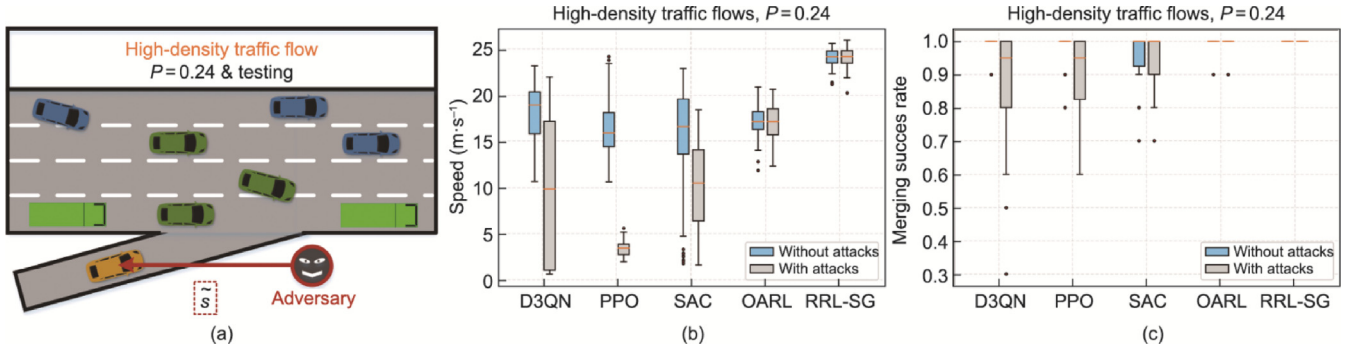


**Fig. 6.** Performance of autonomous driving agents in the on-ramp merging scenario with stochastic dynamic traffic flows under different attack situations. (a) Schematic diagram of the adopted on-ramp merging scenario. (b) Speed and (c) merging success rates of autonomous driving agents in the high-density stochastic dynamic traffic flows with different attack situations.

we empirically evaluate policy robustness against perturbations in environmental dynamics using the mean square deviation of returns for each agent under different attack situations. According to Table 2, the mean square deviations of the returns for the D3QN, PPO, SAC, OARL, and RRL-SG agents under the different attack conditions were 19.81, 7.15, 20.51, 6.59, and 4.04, respectively, implying that the RRL-SG agent exhibited superior policy robustness against environmental changes, thus making it the least susceptible to environmental changes compared to the baselines. These results highlight our pivotal contribution toward trustworthy decision-making for autonomous vehicles.

### 4.4. Experiments with a real autonomous vehicle

We conducted physical platform experiments using a real low-speed autonomous vehicle, Hunter (AgileX Robotics, China), to further verify the trustworthiness of the proposed approach. As shown in Fig. 7(a), Hunter is equipped with a 16-line light

**Table 2**
Statistical results of autonomous driving agents in the on-ramp merging scenario with stochastic dynamic traffic flows under different attack situations.

| Method | Return | | Speed | | Robustness | | Merging success rate | |
|---|---|---|---|---|---|---|---|---|
| | Without attacks | With attacks | Without attacks | With attacks | Without attacks | With attacks | Without attacks | With attacks |
| D3QN | 128.02 ± 12.12 | 97.15 ± 27.49 | 18.28 ± 3.22 | 9.55 ± 7.81 | – | 0.39 ± 0.07 | 0.98 ± 0.04 | 0.87 ± 0.17 |
| PPO | 122.83 ± 11.73 | 75.42 ± 2.57 | 16.90 ± 3.20 | 3.36 ± 0.81 | – | 0.19 ± 0.02 | 0.98 ± 0.04 | 0.91 ± 0.10 |
| SAC | 115.51 ± 23.66 | 97.95 ± 17.35 | 14.82 ± 6.67 | 9.96 ± 4.89 | – | 0.28 ± 0.16 | 0.96 ± 0.08 | 0.94 ± 0.08 |
| OARL | 123.57 ± 6.40 | 123.11 ± 6.77 | 17.13 ± 1.81 | 17.02 ± 1.91 | – | $(0.89 \pm 1.15) \times 10^{-3}$ | 1.00 ± 0.02 | 0.99 ± 0.03 |
| RRL-SG | **149.74 ± 3.93** | **149.61 ± 4.15** | **24.10 ± 1.02** | **24.04 ± 1.10** | – | $\mathbf{(0.87 \pm 1.11) \times 10^{-8}}$ | **1.00 ± 0.00** | **1.00 ± 0.00** |

Bold numbers represent the best in each column.

detection and ranging (LiDAR), two stereo cameras, eight ultrasonic sensors, and one "Jetson Xavier NX 16 GB" edge computing system (NVIDIA, USA). Hence, the RL policy model can generate decision commands in real-time based on the perceived states from the onboard sensors, with all computations performed on the NVIDIA Jetson platform. All models trained in the SUMO simulator were directly deployed in Hunter and tested in a laboratory environment with a free space measuring 8 m × 8 m. Only the trained policy models were tested here and were not trained further (i.e., the model parameters were fixed). The policy model required approximately 0.002 s to perform a single inference. The sampling frequency of Hunter was 30 Hz. As the evaluated policy model executes a decision once it receives a set of sampled states, Hunter's decision-making frequency is 30 Hz.

Figs. 7(b) and (c) illustrate the experimental schemes. Similar to model testing in the simulator, we instantiated five final policy models trained by each algorithm using five different random seeds and evaluated each model under varying conditions, with and without adversarial attacks. In the experimental case shown in Fig. 7(b), Hunter's perception information consisted only of the original environmental observations without any adversarial observational perturbations. In contrast, as shown in Fig. 7(c), Hunter senses the driving environment information containing both the original environmental observations and the adversarial perturbations generated by the trained adversarial models. Additionally, the environmental dynamics change significantly from the simulation environment to the real-world physical platform.

The experimental space was free of static or dynamic obstacles, implying that Hunter should be able to maintain a straight run without attacks from an adversary model. We assessed each policy model during the period (150 time steps) in which Hunter drove from one side to the other. In the test case with adversarial attacks, the attacks started at the 75th time step. Hunter can execute five decision-making behaviors: turning right, turning left, maintaining the current state, accelerating, and decelerating.

Fig. 8 shows the global motion trajectories of autonomous vehicles driven by different agents in different attack situations, wherein all policy models enable Hunter to continue running straight without adversarial attacks. However, in the test case with adversarial attacks, the performance of the baseline models was affected to varying degrees. Specifically, all D3QN autonomous driving agents, four-fifths of the PPO agents, all SAC agents, and one-fifth of the OARL agents make turning decisions under adversarial attacks. In contrast, the five proposed RRL-SG policy models perform consistently in all cases, for example, the RRL-SG-driven Hunter can maintain a straight run even when it suffers from attacks by the adversary model. For more visual results, please refer to Video S1 in Appendix A.

To illustrate the impact of adversarial attacks on the policy model, Fig. 9 shows the probability distribution of actions based on the D3QN and RRL-SG policies before and after encountering adversarial perturbations. We leverage the softmax function to convert the output of the D3QN policy model, which consists of $Q$ values for each action, into a probability distribution over the actions. The action distribution based on the RRL-SG policy shows hardly any change compared with the distribution based on the D3QN policy. Specifically, in the absence of adversarial attacks, the probabilities of the five decision actions based on the D3QN policy were approximately 12.77%, 19.55%, 20.61%, 34.45%, and 12.63%, respectively. Under adversarial attacks, the probabilities of the five decision actions based on the D3QN policy are approximately 38.14%, 15.30%, 17.95%, 15.40%, and 13.21%, respectively, thus explaining why adversarial attacks can cause the D3QN-driven Hunter to continue running suddenly in a straight-line turn. In addition, without adversarial perturbations, the probabilities of the five decision actions regarding the RRL-SG policy are approximately

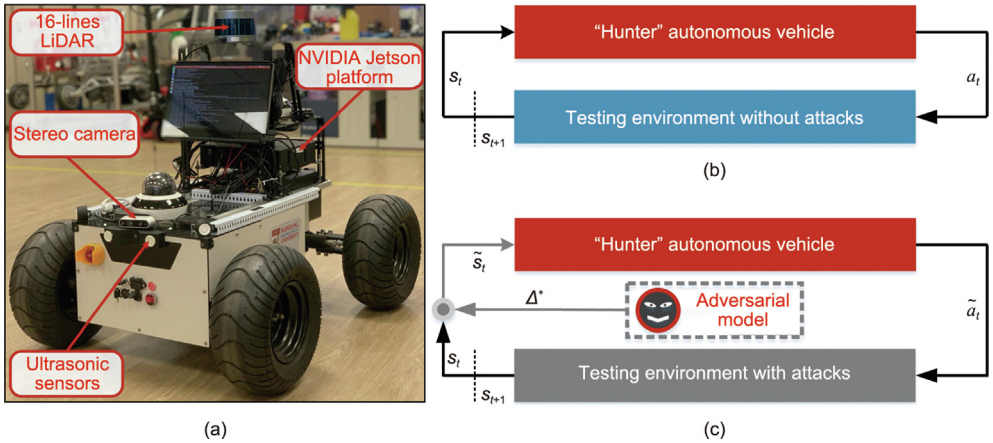


**Fig. 7.** Experimental setup for the real physical system. (a) "Hunter" autonomous vehicle used for experimental validation. (b) Illustration of experimental scheme without adversarial attacks. (c) Illustration of experimental scheme with adversarial attacks. LiDAR: light detection and ranging.

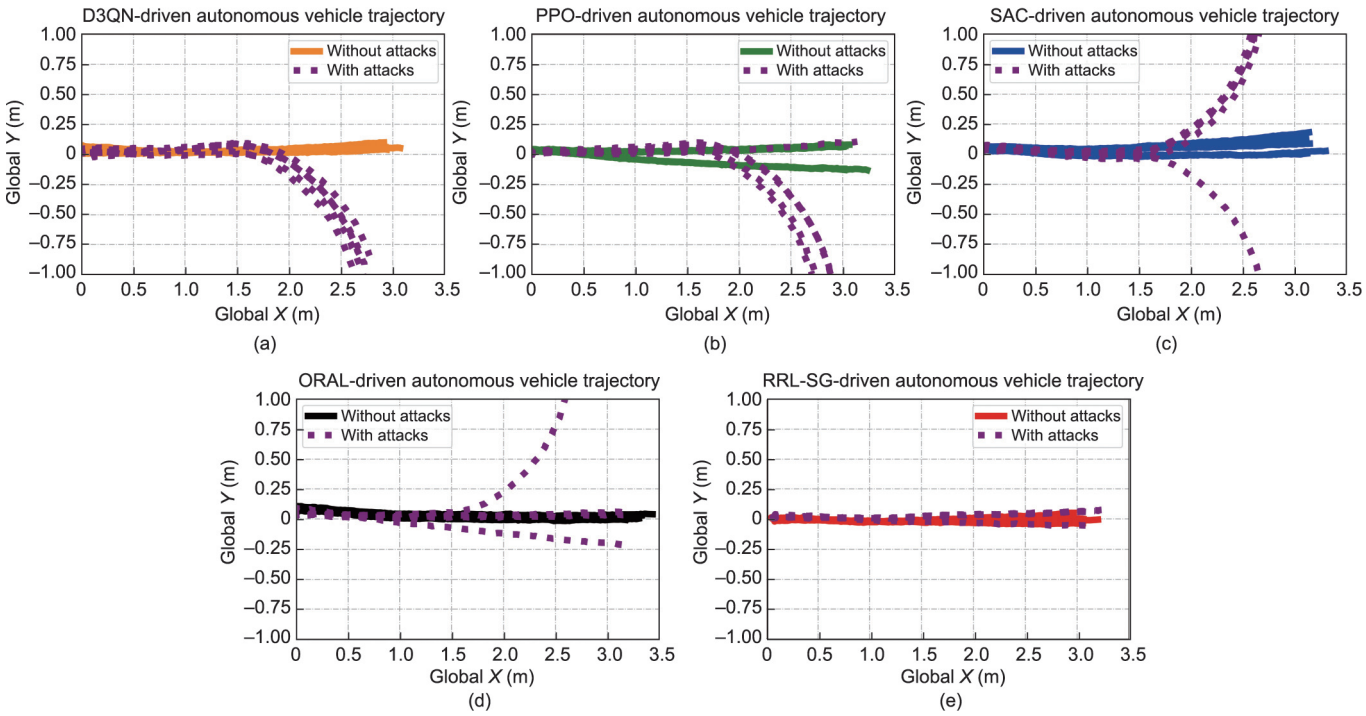


**Fig. 8.** Global motion trajectories of autonomous vehicles driven by different agents under different attack situations. Global motion trajectories of autonomous vehicles based on the (a) D3QN, (b) PPO, (c) SAC, (d) OARL, and (e) RRL-SG agents under different attack situations.

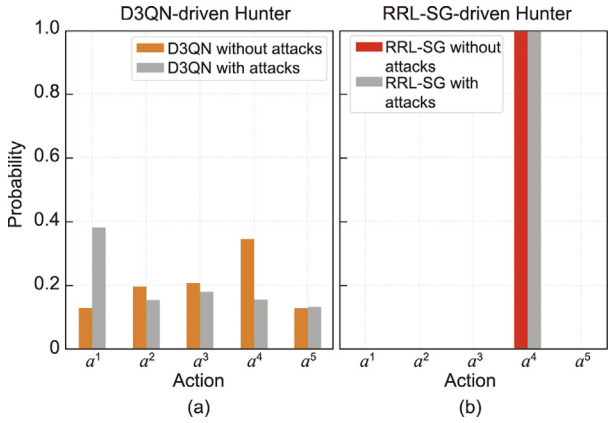


**Fig. 9.** Probability distribution of actions based on different autonomous driving policies under different attack situations. Probability distributions of actions under the (a) D3QN and (b) RRL-SG policies with and without adversarial attacks. $a^1$, $a^2$, $a^3$, $a^4$, and $a^5$ denote turning right, turning left, keeping the current state, accelerating, and decelerating, respectively.

$3.97 \times 10^{-13}\%, 1.74 \times 10^{-13}\%, 2.48 \times 10^{-12}\%$, 100.00%, and $5.57 \times 10^{-13}\%$, respectively. With adversarial perturbations, the probabilities of the five decision actions concerning the RRL-SG policy were approximately $6.83 \times 10^{-8}\%, 3.47 \times 10^{-8}\%, 2.09 \times 10^{-7}\%$, 100.00%, and $6.45 \times 10^{-8}\%$, respectively. Therefore, Hunter, based on the RRL-SG policy, can maintain its running status without being affected by adversarial perturbations.

## 5. Conclusions

In this study, we introduce the RRL-SG technique, which empowers autonomous vehicles to make trustworthy decisions. The proposed paradigm attempts to ensure trustworthiness in terms of policy robustness and collision safety. Specifically, the adversary model is trained online to simulate the worst-case uncertainty by generating optimal adversarial perturbations on observed states and environmental dynamics. Meanwhile, the ARAC approach is advanced to facilitate the agent in learning robust policies against multiple uncertainties from the adversary. In addition, we devise a safety mask to ensure the collision safety of the autonomous driving agent during both the training and testing processes using the interpretable knowledge model RSS.

The evaluation results of the simulations with stochastic dynamic traffic flow and the experiment with a real autonomous vehicle indicate that the proposed RRL-SG scheme enables the autonomous driving agent to learn trustworthy policies against adversarial environmental uncertainties. In addition, compared with the four baselines, the RRL-SG driving policies ensure superior robustness and safety. Notably, our autonomous agent consistently delivers a more stable performance than the baselines in both simulations and experiments.

Although we demonstrated the potential of the proposed approach, one limitation remains. While the RRL-SG solution leverages the worst-case setting and interpretable knowledge model, the provision of theoretical guarantees for robustness and safety of autonomous driving models remains a critical subject for future research. Consequently, in the future, we will investigate certifiable and interpretable decision-making techniques to further enhance the trustworthiness of autonomous driving systems.

## Acknowledgment

This work was supported in part by the Start-Up Grant-Nanyang Assistant Professorship Grant of Nanyang Technological University, the Agency for Science, Technology and Research (A*STAR) under Advanced Manufacturing and Engineering (AME) Young Individual Research under Grant (A2084c0156), the MTC Individual Research Grant (M22K2c0079), the ANR-NRF Joint Grant

(NRF2021-NRF-ANR003 HM Science), and the Ministry of Education (MOE) under the Tier 2 Grant (MOE-T2EP50222-0002).

## Compliance with ethics guidelines

Xiangkun He, Wenhui Huang, and Chen Lv declare that they have no conflict of interest or financial conflicts to disclose.

## Appendix A. Supplementary material

Supplementary material to this article can be found online at https://doi.org/10.1016/j.eng.2023.10.005.

## References

[1] Yang B, Cao X, Xiong K, Yuen C, Guan YL, Leng S, et al. Edge intelligence for autonomous driving in 6G wireless system: design challenges and solutions. IEEE Wireless Commun 2021;28(2):40–7.
[2] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Von Luxburg U, Guyon I, Bengio S, Wallach H, Fergus R, editors. NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems; 2017 Dec 4–9; Long Beach, CA, USA. New York City: Curran Associates Inc.; 2017. p. 6000–10.
[3] Wang J, Huang H, Li K, Li J. Towards the unified principles for level 5 autonomous vehicles. Engineering 2021;7(9):1313–25.
[4] Mollah MB, Zhao J, Niyato D, Guan YL, Yuen C, Sun S, et al. Blockchain for the internet of vehicles towards intelligent transportation systems: a survey. IEEE Internet Things J 2021;8(6):4157–85.
[5] Li J, Shao W, Wang H. Key challenges and Chinese solutions for SOTIF in intelligent connected vehicles. Engineering 2023;31(12):27–30.
[6] Feng S, Yan X, Sun H, Feng Y, Liu HX. Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. Nat Commun 2021;12(1):e748.
[7] Liu J, Luo Y, Zhong Z, Li K, Huang H, Xiong H. A probabilistic architecture of long-term vehicle trajectory prediction for autonomous driving. Engineering 2022;19(12):228–39.
[8] He X, Wu J, Huang Z, Hu Z, Wang J, Sangiovanni-Vincentelli A, et al. Fear-neuro-inspired reinforcement learning for safe autonomous driving. IEEE Trans Pattern Anal Mach Intell 2023 Oct::1–13.
[9] Yuan K, Huang Y, Yang S, Zhou Z, Wang Y, Cao D, et al. Evolutionary decisionmaking and planning for autonomous driving based on safe and rational exploration and exploitation. Engineering. In press.
[10] Huang W, Zhou Y, He X, Lv C. Goal-guided transformer-enabled reinforcement learning for efficient autonomous navigation. IEEE Trans Intell Transp Syst 2023 Sep:1–14.
[11] Zhang Y, Li C, Luan TH, Yuen C, Fu Y. Collaborative driving: learning-aided joint topology formulation and beamforming. IEEE Veh Technol Mag 2022;17(2):103–11.
[12] Wu J, Huang Z, Hu Z, Lv C. Toward human-in-the-loop AI: enhancing deep reinforcement learning via real-time human guidance for autonomous driving. Engineering 2023;21(2):75–91.
[13] Wang H, Khajepour A, Cao D, Liu T. Ethical decision making in autonomous vehicles: challenges and research progress. IEEE Intell Transp Syst Mag 2022;14(1):6–17.
[14] He X, Lv C. Toward personalized decision making for autonomous vehicles: a constrained multi-objective reinforcement learning technique. Transp Res Part C Emerging Technol 2023;156:104352.
[15] Tang X, Yang K, Wang H, Wu J, Qin Y, Yu W, et al. Prediction-uncertainty-aware decision-making for autonomous vehicles. IEEE Trans Intell Veh 2022;7(4):849–62.
[16] Liu J, Wang H, Peng L, Cao Z, Yang D, Li J. PNNUAD: perception neural networks uncertainty aware decision-making for autonomous vehicle. IEEE Trans Intell Transp Syst 2022;23(12):24355–68.
[17] Li G, Qiu Y, Yang Y, Li Z, Li S, Chu W, et al. Lane change strategies for autonomous vehicles: a deep reinforcement learning approach based on transformer. IEEE Trans Intell Veh 2023;8(3):2197–211.
[18] Urmson C, Anhalt J, Bagnell D, Baker C, Bittner R, Clark MN, et al. Autonomous driving in urban environments: boss and the urban challenge. J Field Rob 2008;25(8):425–66.
[19] Montemerlo M, Becker J, Bhat S, Dahlkamp H, Dolgov D, Ettinger S, et al. Junior: the Stanford entry in the urban challenge. J Field Rob 2008;25(9):569–97.
[20] Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, et al. Human-level control through deep reinforcement learning. Nature 2015;518(7540):529–33.
[21] Vinyals O, Babuschkin I, Czarnecki WM, Mathieu M, Dudzik A, Chung J, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. Nature 2019;575(7782):350–4.
[22] He X, Chen H, Lv C. Robust multiagent reinforcement learning toward coordinated decision-making of automated vehicles. SAE Int J Veh Dyn Stab NVH 2023;7(4):2023.
[23] Hieu NQ, Hoang DT, Niyato D, Wang P, Kim DI, Yuen C. Transferable deep reinforcement learning framework for autonomous vehicles with joint radar-data communications. IEEE Trans Commun 2022;70(8):5164–80.
[24] Duan J, Li SE, Guan Y, Sun Q, Cheng B. Hierarchical reinforcement learning for self-driving decision-making without reliance on labelled driving data. IET Intell Transp Syst 2020;14(5):297–305.
[25] Kiran BR, Sobh I, Talpaert V, Mannion P, Al Sallab AA, Yogamani S, et al. Deep reinforcement learning for autonomous driving: a survey. IEEE Trans Intell Transp Syst 2022;23(6):4909–26.
[26] Ye F, Wang P, Chan CY, Zhang J. Meta reinforcement learning-based lane change strategy for autonomous vehicles. In: Proceedings of 2021 IEEE Intelligent Vehicles Symposium (IV); 2021 Jul 11–17; Nagoya, Japan. Piscataway: IEEE; 2021. p. 223–30.
[27] Wang G, Hu J, Li Z, Li L. Harmonious lane changing via deep reinforcement learning. IEEE Trans Intell Transp Syst 2022;23(5):4642–50.
[28] Li G, Yang Y, Li S, Qu X, Lyu N, Li SE. Decision making of autonomous vehicles in lane change scenarios: deep reinforcement learning approaches with risk awareness. Transp Res Part C 2022;134:e103452.
[29] Mirchevska B, Pek C, Werling M, Althoff M, Boedecker J. High-level decision making for safe and reasonable autonomous lane changing using reinforcement learning. In: Proceedings of 2018 21st International Conference on Intelligent Transportation Systems; 2018 Nov 4–7; Maui, HI, USA. Piscataway: IEEE; 2018. p. 2156–62.
[30] Lubars J, Gupta H, Chinchali S, Li L, Raja A, Srikant R, et al. Combining reinforcement learning with model predictive control for on-ramp merging. In: Proceedings of 2021 IEEE International Intelligent Transportation Systems Conference; 2021 Sep 19–22; Indianapolis, IN, USA. Piscataway: IEEE; 2021. p. 942–7.
[31] Wang H, Gao H, Yuan S, Zhao H, Wang K, Wang X, et al. Interpretable decision-making for autonomous vehicles at highway on-ramps with latent space reinforcement learning. IEEE Trans Veh Technol 2021;70(9):8707–19.
[32] Bouton M, Nakhaei A, Fujimura K, Kochenderfer MJ. Cooperation-aware reinforcement learning for merging in dense traffic. In: Proceedings of 2019 IEEE Intelligent Transportation Systems Conference; 2019 Oct 27–30; Auckland, New Zealand. Piscataway: IEEE; 2019. p. 3441–7.
[33] Qiao Z, Tyree Z, Mudalige P, Schneider J, Dolan JM. Hierarchical reinforcement learning method for autonomous vehicle behavior planning. In: Proceedings of 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems; 2020 Oct 24–2021 Jan 24; Las Vegas, NV, USA. Piscataway: IEEE; 2021. p. 6084–9.
[34] He X, Lou B, Yang H, Lv C. Robust decision making for autonomous vehicles at highway on-ramps: a constrained adversarial reinforcement learning approach. IEEE Trans Intell Transp Syst 2023;24(4):4103–13.
[35] Hoel CJ, Driggs-Campbell K, Wolff K, Laine L, Kochenderfer MJ. Combining planning and deep reinforcement learning in tactical decision making for autonomous driving. IEEE Trans Intell Veh 2020;5(2):294–305.
[36] Zhang Y, Gao B, Guo L, Guo H, Chen H. Adaptive decision-making for automated vehicles under roundabout scenarios using optimization embedded reinforcement learning. IEEE Trans Neural Networks Learn Syst 2021;32(12):5526–38.
[37] He X, Lv C. Toward intelligent connected e-mobility: energy-aware cooperative driving with deep multiagent reinforcement learning. IEEE Veh Technol Mag 2023;18(3):101–9.
[38] Nageshrao S, Tseng HE, Filev D. Autonomous highway driving using deep reinforcement learning. In: Proceedings of 2019 IEEE International Conference on Systems, Man and Cybernetics; 2019 Oct 6–9; Bari, Italy. Piscataway: IEEE; 2019. p. 2326–31.
[39] Gangopadhyay B, Soora H, Dasgupta P. Hierarchical program-triggered reinforcement learning agents for automated driving. IEEE Trans Intell Transp Syst 2022;23(8):10902–11.
[40] Cao Z, Xu S, Jiao X, Peng H, Yang D. Trustworthy safety improvement for autonomous driving using reinforcement learning. Transp Res Part C 2022;138:103656.
[41] Shalev-Shwartz S, Shammah S, Shashua A. On a formal model of safe and scalable self-driving cars. 2017. arXiv:1708.06374.
[42] Shalev-Shwartz S, Shammah S, Shashua A. Vision zero: can roadway accidents be eliminated without compromising traffic throughput? 2018. arXiv:1901.05022.
[43] Lopez PA, Behrisch M, Bieker-Walz L, Erdmann J, Flötteröd YP, Hilbrich R, et al. Microscopic traffic simulation using SUMO. In: Proceedings of 2018 21st International Conference on Intelligent Transportation Systems; 2018 Nov 4–7; Maui, HI, USA. Piscataway: IEEE; 2018. p. 2575–82.
[44] Lin J. Divergence measures based on the Shannon entropy. IEEE Trans Inf Theory 1991;37(1):145–51.
[45] Huszár F. How (not) to train your generative model: scheduled sampling, likelihood, adversary? 2015. arXiv:1511.05101.
[46] Huang W, Zhang C, Wu J, He X, Zhang J, Lv C. Sampling efficient deep reinforcement learning through preference-guided stochastic exploration. IEEE Trans Neural Networks Learn Syst 2023 Oct:1–12.
[47] Hoffman AJ, Karp RM. On nonterminating stochastic games. Manage Sci 1966;12(5):359–70.
[48] Hansen TD, Miltersen PB, Zwick U. Strategy iteration is strongly polynomial for 2-player turn-based stochastic games with a constant discount factor. J ACM 2013;60(1):1–16.
[49] Mazalov V. Mathematical game theory and applications. Chichester: John Wiley & Sons Ltd; 2014.

[50] Haarnoja T, Zhou A, Abbeel P, Levine S. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: Proceedings of the 35th International Conference on Machine Learning; 2018. p. 1861–70.
[51] Bae I, Moon J, Jhung J, Suk H, Kim T, Park H, et al. Self-driving like a human driver instead of a robocar: personalized comfortable driving experience for autonomous vehicles. 2020. arXiv:2001.03908.
[52] Wang Z, Schaul T, Hessel M, van Hasselt H, Lanctot M, de Freitas N. Dueling network architectures for deep reinforcement learning. In: Balcan MF, Weinberger KQ, editors. ICML'16: Proceedings of the 33rd International Conference on International Conference on Machine Learning—Volume 48; 2016 Jun 19–24; New York City, NY, USA. JMLR.org; 2016. p. 1995–2003.
[53] Hessel M, Modayil J, van Hasselt H, Schaul T, Ostrovski G, Dabney W, et al. Rainbow: combining improvements in deep reinforcement learning. In: McIlraith SA, Weinberger KQ, editors. AAAI'18/IAAI'18/EAAI'18: Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence; 2018 Feb 2–7; New Orleans, LA, USA. Palo Alto: AAAI Press; 2018. p. 3215–22.
[54] Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal policy optimization algorithms. 2017. arXiv:1707.06347.
[55] Christodoulou P. Soft actor-critic for discrete action settings. 2019. arXiv:1910.07207.
[56] He X, Yang H, Hu Z, Lv C. Robust lane change decision making for autonomous vehicles: an observation adversarial reinforcement learning approach. IEEE Trans Intell Veh 2023;8(1):184–93.