Fear-Neuro-Inspired Reinforcement Learning for Safe Autonomous Driving

Xiangkun He [1], Jingda Wu [2], Zhiyu Huang [2], Zhongxu Hu [2], Jun Wang [2], Alberto Sangiovanni-Vincentelli [2], and Chen Lv [2]

[1]Nanyang Technological University
[2]Affiliation not available

October 31, 2023

Abstract

Ensuring safety and achieving human-level driving performance remain challenges for autonomous vehicles, especially in safety-critical situations. As a key component of artificial intelligence, reinforcement learning is promising and has shown great potential in many complex tasks; however, its lack of safety guarantees limits its real-world applicability. Hence, further advancing reinforcement learning, especially from the safety perspective, is of great importance for autonomous driving. As revealed by cognitive neuroscientists, the amygdala of the brain can elicit defensive responses against threats or hazards, which is crucial for survival in and adaptation to risky environments. Drawing inspiration from this scientific discovery, we present a fear-neuro-inspired reinforcement learning framework to realize safe autonomous driving through modeling the amygdala functionality. This new technique facilitates an agent to learn defensive behaviors and achieve safe decision making with fewer safety violations. Through experimental tests, we show that the proposed approach enables the autonomous driving agent to attain state-of-the-art performance compared to the baseline agents and perform comparably to 30 certified human drivers, across various safety-critical scenarios. The results demonstrate the feasibility and effectiveness of our framework while also shedding light on the crucial role of simulating the amygdala function in the application of reinforcement learning to safety-critical autonomous driving domains.

Fear-Neuro-Inspired Reinforcement Learning for Safe Autonomous Driving

Xiangkun He, *Member, IEEE*, Jingda Wu, *Graduate Student Member, IEEE*, Zhiyu Huang, *Graduate Student Member, IEEE*, Zhongxu Hu, *Member, IEEE*, Jun Wang, Alberto Sangiovanni-Vincentelli, *Life Fellow, IEEE*, and Chen Lv, *Senior Member, IEEE*

Abstract—Ensuring safety and achieving human-level driving performance remain challenges for autonomous vehicles, especially in safety-critical situations. As a key component of artificial intelligence, reinforcement learning is promising and has shown great potential in many complex tasks; however, its lack of safety guarantees limits its real-world applicability. Hence, further advancing reinforcement learning, especially from the safety perspective, is of great importance for autonomous driving. As revealed by cognitive neuroscientists, the amygdala of the brain can elicit defensive responses against threats or hazards, which is crucial for survival in and adaptation to risky environments. Drawing inspiration from this scientific discovery, we present a fear-neuro-inspired reinforcement learning framework to realize safe autonomous driving through modeling the amygdala functionality. This new technique facilitates an agent to learn defensive behaviors and achieve safe decision making with fewer safety violations. Through experimental tests, we show that the proposed approach enables the autonomous driving agent to attain state-of-the-art performance compared to the baseline agents and perform comparably to 30 certified human drivers, across various safety-critical scenarios. The results demonstrate the feasibility and effectiveness of our framework while also shedding light on the crucial role of simulating the amygdala function in the application of reinforcement learning to safety-critical autonomous driving domains.

Index Terms—Trustworthy artificial intelligence, reinforcement learning, safe decision making, autonomous driving

1 Introduction

Autonomous driving has attracted considerable attention from both academia and industry across the globe in recent years. The societal benefits of this paradigm are expected to include safer transportation, reduced congestion and lower emissions. However, the safety aspect of autonomous driving is still a major concern for large-scale deployment. Many real-world scenarios contains inevitable nonstationarity and uncertainty, which may lead autonomous vehicles to exhibit undesirable and unsafe driving behaviors and might even cause fatal casualties. To deal with these potential risks, there is still a long way to go to meet the strict requirements and high expectations with regard to the deployment of autonomous driving in society.

Modern artificial intelligence (AI) technologies have made numerous accomplishments [1], [2], [3], [4], exerting a strong impetus on the advancement of autonomous driving [5], [6]. Noticeably, reinforcement learning (RL) has emerged as a prominent field within AI, demonstrating remarkable achievements across various challenging decision tasks, such as Go [7], StarCraft [8], and autonomous racing [9]. Consequently, researchers have attempted to explore various RL algorithms along with their applications in autonomous driving [10]. Although existing approaches have achieved many compelling results, the lack of safety guarantees limits the applicability of RL in safety-critical autonomous driving domains. In light of this concern, many researchers have made efforts to study safe RL methods for ensuring the safety of autonomous vehicles. A common paradigm is to combine traditional RL algorithms with safety checkers [11] or constraints [12] to optimize driving policies while guaranteeing or encouraging safety. Yet it is inevitable that the agent will encounter numerous hazardous situations before it can effectively learn to avoid safety violations, even with the integration of sophisticated techniques to minimize the likelihood of failures.

Recently, some researchers have advocated for increased research efforts in "NeuroAI" since it holds the promising potential to catalyze the advancement of next-generation AI technologies [13]. RL theory is derived from the neuroscientific and psychological perspectives on organism behavior [14]. A common assumption regarding RL from the brain science perspective is that the dopamine neurons in the midbrain code for reward prediction errors, which enable the striatum to learn rewarding behaviors [15]. Most existing computational RL frameworks can be represented with this mechanism [16]. However, in recent years, many neuroscientists have argued that the amygdala plays a central role in the RL function of the brain, perhaps a more important role than the striatum but certainly a more important role than is attributed to it in current RL frameworks [15], [16]. The amygdala fear circuit in the brain can predict dangers and elicit defensive behavioral responses against threats and harms; this is crucial for survival in and adaptation to

- *Xiangkun He, Jingda Wu, Zhiyu Huang, Zhongxu Hu and Chen Lv are with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore 639798. E-mail: {xiangkun.he, zhongxu.hu, lyuchen}@ntu.edu.sg, {jingda001,zhiyu001}@e.ntu.edu.sg.*
- *Jun Wang is with the Department of Computer Science, University College London, London WC1E 6BT, U.K. E-mail: jun.wang@cs.ucl.ac.uk.*
- *Alberto Sangiovanni-Vincentelli is with the Electrical Engineering and Computer Sciences (EECS) Department, University of California at Berkeley, Berkeley, CA 94720 USA. E-mail: alberto@berkeley.edu.*

Manuscript received December 8, 2022.
(Corresponding author: Chen Lv.)

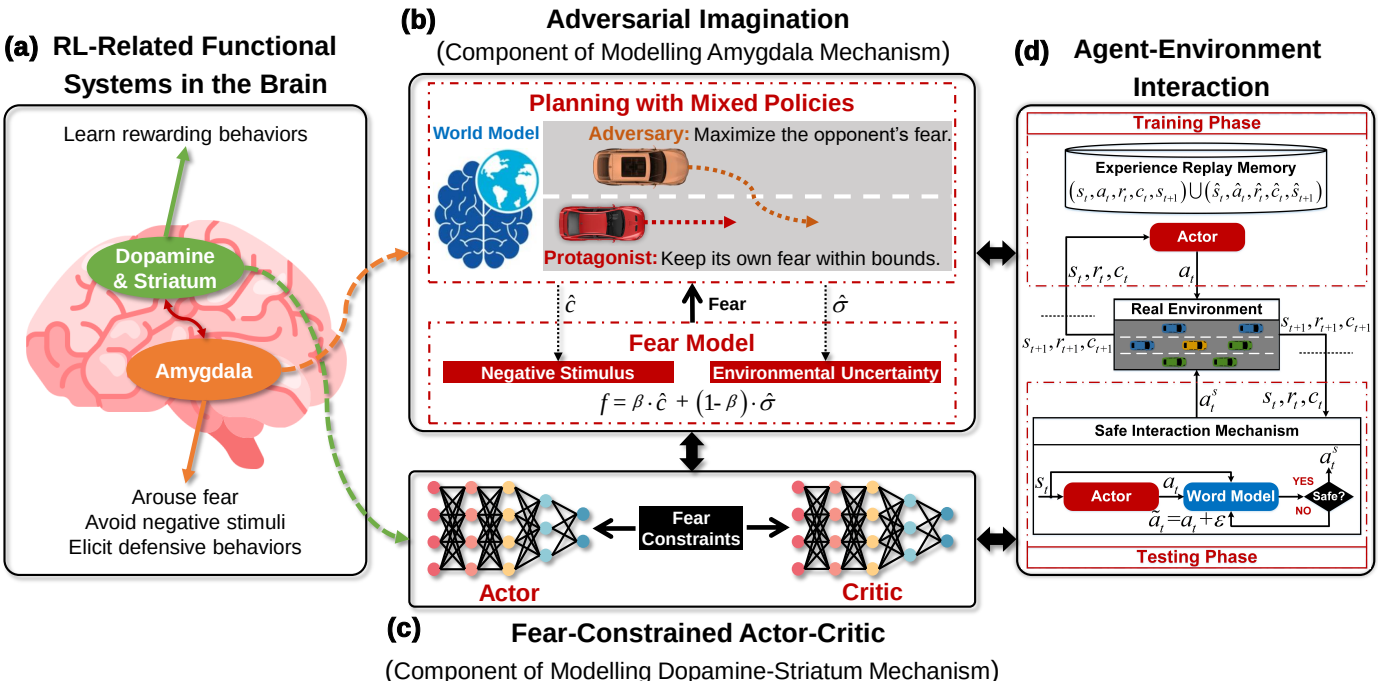


Fig. 1. Schematic of the proposed FNI-RL framework for safe autonomous driving. (a) RL-related functional systems in the brain. (b) Adversarial imagination module for simulating the amygdala mechanism. (c) Fear-constrained actor-critic technique. (d) Agent-environment interaction loop.

potential risky environments [17]. Amygdala lesions inhibit the fear learning and avoidance behavior elicited by threats. Moreover, some studies in neuroscience and psychology have highlighted the necessity of actively forecasting hazards or contingencies via world models to ensure the survival of organisms [17].

Consequently, motivated by the aforementioned insights, in this work, we hope to establish linkages between AI, neuroscience and psychology and explore a novel RL framework by modelling the amygdala functionality of the brain to further advance safe decision making for autonomous vehicles. More specifically, building upon the current computational framework for the dopamine-striatum mechanism, we present a fear-neuro-inspired RL (FNI-RL) technique to model the process of RL in the brain by considering the amygdala functionality, enabling the autonomous driving agent to learn defensive behaviors effectively. We encourage the agent to undertake risky explorations within its own imagination through a model-based setting, while executing safe decisions during interactions with the real environment to the greatest extent possible.

An overview of the proposed approach is illustrated in Fig. 1. In light of the RL-related functional systems in the brain, we first present an adversarial imagination mechanism to simulate safety-critical situations with a learnable adversary and world model, facilitating the agent to cope with unseen hazardous scenarios and enhance policy robustness against uncertainties and nonstationarities. Concretely, we leverage a mixed policy comprising both the agent and the adversary to interact with the learned world model, where the agent seeks to keep its fear within specified bounds while the adversary aims to maximize the agent's fear. Here a fear model is constructed to estimate the fear of the agent in response to the recognition of dangers or contingencies. Based on the findings in neuroscience [17], [18], our fear model incorporates both negative stimuli (e.g., safety violations) and environmental uncertainties. Additionally, we develop a fear-constrained actor-critic (FC-AC) algorithm that enables the agent to learn defensive driving behaviors and ensure safe decision making, via effectively assessing unsafe policy trajectories and adhering to the imposed fear constraints.

Compared with existing studies, the main contributions of this work are summarized as follows[1]. (1) Drawing inspiration from the fear neurons in the brain, we present a computational FNI-RL framework to enhance the safety of autonomous vehicles. (2) An adversarial imagination technique is advanced to simulate safety-critical situations, which facilitates the agent to tackle unseen risky scenarios and improve the policy robustness against uncertainties and nonstationarities. Here a fear model is devised to recognize and estimate dangers and contingencies. (3) An FC-AC algorithm is developed to enable the agent to learn defensive driving behaviors and realize safe decision making with fewer safety violations.

We demonstrate the feasibility and effectiveness of the proposed FNI-RL approach for safe autonomous driving in comparison with state-of-the-art AI agents and 30 certified human drivers. The simulation tests are performed based on the simulation of urban mobility (SUMO) package [19]. In addition, experimental evaluations are also carried out in three critical situations on a human-in-the-loop test platform (Fig. 4B) with a high-fidelity driving simulator, Car Learning to Act (CARLA) [20]. The results indicate that, enhanced by the developed FNI-RL algorithm, the autonomous driving agent can generate defensive decision making behaviors,

1. The code and supplementary video are available at https://github.com/TMIS-Turbo/FNI-RL

thereby significantly improving safety and achieving human drivers' performance in various safety-critical scenarios.

2 Related Work

2.1 Safe Autonomous Driving

In recent years, researchers have endeavored to enhance the safety of autonomous vehicles by various perspectives or methods [21], [22], such as generating diverse driving scenarios [5], [6], imitating human driver behaviors [23], [24], learning safe and robust policies via RL [25], [26].

In [27], a scheme called AdvSim is presented for generating safety-critical scenarios. AdvSim optimizes the vehicle trajectories jointly to perturb the driving paths of surrounding vehicles. Moreover, incorporating AdvSim-generated safety-critical scenarios in training can benefit the safety of autonomous vehicles. In [28], a technique named STRIVE is introduced, which utilizes a graph-based conditional variational autoencoder (CVAE) model to automatically generate challenging scenarios. Here the scenarios generated by STRIVE can be employed to to optimize hyperparameters of a rule-based planner. In [29], a gradient-based scenario generation method called KING is proposed, which utilizes a kinematic motion model to guide the generation of adversarial scenarios. Additionally, the safety of autonomous driving can be enhanced by augmenting the training data with the generated scenarios from KING. However, these methods rely on pre-collected datasets to learn traffic priors. Furthermore, they do not optimize driving policies by integrating generated safety-critical scenarios with RL. In [30], a causal generative model is devised to generate safety-critical scenarios through causal graphs derived from human priors. The authors also empirically demonstrate that incorporating the generated scenarios as additional training samples can enhance the performance of RL-based driving policies. Nevertheless, this technique depends heavily on human priors. In contrast, our FNI-RL approach for learning safe autonomous driving policies does not rely on any pre-collected datasets or human priors. In addition, unlike the aforementioned methods, FNI-RL optimizes both driving policies and the adversarial sample generation module simultaneously in an online learning manner, as the RL agent interacts with the real environment.

An imitation learning (IL) technique with on-policy RL supervisions is developed to enhance the performance of autonomous vehicles in [31]. A human-in-the-loop learning scheme called human-AI copilot optimization is advanced to facilitate the learning of safe driving policies in [32]. This approach integrates interventions from human experts into the interaction between the agent and the environment to guarantee both efficient and safe exploration. Furthermore, some researchers have employed RL methods with safety constraints based on prior knowledge [33] or rules [34] to optimize driving policies while simultaneously guaranteeing the satisfaction of the imposed constraints. In [35], the authors present a constrained adversarial RL algorithm that aims to realize safe autonomous driving from the perspective of robust decision making. While these approaches can effectively improve the safety of autonomous vehicles, they either heavily rely on pre-collected datasets or human priors, or they have to go through a substantial number of safety violations to learn safe driving policies. In contrast, the proposed FNI-RL approach allows the agent to acquire safe driving skills with fewer safety violations, without the requirement for pre-collected datasets or human priors.

2.2 Safe Model-Free Reinforcement Learning

A popular class of safe model-free RL (SMFRL) methods is dedicated to solving the constrained Markov decision process (CMDP) to ensure the acquisition of safe policies [36]. These studies extensively combine model-free RL framework with Lagrangian methods to restrict the cost value of the policy below a predetermined threshold [37]. In the latter case, the policies and Lagrangian multipliers are optimized iteratively via the dual theory [38]. There are also SMFRL algorithms that incorporate reachability analysis [39], [40] or expert information [41], [42]. For instance, in [41], a SMFRL framework with prior knowledge is developed to ensure safe exploration. Although the above methods have achieved many competitive results, they either suffer from a large number of unsafe interactions during training or heavily depend on human priors. In contrast, FNI-RL does not require any prior knowledge and enables the agent to learn safe driving skills with fewer safety violations.

2.3 Safe Model-Based Reinforcement Learning

In safe model-based RL (SMBRL), apart from learning a policy model, an additional environment model is required to be learned, which can be leveraged to generate possible trajectories or evaluate the safety of actions before executing them in the real environment [43], [44], [45]. By incorporating cost constraints throughout the learning process, SMBRL methods have the potential to prevent dangerous exploration behaviors while ensuring sample efficiency [46], [47], [48]. For example, in [45], a SMBRL scheme is proposed to minimize safety violations during training. This method involves learning an ensemble of probabilistic dynamics models to plan ahead a short time into the future and applies heavy penalties to unsafe trajectories. In [47], a SMBRL technique is introduced to cope with safety-critical tasks, which adopts the learned Bayesian world model to generate trajectories and estimate an optimistic bound for the task objective and pessimistic bounds for the constraints. Then, the augmented Lagrangian approach is employed to solve the constrained optimization problem with the estimated bounds. In [48], a SMBRL algorithm is developed with a Lagrangian relaxation-based proximal policy optimization technique and an ensemble of environment model. In this framework, both epistemic and aleatoric uncertainties are simultaneously taken into account during the learning of the dynamics models. Unlike the methods mentioned above, drawing inspiration from the fear neurons in the brain, FNI-RL incorporates the adversarial imagination technique that can simulate safety-critical situations via the learned adversary and world model, assisting the agent in handling unseen risky scenarios and enhancing policy robustness against uncertainties and nonstationarities. Additionally, in FNI-RL, the agent is required to comply with the fear constraint that encompasses the dangers and uncertainties estimated by the adversarial imagination.

3 METHODOLOGY

The proposed FNI-RL framework for the safe decision making of autonomous vehicles is mainly composed of the adversarial imagination technique and the FC-AC algorithm. The framework of our approach is illustrated in Fig. 1.

3.1 Adversarial Imagination

We develop the adversarial imagination technique by combining the adversarial agent with the world model to simulate the worst-case situations in the imagination, enabling our autonomous driving agent to tackle unseen critical scenarios and improve policy robustness. Here a mixed policy $\pi^{\text{mix}}(\cdot)$ is defined as:

$$\pi^{\text{mix}}(\cdot|s) \equiv \alpha \cdot \pi(\cdot|s;\theta) + (1-\alpha)\cdot \bar{\pi}(\cdot|s;\bar{\theta}), \tag{1}$$

where α is a weight between 0 and 1, $\pi(\cdot)$ and $\bar{\pi}(\cdot)$ represent the stochastic policies of the protagonist and the adversary, θ and $\bar{\theta}$ are the parameters of the policy network and the adversarial policy network, and s denotes the state of the agent, respectively. An action perturbed by the adversary, denoted as $\tilde{a}$, can be sampled from the mixed policy, i.e., $\tilde{a} \sim \pi^{\text{mix}}(\cdot|s)$. The protagonist endeavors to optimize the expected return while ensuring that its fear remains within predefined bounds. Conversely, the adversary aims to maximize the protagonist's fear.

In organisms, fear can be elicited by certain negative stimuli [17]. For instance, watching or experiencing a frightening traumatic accident is capable of arousing fear in humans. In RL, the reward function serves as an incentive used to evaluate the behaviors of the agent. Similarly, in constrained RL [36], we can view the cost function as a form of negative stimulus, such as collisions. Furthermore, fear can also be caused by uncertainties [49], [50]. For example, a human being may feel fear in an uncertain environment. Consequently, we construct the fear model to incorporate both the anticipated negative stimuli and epistemic uncertainties simultaneously, and it can be expressed as follows:

$$f(s,\tilde{a}) = \beta \cdot \hat{c}(s,\tilde{a}) + (1-\beta)\cdot \hat{\sigma}(s,\tilde{a}), \tag{2}$$

where β represents a weight that ranges from 0 to 1. $\hat{c}(\cdot)$ and $\hat{\sigma}(\cdot)$ denote the cost function and epistemic uncertainty estimated via the world model, respectively. From Eq. (2), the higher estimated cost and uncertainty will arouse a more intense fear in the agent. $\underline{f}$ and $\bar{f}$ denote the lower and upper bounds of the fear, respectively. In our setting, we utilize the probability of safety violations as the cost function, i.e., $\hat{c}(\cdot) \in [0,1]$. Moreover, the minimum of $\hat{\sigma}(\cdot)$ is equal to zero. We constrain the maximum of $\hat{\sigma}(\cdot)$ as 1. Consequently, we can draw the following conclusion: $\underline{f} = 0$ and $\bar{f} = 1$, namely, $f(\cdot) \in [0,1]$.

The world model aims to provide an internal representation of the contingencies of the real environment. Here, we leverage an ensemble of diagonal Gaussian world models to effectively acquire both aleatoric and epistemic uncertainties [45], [51]. This ensemble can be denoted as $\{\hat{T}_{\phi_k}\}_{k=1}^{K}$, where $\hat{T}_{\phi_k}(s',c|s,a) = \mathcal{N}\big(\mu_{\phi_k}(s,a), \sigma^2_{\phi_k}(s,a)\big)$. s' and K are the next state and the number of the world models, respectively. Moreover, $\mu_{\phi_k}(\cdot)$ and $\sigma_{\phi_k}(\cdot)$ represent the mean and standard deviation of the Gaussian distribution $\mathcal{N}(\cdot)$ parameterized by ϕ_k, respectively. In contrast to the majority of existing environmental models, our world model predicts a cost c rather than a reward r. For the kth world model, it can be trained by minimizing the following objective function based on negative log-likelihood:

$$J_w(\phi_k) = -\mathop{\mathbb{E}}_{(s,a,c,s')\sim\mathcal{M}}[\log \hat{T}_{\phi_k}(s',c|s,a)], \tag{3}$$

where $\mathcal{M}$ denotes an experience replay memory. Random differences in initialization and mini-batch paradigm during training give rise to distinct models. The model ensemble is able to be employed to produce predictions incorporating uncertainties. By combining the ensemble with the mixed policy, the set-valued cost and uncertainty can be obtained:

$$(\hat{s}',\hat{c}) \sim \frac{1}{K}\sum_{k=1}^{K}\hat{T}_{\phi_k}(s,\tilde{a}), \;\; \hat{\sigma} = \frac{1}{K}\sum_{k=1}^{K}\sigma_{\phi_k}(s,\tilde{a}), \tag{4}$$

where $\hat{s}$ and $\hat{s}'$ represent the state and next state estimated by the world model, respectively. With a short prediction horizon m, the fear of the agent can be denoted as:

$$f(\hat{s}^m,\tilde{a}^m) = \beta\cdot\hat{c}(\hat{s}^m,\tilde{a}^m) + (1-\beta)\cdot\hat{\sigma}(\hat{s}^m,\tilde{a}^m), \tag{5}$$

where $\hat{s}^m$ and $\tilde{a}^m$ represent the state and action obtained after m steps of forward planning based on the world model and mixed policy, respectively. We collect the generated virtual transitions into a virtual experience replay memory $\hat{\mathcal{M}}$, enhancing the performance of the agent. Additionally, the adversary model can be learned by maximizing the following objective function:

$$J_{\bar{a}}(\bar{\theta}) = \mathop{\mathbb{E}}_{s\sim\mathcal{M}}[f(\hat{s}^m,\tilde{a}^m)]. \tag{6}$$

3.2 Fear-Constrained Actor-Critic

In this section, the proposed FC-AC algorithm is introduced to optimize the driving policies of our agent while keeping its fear within preset bounds.

A CMDP is an augmentation of a Markov Decision Process (MDP) by incorporating a cost function, which can be represented by a 6-tuple $\langle \mathcal{S}, \mathcal{A}, p, r, c, \gamma\rangle$. $\mathcal{S}$ is the set of states called the state space. $\mathcal{A}$ is the set of actions called the action space. p is the transition probability distribution. $r : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ denotes the reward function, and $c : \mathcal{S}\times\mathcal{A}\to\mathbb{R}$ represents the cost function. $\gamma \in (0,1)$ is the discount factor.

According to CMDP, FC-AC seeks to solve the following constrained optimization problem:

$$\max_{\pi} \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t)\right], \;\; \text{s.t.} \;\; \mathbb{E}\left[f(\hat{s}^m,\tilde{a}^m)\right] \le f_0, \tag{7}$$

where t is the time step, and f_0 is a prescribed threshold.

A policy iteration algorithm, named fear-constrained policy iteration (FC-PI), is developed to approximate the optimal policies. The FC-PI method comprises two learning processes: policy evaluation and policy improvement. These two processes are updated alternately until the policy converges. FC-PI can provably converge to the optimal policy (see the supplementary). Moreover, the Lagrangian of the constrained optimization problem can be written as:

$$L(\pi,\lambda) = \mathbb{E}\left[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t) + \lambda\big(f_0 - f(\hat{s}^m,\tilde{a}^m)\big)\right], \tag{8}$$

where λ denotes the dual variable, and $\lambda \ge 0$.

3.2.1 Fear-Constrained Policy Evaluation

The action-value function $Q^\pi(s,a)$ can be iteratively computed under the fixed policies of the agent via a Bellman backup operator $\mathcal{T}$:

$$\mathcal{T}Q^\pi(s,a) \equiv r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim p}[V^\pi(s')], \tag{9}$$

where $V^\pi(\cdot)$ denotes a value function, and it is designed as:

$$V^\pi(s) = \mathop{\mathbb{E}}_{a\sim\pi(\cdot|s)}[Q^\pi(s,a) - \lambda f(\hat{s}^m, \tilde{a}^m)]. \tag{10}$$

The FC-AC algorithm employs two parameterized action-value functions with network parameters ϕ^z, $z \in \{1,2\}$ to speed up the model training process [52]. The parameters of the action-value function can be learned by minimizing the following loss function of the critic network:

$$J_c(\phi^z) = \mathop{\mathbb{E}}_{(s,a,r,s')\sim\mathcal{M}}\left[\|y - Q^\pi(s,a;\phi^z)\|_2^2\right], \tag{11}$$

where y denotes a target value. According to the results in [53] and our empirical findings, the training of the action-value function network requires relatively high data quality. Therefore, we only employ real interaction data to train the action-value function network, reducing the reliance on the accuracy of the world model.

To ensure safety, it is imperative to guarantee that the Q-values of actions causing unsafe states are lower than the Q-values of safe actions. We follow the assumption regarding the existence of a special horizon H in [45]. According to this assumption, after the agent completes H steps of safe interaction with the environment, it will inevitably transition into an unsafe state (i.e., with a safety violation). Then, the agent can no longer recover to the safe state (i.e., without a safety violation).

In theory, we can devise a specific cost c^* as a penalty of the agent for safety violations to avoid the occurrence of the hazardous situation described in the above assumption. Under the given assumption, the maximum of the infinite-horizon discounted return with the agent's fear is as follows:

$$\sum_{t=0}^{H-1}\gamma^t\bar{r} + \sum_{t=H}^{\infty}\gamma^t(-\underline{c}^*) - \lambda\underline{f} = \frac{\bar{r}(1-\gamma^H) - \underline{c}^*\gamma^H}{1-\gamma} - \lambda\underline{f}, \tag{12}$$

where $\bar{r}$ is the upper bound of the reward r, and $\underline{c}^*$ denotes the lower bound of the cost c^*. In contrast, in the absence of any safety violations, the minimum of the infinite-horizon discounted return considering the fear is as follows:

$$\sum_{t=0}^{\infty}\gamma^t\underline{r} - \lambda\bar{f} = \frac{\underline{r}}{1-\gamma} - \lambda\bar{f}, \tag{13}$$

where $\underline{r}$ represents the the lower bound of the reward r.

To ensure a reasonable evaluation of the safety of decisions, it is desirable for the following inequality to hold:

$$\frac{\bar{r}(1-\gamma^H) - \underline{c}^*\gamma^H}{1-\gamma} - \lambda\underline{f} < \frac{\underline{r}}{1-\gamma} - \lambda\bar{f}. \tag{14}$$

With Eq. (14), we can derive the following conclusion:

$$\underline{c}^* > \frac{\bar{r} - \underline{r} + \lambda(1-\gamma)(\bar{f} - \underline{f})}{\gamma^H} - \bar{r}. \tag{15}$$

Since $\bar{f}$ and $\underline{f}$ are bounded, and to satisfy the above inequality, we can design the cost c^* as:

$$c^* = \frac{\bar{r} - \underline{r}}{\gamma^H(1-\gamma)} + \frac{\lambda}{\gamma^H} - \frac{\bar{r}}{1-\gamma}. \tag{16}$$

To prevent overestimation in the action-value function, the minimum estimation among the two target parameterized action-value functions is leveraged to train the critic network. Hence, y can be devised as:

$$y = \begin{cases} r + \gamma\tilde{Q}^\pi(s',a'), & \text{without a safety violation,} \\ -c^*, & \text{with a safety violation,} \end{cases} \tag{17}$$

where $\hat{Q}^\pi(\cdot)$ represents a target action-value function, $\tilde{Q}^\pi(s,a) = \min_{z\in\{1,2\}} \hat{Q}^\pi(s,a;\bar{\phi}^z) - \lambda f(\hat{s}^m, \tilde{a}^m)$, $\bar{\phi}^z$ denotes the network parameter of the target action-value function, $z \in \{1,2\}$, and $a' \sim \pi(\cdot|s')$.

The network parameters $\bar{\phi}^z$ of the target action-value function can be updated by Polyak averaging: $\bar{\phi}^z \leftarrow \rho\bar{\phi}^z + (1-\rho)\phi^z$, where ρ is a scale coefficient between 0 and 1.

3.2.2 Fear-Constrained Policy Improvement

In FC-PI, the policy improvement aims to maximize the expected return while adhering to the fear constraint.

According to Lagrange duality theory and Eq. (8), the Lagrange dual problem associated with the constrained optimization problem in Eq. (7) can be derived as:

$$\min_\lambda \max_\pi L(\pi,\lambda) = \min_\lambda \max_\pi \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t r(s_t,a_t) + \lambda\big(f_0 - f(\hat{s}^m,\tilde{a}^m)\big)\Big]. \tag{18}$$

In order to effectively tackle unseen safety-critical scenarios and enhance the policy diversity, we optimize the policy of the agent using data from both virtual and real experience replay memories. Hence, the optimal policy of the agent can be approximated by maximizing the following objective function for the actor network:

$$J_a(\theta) = \mathop{\mathbb{E}}_{s\sim\mathcal{M}\cup\hat{\mathcal{M}},a\sim\pi(\cdot|s;\theta)}[Q^\pi(s,a) - \lambda f(\hat{s}^m,\tilde{a}^m)]. \tag{19}$$

Additionally, the dual variable λ can be updated by minimizing the following objective function:

$$J_d(\lambda) = \mathop{\mathbb{E}}_{s\sim\mathcal{M}\cup\hat{\mathcal{M}}}\big[\lambda\big(f_0 - f(\hat{s}^m,\tilde{a}^m)\big)\big]. \tag{20}$$

In our setting, the cost $\hat{c}$ returned by the world model represents the probability of a safety violation. Hence, during the model testing phase, to further diminish the risk, the agent can assess the safety of decisions using the learned world model. For instance, in Fig. 1, if the agent's action is evaluated by the world model as having a high collision risk, then a Gaussian noise ϵ will be added to this action.

4 Results

To benchmark FNI-RL, we set up experimental comparisons with state-of-the-art AI agents and certified human drivers in complex and critical traffic scenes.

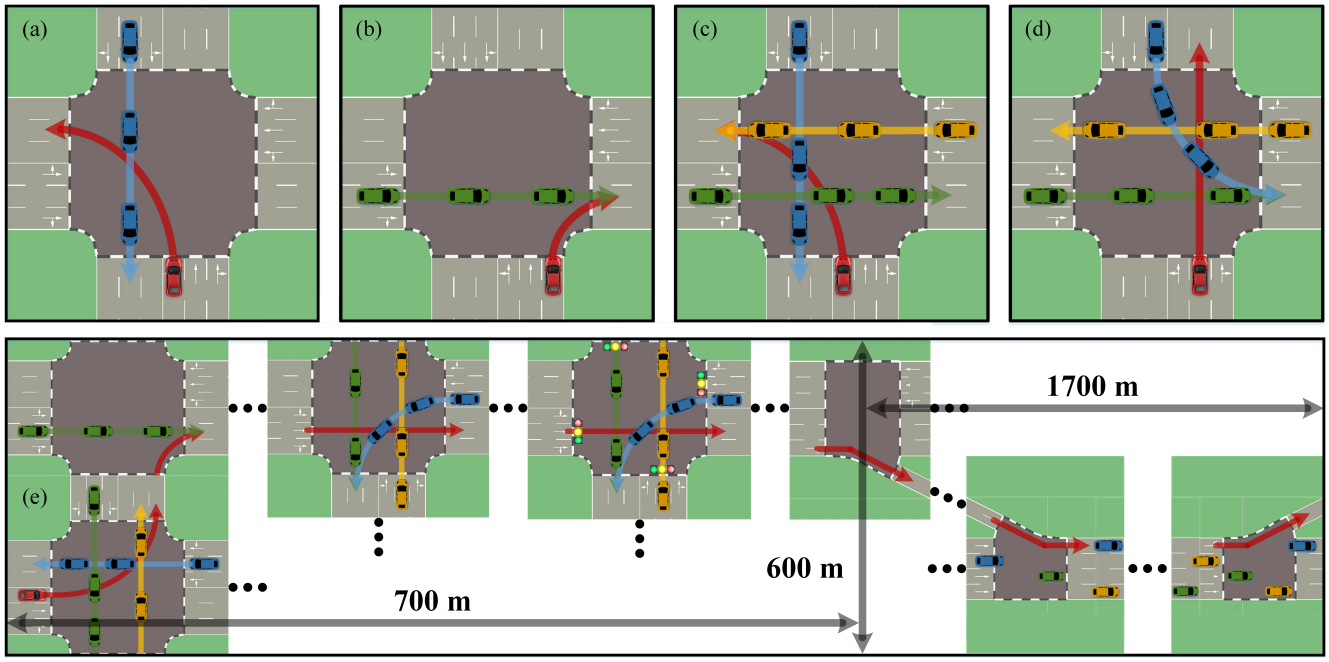


Fig. 2. Experimental traffic environments. (a) Unprotected left turn at an unsignalized intersection with oncoming traffic. (b) Right turn at an unsignalized intersection with crossing traffic. (c) Unprotected left turn at an unsignalized intersection with mixed traffic flows. (d) Crossing negotiation at an unsignalized intersection with mixed traffic flows. (e) Long-term goal-driven navigation with mixed traffic flows.

4.1 Baselines

Rule-based driver: An intelligent driver model (IDM) in SUMO is leveraged as a rule-based baseline.

Vanilla RL: We employ proximal policy optimization (PPO) [54] and soft actor-critic (SAC) [55] as two vanilla RL baselines, representing on-policy and off-policy methods.

SMFRL: Constraint policy optimization (CPO) [36] and SAC-Lagrangian (SAC-Lag) [38] are adopted as two SMFRL baselines.

SMBRL: We utilize safe model-based policy optimization (SMBPO) [45] and safe model-based PPO (SMBPPO) [48] as two SMBRL baselines.

IL: Generative adversarial imitation learning (GAIL) [56] and RL coach (Roach) [31] are employed as two IL baselines. We utilize the next generation simulation (NGSIM) dataset [57] along with the behavior cloning (BC) technique to train a policy model as the initial model for the two IL baselines. This ensures that the IL agents possess basic driving skills right from the start of the training phase. Furthermore, during the training process, the GAIL agent learn expert behaviors by leveraging the demonstration data from IDM.

Human driver: We recruit 30 human participants for the experiments, all of whom hold valid driving licenses.

4.2 Metrics

To assess the overall driving quality, we introduce a driving score (DS) defined as follows:

$$\mathrm{DS} = \eta \cdot \mathrm{SR} + (1-\eta) \cdot \frac{v}{v_{\max}}, \tag{21}$$

where SR is a success rate, v and $v_{\max}$ denote the agent's speed and the permissible maximum speed. The weight η is set to 0.8. Successful driving here refers to the vehicle's ability to reach the target lane without any safety violations including collisions and running a red light. Obviously, DS $\in [0, 1]$. In the scenarios (a)-(d) depicted in Fig. 2, a safety violation rate (SVR) denotes a collision rate (CR). In the scenario (e), SVR includes not only CR but also a red-light violation rate (RVR). Furthermore, the training-time safety is measured by the total number of safety violations (TNSV) in the training.

In the human-in-the-loop experiment, apart from SR, a time-to-collision (TTC) metric is utilized to evaluate potential collision risks or driving safety. The acceleration of the ego vehicle is utilized as a metric to measure driving smoothness and comfort. Additionally, the acceleration of the following vehicle is leveraged to analyze the influence of the ego vehicle's driving behaviors on surrounding traffic.

4.3 General Settings

All agents are trained for 2000 episodes in SUMO using five different random seeds. Except for the navigation task, where each episode includes a maximum of 300 time steps, all other tasks have episodes with a maximum of 30 time steps. For a comprehensive evaluation, we set up three traffic flows with different densities, namely flow-0, flow-1, and flow-2. In the flow-0, flow-1 and flow-2, the probabilities of emitting a vehicle each second are set to 0.5, 0.3 and 0.7, respectively. All agents are trained in the flow-0, while the flow-1 and flow-2 are solely leveraged for testing. During the model testing phase, we evaluate the final policy models trained with all the algorithms and different random seeds. All the methods utilize the same policy network configuration. For further details such as reward function and hyperparameters, please refer to the supplementary.

TABLE 1
Statistical results of different autonomous driving agents in the traffic scenarios (a)-(d), including the mean and standard deviation (in brackets).

Method	Metric	Scenario (a)		Scenario (b)		Scenario (c)		Scenario (d)		Summary
		Flow-1	Flow-2	Flow-1	Flow-2	Flow-1	Flow-2	Flow-1	Flow-2	
IDM	DS	0.99 (0.00)	0.98 (0.05)	**0.98 (0.00)**	**0.98 (0.00)**	0.88 (0.17)	0.86 (0.20)	0.99 (0.04)	0.98 (0.04)	0.96 (0.05)
	SR	**1.00 (0.00)**	0.99 (0.06)	**1.00 (0.00)**	**1.00 (0.00)**	0.87 (0.20)	0.85 (0.25)	0.99 (0.04)	0.98 (0.05)	0.96 (0.06)
	CR	**0.00 (0.00)**	0.01 (0.06)	**0.00 (0.00)**	**0.00 (0.00)**	0.13 (0.20)	0.15 (0.25)	0.01 (0.04)	0.02 (0.05)	0.04 (0.06)
PPO	DS	0.86 (0.07)	0.85 (0.06)	0.70 (0.16)	0.61 (0.19)	0.56 (0.20)	0.40 (0.30)	0.94 (0.10)	0.70 (0.31)	0.70 (0.17)
	SR	0.97 (0.09)	0.97 (0.08)	0.68 (0.20)	0.57 (0.23)	0.55 (0.23)	0.35 (0.32)	0.93 (0.12)	0.68 (0.34)	0.71 (0.21)
	CR	0.03 (0.09)	0.03 (0.08)	0.32 (0.20)	0.43 (0.23)	0.10 (0.10)	0.27 (0.19)	0.07 (0.12)	0.10 (0.13)	0.17 (0.14)
SAC	DS	0.85 (0.05)	0.84 (0.06)	0.94 (0.04)	0.93 (0.04)	0.92 (0.06)	0.91 (0.06)	0.75 (0.08)	0.73 (0.09)	0.86 (0.08)
	SR	0.97 (0.06)	0.96 (0.07)	0.98 (0.05)	0.98 (0.05)	0.96 (0.07)	0.95 (0.08)	0.89 (0.10)	0.87 (0.11)	0.95 (0.04)
	CR	**0.00 (0.00)**	0.01 (0.03)	0.01 (0.04)	0.01 (0.03)	0.03 (0.05)	0.04 (0.07)	0.08 (0.08)	0.09 (0.10)	0.03 (0.03)
CPO	DS	0.99 (0.03)	0.98 (0.04)	0.83 (0.12)	0.81 (0.10)	0.83 (0.10)	0.75 (0.10)	0.94 (0.06)	0.92 (0.07)	0.88 (0.08)
	SR	0.99 (0.04)	0.97 (0.04)	0.81 (0.15)	0.79 (0.12)	0.79 (0.13)	0.75 (0.12)	0.94 (0.07)	0.90 (0.09)	0.87 (0.09)
	CR	0.01 (0.04)	0.03 (0.04)	0.19 (0.15)	0.21 (0.12)	0.21 (0.13)	0.25 (0.12)	0.06 (0.07)	0.10 (0.09)	0.13 (0.09)
SAC-Lag	DS	0.95 (0.03)	0.92 (0.05)	0.94 (0.03)	0.93 (0.04)	0.89 (0.07)	0.88 (0.08)	0.85 (0.08)	0.85 (0.09)	0.90 (0.04)
	SR	0.99 (0.04)	0.96 (0.06)	0.99 (0.03)	0.97 (0.05)	0.93 (0.08)	0.91 (0.10)	0.88 (0.10)	0.88 (0.11)	0.94 (0.04)
	CR	**0.00 (0.02)**	0.01 (0.02)	**0.00 (0.02)**	0.02 (0.04)	0.04 (0.07)	0.05 (0.07)	0.07 (0.08)	0.07 (0.08)	0.03 (0.03)
SMBPO	DS	0.98 (0.03)	0.98 (0.03)	0.71 (0.18)	0.65 (0.19)	0.96 (0.04)	0.95 (0.05)	0.97 (0.10)	0.94 (0.06)	0.89 (0.12)
	SR	0.99 (0.03)	0.98 (0.04)	0.72 (0.18)	0.66 (0.19)	0.98 (0.05)	0.96 (0.06)	0.97 (0.10)	0.94 (0.07)	0.90 (0.12)
	CR	0.01 (0.02)	0.01 (0.03)	0.01 (0.04)	0.03 (0.06)	0.02 (0.05)	0.04 (0.06)	0.02 (0.04)	0.05 (0.07)	0.02 (0.01)
SMBPPO	DS	0.98 (0.03)	0.98 (0.04)	0.92 (0.09)	0.83 (0.08)	0.94 (0.12)	0.78 (0.12)	0.95 (0.05)	0.92 (0.08)	0.91 (0.07)
	SR	0.98 (0.04)	0.98 (0.05)	0.94 (0.12)	0.88 (0.10)	0.81 (0.12)	0.76 (0.15)	0.95 (0.06)	0.91 (0.09)	0.90 (0.08)
	CR	0.02 (0.04)	0.02 (0.05)	0.06 (0.12)	0.11 (0.10)	0.19 (0.12)	0.24 (0.14)	0.05 (0.06)	0.09 (0.09)	0.10 (0.08)
GAIL	DS	0.92 (0.06)	0.91 (0.09)	0.75 (0.15)	0.51 (0.13)	0.70 (0.11)	0.66 (0.15)	0.83 (0.09)	0.56 (0.15)	0.73 (0.14)
	SR	0.96 (0.07)	0.95 (0.10)	0.73 (0.19)	0.48 (0.17)	0.72 (0.13)	0.67 (0.17)	0.85 (0.11)	0.55 (0.18)	0.74 (0.16)
	CR	0.01 (0.02)	0.01 (0.02)	0.25 (0.18)	0.38 (0.16)	0.09 (0.10)	0.07 (0.07)	0.10 (0.09)	0.13 (0.11)	0.13 (0.12)
Roach	DS	0.98 (0.03)	0.96 (0.04)	0.94 (0.03)	0.80 (0.11)	0.89 (0.07)	0.84 (0.09)	0.42 (0.17)	0.15 (0.08)	0.75 (0.28)
	SR	0.99 (0.03)	0.98 (0.05)	0.98 (0.04)	0.83 (0.14)	0.90 (0.09)	0.88 (0.11)	0.40 (0.21)	0.10 (0.09)	0.76 (0.31)
	CR	0.01 (0.03)	0.01 (0.04)	0.02 (0.04)	0.16 (0.13)	0.09 (0.09)	0.07 (0.08)	0.07 (0.08)	0.02 (0.03)	0.06 (0.05)
FNI-RL	DS	**1.00 (0.00)**	**0.99 (0.01)**	**0.98 (0.00)**	0.97 (0.01)	**0.97 (0.00)**	**0.97 (0.00)**	**1.00 (0.00)**	**0.99 (0.02)**	**0.98 (0.01)**
	SR	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.00)**	**1.00 (0.02)**	**1.00 (0.00)**
	CR	**0.00 (0.00)**	**0.00 (0.00)**	**0.00 (0.00)**	**0.00 (0.00)**	**0.00 (0.00)**	**0.00 (0.00)**	**0.00 (0.00)**	**0.00 (0.02)**	**0.00 (0.00)**

4.4 Traffic Negotiation at Unsignalized Intersections

Task. In the scenario (a) depicted in Fig. 2, the ego vehicle (i.e., the red-colored vehicle) is executing an unprotected left turn at an unsignalized intersection while interacting with an oncoming dynamic traffic flow. In the scenario (b), the ego vehicle is carrying out a right turn at an unsignalized intersection while interacting with a crossing dynamic traffic flow. In the scenario (c), the ego vehicle is performing an unprotected left turn at an unsignalized intersection while interacting with an oncoming dynamic traffic flow and two crossing dynamic traffic flows. In the scenario (d), the ego vehicle is required to negotiate with an oncoming dynamic traffic flow and two crossing dynamic traffic flows in order to cross an unsignalized intersection.

State and action. We adopt the information from the 6 nearest vehicles within a 200-meter distance from the ego vehicle, encompassing the relative distance, orientation, speed, and velocity direction of the front, back, left-front, left-back, right-front, and right-back vehicles. Moreover, we incorporate the speed and velocity direction of the ego vehicle, resulting in a state representation of the agent with a total of 26 dimensions. Here, the action of agents is continuous longitudinal acceleration or deceleration.

Evaluation. Here, we assess and compare the performance of FNI-RL against 9 baselines. We test each agent on each situation for 500 episodes and report the average metrics in Table 1. Overall, FNI-RL performs consistently across all tasks, surpassing the rule-based, vanilla RL, SMFRL, SMBRL, and IL baselines in terms of comprehensive performance and safety in the majority of test cases. For instance, compared with the IDM, PPO, SAC, CPO, SAC-Lag, SMBPO, SMBPPO, GAIL and Roach agents, FNI-RL gains approximately 1.55%, 16.57%, 18.64%, 1.94%, 7.98%, 1.55%, 1.78%, 9.33% and 3.24% improvements with respect to DS in the scenario (a) with flow-2, respectively. In the scenario (b), FNI-RL performs comparably to IDM and outperforms all other baselines with a large margin, in terms of DS, SR and CR. In scenario (c) with flow-2, FNI-RL surpasses the IDM, PPO, SAC, CPO, SAC-Lag, SMBPO, SMBPPO, GAIL, and Roach agents, exhibiting DS improvements of approximately 12.74%, 144.59%, 6.39%, 29.50%, 10.71%, 2.33%, 24.55%, 46.45% and 15.41% respectively, along with SR enhancements of approximately 18.20%, 182.49%, 4.82%, 32.63%, 9.41%, 3.95%, 31.93%, 48.81% and 13.90%. Moreover, in the scenario (d) with flow-1, compared with IDM, PPO, SAC, CPO, SAC-Lag, SMBPO, SMBPPO, GAIL and Roach, the SR metric of FNI-RL is enhanced by approximately 1.22%, 7.76%, 12.87%, 6.84%, 13.90%, 3.31%, 5.49% 17.37%, and 152.53%, respectively. Agents generally perform better in flow-1 than in flow-2. While certain baselines, such as

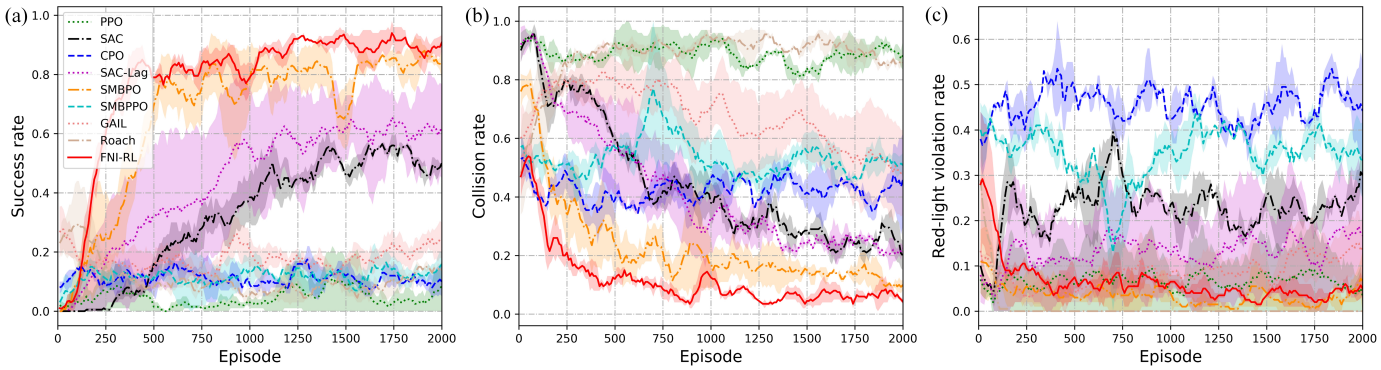


Fig. 3. The training performance of the different autonomous driving agents on the long-term goal-driven navigation task based on the stochastic dynamic traffic flows. (a) Success rate. (b) Collision rate. (c) Red-light violation rate.

TABLE 2
Assessment results of the rule-based and learning-based autonomous driving agents in the long-term goal-driven navigation benchmark.

Metric	IDM	PPO	SAC	CPO	SAC-Lag	SMBPO	SMBPPO	GAIL	Roach	FNI-RL
DS	0.47 (0.13)	0.17 (0.11)	0.51 (0.13)	0.18 (0.08)	0.60 (0.17)	0.73 (0.13)	0.19 (0.08)	0.26 (0.11)	0.25 (0.08)	**0.84 (0.08)**
SR	0.46 (0.16)	0.06 (0.10)	0.49 (0.16)	0.11 (0.10)	0.60 (0.21)	0.81 (0.16)	0.13 (0.10)	0.18 (0.14)	0.09 (0.10)	**0.90 (0.10)**
CR	0.54 (0.16)	0.88 (0.11)	0.29 (0.15)	0.43 (0.16)	0.27 (0.15)	0.15 (0.11)	0.51 (0.14)	0.61 (0.28)	0.91 (0.10)	**0.06 (0.08)**
RVR	**0.00 (0.00)**	0.07 (0.10)	0.23 (0.13)	0.45 (0.16)	0.14 (0.15)	0.02 (0.06)	0.37 (0.14)	0.10 (0.12)	**0.00 (0.00)**	0.04 (0.06)
TNSV ($\times 10^2$)	N/A	19.09 (0.85)	13.51 (0.52)	17.58 (0.32)	11.11 (3.38)	5.33 (1.51)	17.65 (0.76)	15.33 (2.02)	17.80 (0.48)	**3.57 (0.14)**

IDM, SAC, SAC-Lag, SMBPO, and SMBPPO, may exhibit comparable performance to FNI-RL in several situations, they do not achieve an equivalent level of safety. This distinction underscores the primary contribution of this work.

Additionally, in Table 1, we present summary statistics that assess the average performance of each method across all testing conditions. For instance, according to the average DS metric in the last column of Table 1, in contrast to the IDM, PPO, SAC, CPO, SAC-Lag, SMBPO, SMBPPO, GAIL and Roach agents, FNI-RL gains approximately 2.08%, 40.00%, 13.95%, 11.36%, 8.89%, 10.11%, 7.69%, 34.25% and 30.67% improvements with respect to DS, respectively. We find that the rule-based IDM agent exhibits strong competitiveness. Specifically, FNI-RL performs comparably to IDM on the easier tasks and surpasses IDM on the more challenging tasks in terms of the overall driving performance.

4.5 Long-Term Goal-Driven Navigation

Task. In the scenario (e) of Fig. 2, the ego vehicle first executes an unprotected left turn at an unsignalized intersection while interacting with an oncoming dynamic traffic flow and two crossing dynamic traffic flows. Then, the ego vehicle performs a right turn at an unsignalized intersection while navigating a crossing dynamic traffic flow. Following that, the ego vehicle is required to sequentially traverse an unsignalized intersection and a signalized intersection while interacting with dynamic traffic flows. Afterward, the ego vehicle merges into moving highway traffic from a highway on-ramp and engages in a high-speed cruising task with dynamic traffic flows. Finally, the ego vehicle is tasked with exiting the highway at an off-ramp. Here successful driving refers to the vehicle arriving at the off-ramp from the starting point without any collisions or running red lights. The total length of the task is 2400m (700m + 1700m) in the east-west direction and 600m in the north-south direction.

State and action. In this task, apart from utilizing the 26-dimensional state in the scenarios (a)-(d), the agent incorporates three additional states: the distance from the traffic light, the status of the traffic light, and the distance from the navigation target. Consequently, the agent's state encompasses a total of 29 dimensions. Furthermore, the action of the agent includes continuous longitudinal acceleration (or deceleration) as well as lane change direction.

Evaluation. Here, we assess and compare the performance of FNI-RL against the nine baseline approaches. Fig. 3 illustrates the training performance of the nine learning-based autonomous driving agents on the long-term goal-driven navigation task under the flow-0 condition. Quantitatively, we provide the average metrics of the last 100 training episodes for each learning-based method under different random seeds, as shown in Table 2. Correspondingly, we assess the rule-based IDM baseline using the test results from 500 episodes. Fig. 3 and Table 2 demonstrate that, overall, FNI-RL surpasses the baselines with a large margin, in terms of the DS, SR, CR, and TNSV metrics, while performing comparably to the competitive baseline methods in terms of RVR. Specifically, in comparison with the IDM, SAC, SAC-Lag and SMBPO agents, the DS metric of FNI-RL is improved by approximately 78.72%, 64.19%, 39.20% and 14.49%, respectively. Compared with the IDM, SAC, SAC-Lag and SMBPO agents, FNI-RL gains approximately 95.65%, 83.16%, 50.97% and 10.67% improvements with respect to the SR metric, respectively. It is evident that on this challenging long-term goal-driven navigation task, autonomous driving agents trained using baseline methods struggle to effectively avoid collision incidents compared to FNI-RL. In contrast to the PPO, SAC, CPO, SAC-Lag, SMBPO, SMBPPO, GAIL, and Roach agents, the TNSV metric of FNI-RL is approximately reduced by 81.30%, 73.58%,

79.69%, 67.87%, 32.96%, 79.77%, 76.71% and 79.94% in 2000 training episodes, respectively.

We observe that the majority of the autonomous driving agents excel at avoiding running red lights rather than avoiding collisions in the random and dynamic traffic environment. For instance, the rule-based IDM and learning-based Roach methods can ensure complete compliance with red light instructions; however, they prove less effective in enabling autonomous driving agents to avoid collisions effectively. Additionally, we find that the three on-policy RL baselines (i.e., PPO, CPO and SMBPPO) fail to make distinct progress in terms of DS and SR. Unlike off-policy RL methods, which store experiences in a replay buffer for learning, on-policy RL approaches directly update their policy based on the experiences collected during each episode or trajectory. This distinction may be a disadvantage for solving the challenging long-term goal-driven navigation task. In addition, since both GAIL and Roach are based on on-policy RL and the IDM-based demonstration data is of insufficient quality, they similarly fail to achieve the competitive outcomes on this complicated task.

4.6 Human-in-the-Loop Experiment

Task. In Fig. 4(a), we construct three cut-in scenarios (scene-0, scene-1 and scene-2) with different levels of aggressiveness (normal, aggressive and extremely aggressive) to assess the performance of our FNI-RL agent in safety-critical situations compared to 30 certified human drivers. The aggressiveness of the cut-in vehicle is manifested differently in the hesitation time and the longitudinal distance to the maneuver endpoint. The hesitation time is defined as maintaining the original velocity and not initiating any lane changes, and the maneuver endpoint is the longitudinal position at which the cut-in vehicle completes its lane change. The ego vehicle is in the leftmost lane. For the formal experiment, each scenario is repeated five times to assess the average performance of the human and FNI-RL drivers. Finally, we analyze and assess the data derived from the human drivers and the FNI-RL agents, with each participant conducting 5 repeated trials. Since it would be extremely dangerous to perform emergency collision avoidance tasks in a real vehicle, the experiment is conducted in safety-critical situations with the human-in-the-loop platform with the high-fidelity CARLA simulator. The detailed description of the experiment can be found in the supplementary.

State and action. To demonstrate the advantages of our method, for the cut-in scene we constructed, the FNI-RL agent only adopts the information from the 3 nearest vehicles within a 200-meter distance from the ego vehicle, consisting of 7 dimensions, including the ego vehicle's speed, the speed and relative distance of the nearest front and rear vehicles, and the speed and relative distance of the nearest right-side vehicle. Instead, the human drivers can observe relevant information such as the distance and speed of almost all surrounding vehicles in the traffic environment through the screens on the platform. Here, the action of our autonomous driving agent is a continuous control of longitudinal acceleration or deceleration.

Evaluation. The experimental results obtained from three distinct scenarios are evaluated using four different

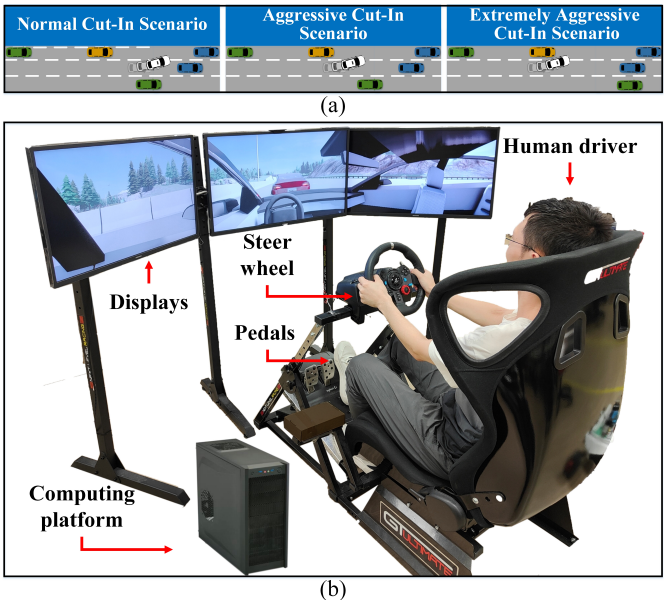


Fig. 4. Human-in-the-loop experiment. (a) Cut-in scenarios with three levels of aggressiveness. The ego vehicle (i.e., the golden-colored vehicle in the leftmost lane) performs a high-speed cruising task while a nearby vehicle suddenly cuts into its lane. The ego vehicle should stay in its lane and avoid collisions to the greatest extent possible. (b) Experimental platform. The human drivers manipulate the steering wheel and pedals to control the ego vehicle. A computing platform and three heads-up displays provide a real-time, high-fidelity in-vehicle view.

metrics. In Fig. 5(a), the success rate is computed by the ratio of successful runs to total runs. A successful run is defined as a trial where the ego vehicle avoids collision with any of the surrounding social vehicles throughout the course of the run. The human drivers recorded success rates of 81.3%, 76.0%, and 70.0% for each scenario respectively. Surprisingly, our FNI-RL agent consistently outperforms the human drivers in all scenarios, achieving a success rate of 100% in each case. Statistical analysis, employing a paired t-test, confirms the superior performance of the FNI-RL agent, where $p <$ 1e-4 for all cases. Fig. 5(b) illustrates the average reciprocal TTC of the ego vehicle with respect to the cut-in vehicle; a higher value suggests a higher risk. The FNI-RL agent consistently exhibits greater safety than human drivers, as evidenced by lower reciprocal TTC values across all scenarios. Statistical significance of this superiority is validated with $p <$ 1e-4 for all cases. In Fig. 5(c), the FNI-RL agent showcases smoother driving across all scenarios, as supported by its lower average acceleration values in comparison to human drivers. Statistical tests confirm the significance of this difference, with $p <$ 1e-2 for scene-0 and $p <$ 1e-4 for scenes-1 and scenes-2. In Fig. 5(d), compared to human drivers, the FNI-RL agent maintains a smaller and more stable effect on the rear vehicle, consequently enhancing overall traffic performance. This improvement is substantiated through t-tests, as depicted in Fig. 5(d).

4.7 Ablation Study

We implement 8 ablation schemes by removing different components and setting various hyperparameters, on the traffic environment (c) shown in Fig. 2. In Table 3, "$-\bar{\pi}(\cdot)$", "$-f(\cdot)$" and "$-\hat{\mathcal{M}}$" correspond to removing the adversary

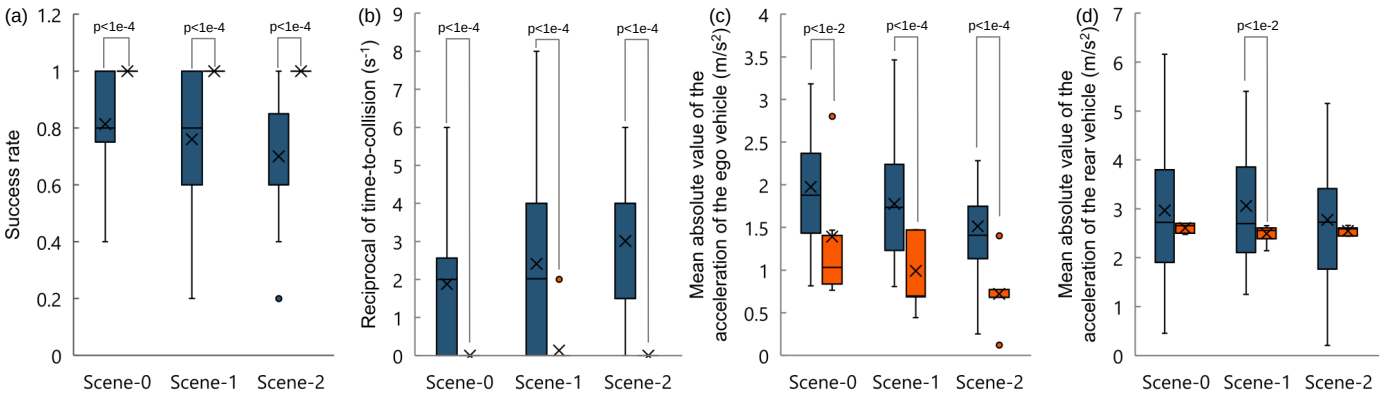


Fig. 5. Statistical results produced by the human drivers (blue bars) and the FNI-RL agents (orange bars). (a) Bar plot of the success rates of the human drivers and the FNI-RL agent. (b) Boxplot of the reciprocal of the time-to-collision values produced by the human drivers and the FNI-RL agent, where the time-to-collision is calculated based on the moment at which the cut-in vehicle reaches the ego lane, and a small but nonzero constant (0.1s) is leveraged as the time-to-collision value for the unsuccessful trials. (c) Boxplot of the mean absolute value of the acceleration of the ego vehicle, where the counting range is 2s from the time at which the cut-in behavior occurs. (d) Boxplot of the mean absolute value of the acceleration of the rear vehicle, where the counting range is 2s from the time at which the cut-in behavior occurs.

TABLE 3
Final performance of different autonomous driving agents in training.

Method	DS	SR	CR	TNSV
FNI-RL	0.97 (0.00)	**1.00 (0.00)**	**0.00 (0.00)**	**90.00 (8.00)**
$-\hat{\mathcal{M}}$	0.90 (0.05)	0.97 (0.05)	0.03 (0.05)	121.00 (13.00)
$-\bar{\pi}(\cdot)$	0.93 (0.03)	0.95 (0.05)	0.05 (0.05)	114.00 (4.00)
$-f(\cdot)$	0.89 (0.06)	**1.00 (0.00)**	**0.00 (0.00)**	160.00 (20.00)
$-\bar{\pi}(\cdot) - f(\cdot)$	0.73 (0.13)	0.80 (0.10)	0.10 (0.10)	169.00 (8.00)
$\alpha = 0.5$	0.97 (0.00)	**1.00 (0.00)**	**0.00 (0.00)**	93.00 (5.00)
$\beta = 0.8$	**0.98 (0.00)**	**1.00 (0.00)**	**0.00 (0.00)**	97.00 (6.00)
$f_0 = 0.1$	0.94 (0.04)	0.95 (0.05)	0.05 (0.05)	103.00 (3.00)
$m = 10$	**0.98 (0.01)**	**1.00 (0.00)**	**0.00 (0.00)**	102.00 (2.00)

policy, the fear model and the virtual experience replay memory from FNI-RL, respectively. "$-\bar{\pi}(\cdot) - f(\cdot)$" represents the baseline that excludes the adversary policy and fear model components from FNI-RL. To analyze the effect of several key hyperparameters, "$\alpha = 0.8$", "$\beta = 0.9$", "$f_0 = 0.5$" and "$m = 5$" in FNI-RL are set as "$\alpha = 0.5$", "$\beta = 0.8$", "$f_0 = 0.1$" and "$m = 10$", respectively. For the "$-\bar{\pi}(\cdot)$" and "$\alpha = 1.0$" baselines, these two are equivalent.

Overall, FNI-RL performs comparably to the baselines of changing hyperparameters and outperforms the baselines of removing critical components, in terms of the final DS, SR and CR. Most notably, FNI-RL exhibits significant advantages in terms of safety, especially in terms of training-time safety. Specifically, compared with the "$-\hat{\mathcal{M}}$", "$-\bar{\pi}(\cdot)$", "$-f(\cdot)$" and "$-\bar{\pi}(\cdot) - f(\cdot)$" baselines, the TNSV metric of FNI-RL is approximately reduced by 25.62%, 21.05%, 43.75%, and 46.75% in 2000 training episodes, respectively. From the results in Table 3, we can see that the component regarding the fear model has a significant impact on the performance of FNI-RL, especially in safety. In addition, by comparing the "$\alpha = 0.5$", "$\beta = 0.8$", "$f_0 = 0.1$" and "$m = 10$" baselines, we can find that hyperparameters have a certain impact on the performance of FNI-RL, but in general FNI-RL is not very sensitive to changes in hyperparameters. Consequently, the results of the ablation analysis demonstrate that the components or setting in FNI-RL are critical. More results can be found in the supplementary.

5 Discussion and Conclusion

Performance. Inspired by the amygdala, which arouses the fear and defensive behaviors of organisms in response to the recognition of dangers or contingencies, we propose the FNI-RL framework to realize safe autonomous driving.

The results demonstrate the effectiveness of FNI-RL via simulations and experiments. In the scenarios (a)-(e), FNI-RL achieves superior performance to that of the competitive AI agents, especially in terms of safety. In the human-in-the-loop experiment, one obstacle to evaluating our agent is the "transfer gap": the performance of the well-trained agent in the SUMO-based simulation can be easily degraded in the experiment. One major reason for this problem may be the differences in the vehicle models between the two environments. Surprisingly, the experimental results indicate that FNI-RL can achieve the performance of the 30 certified human drivers in three safety-critical scenarios. Additionally, the ablation studies show that the components in FNI-RL to simulate the amygdala mechanism are critical.

Diving deeper into the results. We find four possible explanations for the above results. (1) Threats and contingencies can be recognized or estimated with the fear model. FNI-RL selects the action that minimizes fear during interactions with the real environment. (2) While prediction error is unavoidable, by combining the adversarial agent with the world model, the adversarial imagination technique is able to simulate the worst-case situations in the imagination, enabling the agent to tackle unseen critical situations and improve its policy robustness against the "transfer gap" or uncertainties. (3) The FC-AC algorithm enables the agent to learn defensive driving behaviors that ensure safety or performance during emergencies. (4) Compared with human drivers, autonomous driving systems have faster reaction times and are fatigue-proof in terms of their functioning.

Broader impact. RL has been an impressive component of modern AI and is still under vigorous development. Nonetheless, unlike supervised learning, which has found extensive application in various commercial and industrial domains, RL has not gained widespread acceptance and deployment in real-world tasks. One important aspect is

the trustworthiness, where safety plays a critical role. Compared to AI, especially RL, human intelligence is considered safer and more trustworthy. Our framework inspired by the brain fear circuit contributes to the foundation for realizing safe AI, potentially bringing RL closer to safety-critical real-world applications. Moreover, this work establishes linkages between AI, neuroscience and psychology, which may be beneficial for interpreting the RL process in the brain.

Limitations and future work. Our algorithm implementation has several simplifications (e.g., its network structure and limited states) for the convenience of simulation and experimentation. We believe that neural networks considering temporal sequences, e.g., transformer [1], could improve the performance of FNI-RL, and this topic will be studied in the future. Additionally, the amygdala enables organisms to learn at fast rates and track rapid changes in environments, while the striatum is more robust to noise [14]. However, since the internal structure and mechanism of the amygdala and striatum remain unclear, FNI-RL has not lived up to its full potential. An additional investigation is required to elucidate the fundamental principles of the amygdala and striatum, fostering the development of RL-based computational models and high-level autonomous driving.

ACKNOWLEDGMENTS

The authors appreciate the contributions of all participants in the experiments. This work was supported in part by the Start-Up Grant-Nanyang Assistant Professorship Grant of Nanyang Technological University, and the Agency for Science, Technology and Research (A*STAR), Singapore, under Advanced Manufacturing and Engineering (AME) Young Individual Research Grant (No. A2084c0156).

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[2] C. König, M. Turchetta, J. Lygeros, A. Rupenyan, and A. Krause, "Safe and efficient model-free adaptive control via bayesian optimization," in *2021 IEEE International Conference on Robotics and Automation (ICRA)*, 2021, pp. 9782–9788.

[3] S. Baik, M. Choi, J. Choi, H. Kim, and K. M. Lee, "Learning to learn task-adaptive hyperparameters for few-shot learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2023.

[4] J. Rothfuss, C. Koenig, A. Rupenyan, and A. Krause, "Meta-learning priors for safe bayesian optimization," in *Conference on Robot Learning*. PMLR, 2023, pp. 237–265.

[5] Q. Li, Z. Peng, L. Feng, Q. Zhang, Z. Xue, and B. Zhou, "Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 3, pp. 3461–3475, 2023.

[6] S. Feng, H. Sun, X. Yan, H. Zhu, Z. Zou, S. Shen, and H. X. Liu, "Dense reinforcement learning for safety validation of autonomous vehicles," *Nature*, vol. 615, no. 7953, pp. 620–627, 2023.

[7] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *nature*, vol. 529, no. 7587, pp. 484–489, 2016.

[8] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev *et al.*, "Grandmaster level in starcraft ii using multi-agent reinforcement learning," *Nature*, vol. 575, no. 7782, pp. 350–354, 2019.

[9] P. R. Wurman, S. Barrett, K. Kawamoto, J. MacGlashan, K. Subramanian, T. J. Walsh, R. Capobianco, A. Devlic, F. Eckert, F. Fuchs *et al.*, "Outracing champion gran turismo drivers with deep reinforcement learning," *Nature*, vol. 602, no. 7896, pp. 223–228, 2022.

[10] B. R. Kiran, I. Sobh, V. Talpaert, P. Mannion, A. A. Al Sallab, S. Yogamani, and P. Pérez, "Deep reinforcement learning for autonomous driving: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 6, pp. 4909–4926, 2021.

[11] B. Mirchevska, C. Pek, M. Werling, M. Althoff, and J. Boedecker, "High-level decision making for safe and reasonable autonomous lane changing using reinforcement learning," in *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2018, pp. 2156–2162.

[12] L. Wen, J. Duan, S. E. Li, S. Xu, and H. Peng, "Safe reinforcement learning for autonomous vehicles through parallel constrained policy optimization," in *2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC)*. IEEE, 2020, pp. 1–7.

[13] A. Zador, S. Escola, B. Richards, B. Ölveczky, Y. Bengio, K. Boahen, M. Botvinick, D. Chklovskii, A. Churchland, C. Clopath *et al.*, "Catalyzing next-generation artificial intelligence through neuroai," *Nature communications*, vol. 14, no. 1, p. 1597, 2023.

[14] E. O. Neftci and B. B. Averbeck, "Reinforcement learning in artificial and biological systems," *Nature Machine Intelligence*, vol. 1, no. 3, pp. 133–143, 2019.

[15] V. D. Costa, O. Dal Monte, D. R. Lucas, E. A. Murray, and B. B. Averbeck, "Amygdala and ventral striatum make distinct contributions to reinforcement learning," *Neuron*, vol. 92, no. 2, pp. 505–517, 2016.

[16] B. B. Averbeck and V. D. Costa, "Motivational neural circuits underlying reinforcement learning," *Nature Neuroscience*, vol. 20, no. 4, pp. 505–512, 2017.

[17] J. LeDoux and N. D. Daw, "Surviving threats: neural circuit and computational implications of a new taxonomy of defensive behaviour," *Nature Reviews Neuroscience*, vol. 19, no. 5, pp. 269–282, 2018.

[18] D. Schiller, I. Levy, Y. Niv, J. E. LeDoux, and E. A. Phelps, "From fear to safety and back: reversal of fear in the human brain," *Journal of Neuroscience*, vol. 28, no. 45, pp. 11 517–11 525, 2008.

[19] P. A. Lopez, M. Behrisch, L. Bieker-Walz, J. Erdmann, Y.-P. Flötteröd, R. Hilbrich, L. Lücken, J. Rummel, P. Wagner, and E. Wießner, "Microscopic traffic simulation using sumo," in *2018 21st international conference on intelligent transportation systems (ITSC)*. IEEE, 2018, pp. 2575–2582.

[20] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun, "Carla: An open urban driving simulator," in *Conference on robot learning*. PMLR, 2017, pp. 1–16.

[21] H. Shao, L. Wang, R. Chen, H. Li, and Y. Liu, "Safety-enhanced autonomous driving using interpretable sensor fusion transformer," in *Conference on Robot Learning*. PMLR, 2023, pp. 726–737.

[22] K. Muhammad, A. Ullah, J. Lloret, J. D. Ser, and V. H. C. de Albuquerque, "Deep learning for safe autonomous driving: Current challenges and future directions," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 7, pp. 4316–4336, 2021.

[23] E. Bronstein, M. Palatucci, D. Notz, B. White, A. Kuefler, Y. Lu, S. Paul, P. Nikdel, P. Mougin, H. Chen *et al.*, "Hierarchical model-based imitation learning for planning in autonomous driving," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2022, pp. 8652–8659.

[24] L. Le Mero, D. Yi, M. Dianati, and A. Mouzakitis, "A survey on imitation learning techniques for end-to-end autonomous vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 9, pp. 14 128–14 147, 2022.

[25] Z. Gu, L. Gao, H. Ma, S. E. Li, S. Zheng, W. Jing, and J. Chen, "Safe-state enhancement method for autonomous driving via direct hierarchical reinforcement learning," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–18, 2023.

[26] X. He, H. Yang, Z. Hu, and C. Lv, "Robust lane change decision making for autonomous vehicles: An observation adversarial reinforcement learning approach," *IEEE Transactions on Intelligent Vehicles*, vol. 8, no. 1, pp. 184–193, 2023.

[27] J. Wang, A. Pun, J. Tu, S. Manivasagam, A. Sadat, S. Casas, M. Ren, and R. Urtasun, "Advsim: Generating safety-critical scenarios for self-driving vehicles," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9909–9918.

[28] D. Rempe, J. Philion, L. J. Guibas, S. Fidler, and O. Litany, "Generating useful accident-prone driving scenarios via a learned traffic

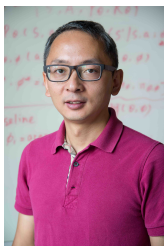
prior," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 17 305–17 315.
[29] N. Hanselmann, K. Renz, K. Chitta, A. Bhattacharyya, and A. Geiger, "King: Generating safety-critical driving scenarios for robust imitation via kinematics gradients," in *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXVIII*. Springer, 2022, pp. 335–352.
[30] W. Ding, H. Lin, B. Li, and D. Zhao, "Causalaf: Causal autoregressive flow for safety-critical driving scenario generation," in *Conference on Robot Learning*. PMLR, 2023, pp. 812–823.
[31] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. Van Gool, "End-to-end urban driving by imitating a reinforcement learning coach," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 15 222–15 232.
[32] Q. Li, Z. Peng, and B. Zhou, "Efficient learning of safe driving policy via human-AI copilot optimization," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=0cgU-BZp2ky
[33] L. Zhang, R. Zhang, T. Wu, R. Weng, M. Han, and Y. Zhao, "Safe reinforcement learning with stability guarantee for motion planning of autonomous vehicles," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 12, pp. 5435–5444, 2021.
[34] J. Wang, Q. Zhang, D. Zhao, and Y. Chen, "Lane change decision-making through deep reinforcement learning with rule-based constraints," in *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019, pp. 1–6.
[35] X. He, B. Lou, H. Yang, and C. Lv, "Robust decision making for autonomous vehicles at highway on-ramps: A constrained adversarial reinforcement learning approach," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 4, pp. 4103–4113, 2023.
[36] J. Achiam, D. Held, A. Tamar, and P. Abbeel, "Constrained policy optimization," in *International conference on machine learning*. PMLR, 2017, pp. 22–31.
[37] Y. Zhang, Q. Vuong, and K. Ross, "First order constrained optimization in policy space," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 338–15 349, 2020.
[38] S. Ha, P. Xu, Z. Tan, S. Levine, and J. Tan, "Learning to walk in the real world with minimal human effort," in *Conference on Robot Learning*. PMLR, 2021, pp. 1110–1120.
[39] D. Yu, H. Ma, S. Li, and J. Chen, "Reachability constrained reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2022, pp. 25 636–25 655.
[40] N. Kochdumper, H. Krasowski, X. Wang, S. Bak, and M. Althoff, "Provably safe reinforcement learning via action projection using reachability analysis and polynomial zonotopes," *IEEE Open Journal of Control Systems*, vol. 2, pp. 79–92, 2023.
[41] M. Turchetta, A. Kolobov, S. Shah, A. Krause, and A. Agarwal, "Safe reinforcement learning via curriculum induction," *Advances in Neural Information Processing Systems*, vol. 33, pp. 12 151–12 162, 2020.
[42] B. Thananjeyan, A. Balakrishna, S. Nair, M. Luo, K. Srinivasan, M. Hwang, J. E. Gonzalez, J. Ibarz, C. Finn, and K. Goldberg, "Recovery rl: Safe reinforcement learning with learned recovery zones," *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 4915–4922, 2021.
[43] F. Berkenkamp, M. Turchetta, A. Schoellig, and A. Krause, "Safe model-based reinforcement learning with stability guarantees," *Advances in neural information processing systems*, vol. 30, 2017.
[44] Y. J. Ma, A. Shen, O. Bastani, and J. Dinesh, "Conservative and adaptive penalty for model-based safe reinforcement learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 5, 2022, pp. 5404–5412.
[45] G. Thomas, Y. Luo, and T. Ma, "Safe reinforcement learning by imagining the near future," *Advances in Neural Information Processing Systems*, vol. 34, pp. 13 859–13 869, 2021.
[46] M. A. Zanger, K. Daaboul, and J. M. Zöllner, "Safe continuous control with constrained model-based policy optimization," in *2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2021, pp. 3512–3519.
[47] Y. As, I. Usmanova, S. Curi, and A. Krause, "Constrained policy optimization via bayesian world models," in *International Conference on Learning Representations*, 2022. [Online]. Available: https://openreview.net/forum?id=PRZoSmCinhf
[48] A. K. Jayant and S. Bhatnagar, "Model-based safe deep reinforcement learning via a constrained proximal policy optimization algorithm," *Advances in Neural Information Processing Systems*, vol. 35, pp. 24 432–24 445, 2022.
[49] L. E. Williams, J. A. Oler, A. S. Fox *et al.*, "Fear of the unknown: uncertain anticipation reveals amygdala alterations in childhood anxiety disorders," *Neuropsychopharmacology*, vol. 40, no. 6, pp. 1428–1435, 2015.
[50] R. N. Carleton, "Fear of the unknown: One fear to rule them all?" *Journal of Anxiety Disorders*, vol. 41, pp. 5–21, 2016, fearing the Unknown.
[51] K. Chua, R. Calandra, R. McAllister, and S. Levine, "Deep reinforcement learning in a handful of trials using probabilistic dynamics models," *Advances in neural information processing systems*, vol. 31, 2018.
[52] H. Van Hasselt, A. Guez, and D. Silver, "Deep reinforcement learning with double q-learning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016.
[53] M. Janner, J. Fu, M. Zhang, and S. Levine, "When to trust your model: Model-based policy optimization," *Advances in neural information processing systems*, vol. 32, 2019.
[54] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.
[55] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine, "Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor," in *International conference on machine learning*. PMLR, 2018, pp. 1861–1870.
[56] J. Ho and S. Ermon, "Generative adversarial imitation learning," *Advances in neural information processing systems*, vol. 29, 2016.
[57] U. D. of Transportation Federal Highway Administration, "Next generation simulation (ngsim) vehicle trajectories and supporting data," 2016. [Online]. Available: http://doi.org/10.21949/1504477

Xiangkun He (Member, IEEE) received his PhD degree in 2019 from the School of Vehicle and Mobility, Tsinghua University, Beijing, China. From 2019 to 2021, he served as a Senior Researcher at Huawei Noah's Ark Lab. He is currently a Research Fellow at Nanyang Technological University, Singapore. His research interests include autonomous driving, reinforcement learning, trustworthy AI, decision and control. He received many awards or honors, selectively including the Tsinghua University Outstanding Doctoral Thesis Award in 2019, Best Paper Finalist at 2020 IEEE ICMA, 1st Class Outstanding Paper of China Journal of Highway and Transport in 2021, Huawei Major Technological Breakthrough Award in 2021, Best Paper Runner-Up Award at 2022 6th CAA International Conference on Vehicular Control and Intelligence, and Runner-Up at Intelligent Algorithm Final of 2022 Alibaba Global Future Vehicle Challenge.

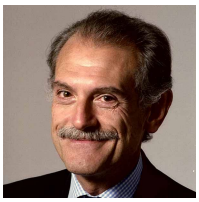
Wu Jingda (Graduate Student Member, IEEE) received his B.S. (2016) and M.S. (2019) in mechanical engineering from Beijing Institute of Technology, China. He is currently working on his Ph.D. degree with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore. His research interests include human guidance-based reinforcement learning algorithms, human-artificial intelligence (AI) collaborated driving strategy design, and decision-making of autonomous vehicles.

Zhiyu Huang (Graduate Student Member, IEEE) received his B.E. degree from the School of Automobile Engineering, Chongqing University, Chongqing, China, in 2019. He is currently pursuing his Ph.D. degree with the School of Mechanical and Aerospace Engineering, Nanyang Technological University, Singapore. His current research focuses on machine learning-based methods for decision-making in autonomous driving, including reinforcement learning, behavior prediction, and data-driven motion planning.

Zhongxu Hu (Member, IEEE) received a mechatronic Ph.D. degree from the Huazhong University of Science and Technology of China, in 2018. He was a senior engineer at Huawei. He is currently a Research Fellow at Nanyang Technological University, Singapore. His current research interests include human-machine collaboration, computer vision, and deep learning applied to autonomous vehicles. Dr. Hu serves as a Lead Guest Editor for Computational Intelligence and Neuroscience, an Academic Editor/Editorial Board for Automotive Innovation, Journal of Electrical and Electronic Engineering, Advances in Multimedia.



Jun Wang is Chair Professor, Computer Science, University College London, and Founding Director of MSc Web Science and Big Data Analytics. He is also Co-founder and Chief Scientist in MediaGamma Ltd, a UCL start-up company focusing on AI for intelligent audience decision making. Prof. Jun Wang's main research interests are in the areas of AI and intelligent systems, including (multiagent) reinforcement learning, deep generative models, and their diverse applications on information retrieval, recommender systems and personalization, data mining, smart cities, bot planning, computational advertising etc. His team won the first global real-time bidding algorithm contest with 80+ participants worldwide. Jun has published over 100 research papers and is a winner of multiple "Best Paper" awards. He was a recipient of the Beyond Search – Semantic Computing and Internet Economics award by Microsoft Research and also received Yahoo! FREP Faculty award. He has served as an Area Chair in ACM CIKM and ACM SIGIR. His recent service includes co-chair of Artificial Intelligence, Semantics, and Dialog in ACM SIGIR 2018. MediaGamma has received the UCLB One-to-Watch award 2016.



Alberto Sangiovanni-Vincentelli (Life Fellow, IEEE) is the Edgar L. and Harold H. Buttner Chair with EECS Department, UC Berkeley, Berkeley, CA, USA. He co-founded Cadence and Synopsys, the two leading EDA companies. He is on the Board of Directors of Cadence, KPIT, Expert Systems, Cy4Gate, Exein, Quantum Motion, Phononic Vibes and Phoelex. He is a Member of the advisory board of Walden International and Xseed, of the Scientific Advisory Board of the Italian Institute of Technology and the Chair of the International Advisory Board for the Milano Innovation District. He is a Member of the Advisory Board of the Politecnico di Milano and honorary Professor at Politecnico di Torino. He was the President of the Comitato Nazionale dei Garanti della Ricerca and of the Strategy Committee of Fondo Strategico Italiano. He consulted for companies such as Intel, HP, Bell Labs, IBM, Lendlease, Samsung, UTC, Lutron, Camozzi Group, Kawasaki Steel, Fujitsu, Telecom Italia, Pirelli, GM, BMW, Mercedes, Magneti Marelli, ST Microelectronics, and ELT. He has authored 1,120 papers, 19 books, and two patents. He is Fellow of the ACM and a Member of the National Academy of Engineering. He earned the IEEE/RSE Maxwell Award for groundbreaking contributions that have had an exceptional impact on the development of electronics and electrical engineering or related fields, the Kaufmann Award, the EDAA lifetime Achievement Award, the IEEE/ACM R. Newton Impact Award, the UC Distinguished Teaching Award, the IEEE TC-CPS Technical Achievement Award, the IEEE Leon Kirchmayer Graduate Teaching Award, and the ISPD Lifetime Achievement Award.

Chen Lv (Senior Member, IEEE) is a Nanyang Assistant Professor at School of Mechanical and Aerospace Engineering, and the Cluster Director in Future Mobility Solutions, Nanyang Technological University, Singapore. He received his PhD degree at Department of Automotive Engineering, Tsinghua University, China in Jan 2016. He was a joint PhD researcher at UC Berkeley, USA during 2014-2015, and worked as a Research Fellow at Cranfield University, UK during 2016-2018. He joined NTU and founded the Automated Driving and Human-Machine System (AutoMan) Research Lab since June 2018. His research focuses on intelligent vehicles, automated driving, and human-machine systems, where he has contributed 2 books, over 100 papers, and obtained 12 granted patents. He serves as Associate Editor for IEEE T-ITS, IEEE TVT, and IEEE T-IV. He received many awards and honors, selectively including the Highly Commended Paper Award of IMechE UK in 2012, Japan NSK Outstanding Mechanical Engineering Paper Award in 2014, Tsinghua University Outstanding Doctoral Thesis Award in 2016, IEEE IV Best Workshop/Special Session Paper Award in 2018, Automotive Innovation Best Paper Award in 2020, the winner of Waymo Open Dataset Challenges at CVPR 2021, Machines Young Investigator Award in 2022, and Best Paper Runner Up Award at CVCI 2022.

Fear-Neuro-Inspired Reinforcement Learning for Safe Autonomous Driving

Xiangkun He[1], Jingda Wu[1], Zhiyu Huang[1], Zhongxu Hu[1], Jun Wang[2], Alberto Sangiovanni-Vincentelli[3], Chen Lv[1,*]

[1]School of Mechanical and Aerospace Engineering, Nanyang Technological University

[2]Department of Computer Science, University College London

[3]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley

[*]Corresponding author. E-mail: lyuchen@ntu.edu.sg

Supplementary materials

Supplementary Notes

Supplementary Note 1-Implementation of Reward and Cost Functions

Algorithm 1 Reward and cost functions for the scenarios (a)-(d) and the cut-in scenarios

Input: State and action of the RL agent.

1: $r(s,a) = v_0/10$. ▷ Encourage agent to be more efficient
2: **if** Collision **then**
3: $\quad c(s,a) = 1.00$. ▷ Penalize collisions
4: **else**
5: $\quad c(s,a) = 0.00$.
6: **end if**

Output: $r(s,a) = r(s,a) - c(s,a)$, $c(s,a)$.

In Algorithms 1, v_0 represents the speed of the ego vehicle.

In Algorithms 2, c_1 and c_2 correspond to the cost functions associated with collisions and running a red light, respectively. Moreover, Δd and Δd_{max} denote the distance from the target off-ramp and the maximum distance from the target off-ramp, respectively.

Algorithm 2 Reward and cost functions for the long-term goal-driven navigation task

Input: State and action of the RL agent.

$r(s,a) = v_0/5.$ ▷ Encourage agent to be more efficient

if Collision **then**

  $c_1(s,a) = 1.00.$ ▷ Penalize collisions

else

  $c_1(s,a) = 0.00.$

end if

if Running a red light **then**

  $c_2(s,a) = 1.00.$ ▷ Penalize red-light violations

else

  $c_2(s,a) = 0.00.$

end if

if Arriving at the off-ramp **then**

  $r(s,a) = r(s,a) + 100.00.$ ▷ Encourage agent arriving at the off-ramp

else

  $r(s,a) = r(s,a) - \log(1.00 + \Delta d/\Delta d_{\text{max}})\text{-}1.00.$ ▷ Lead to the off-ramp

end if

Output: $r(s,a) = r(s,a) - c_1(s,a) - c_2(s,a)$, $c_1(s,a)$, $c_2(s,a)$.

Supplementary Note 2-Proof of Convergence of Fear-Constrained Policy Iteration

We provide the following proof to show that the fear-constrained policy iteration (FC-PI) can converge to the optimal policy.

Lemma 1 *(Fear-Constrained Policy Evaluation). Consider the bellman backup operator $\mathcal{T}$ in Eq. (9) and a state-action function $Q_0^\pi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ with $|\mathbb{R}| < \infty$, and define $Q_{i+1}^\pi = \mathcal{T}Q_i^\pi$. Then, the sequence Q_i^π can converge to a unique fixed the Q-value of π as $i \to \infty$.*

Proof: We can rewrite the update rule as via Eq. (9):

$$\begin{aligned}\mathcal{T}Q^\pi(s,a) &\equiv r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim p, a'\sim\pi(\cdot|s')}[Q^\pi(s,a) - \lambda f^\pi] \\ &= r_\pi(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim p, a'\sim\pi(\cdot|s')}[Q^\pi(s',a')], \end{aligned} \tag{1}$$

where $r_\pi(s,a) = r(s,a) - \gamma\lambda \mathop{\mathbb{E}}_{s'\sim p, a'\sim\pi(\cdot|s')}[f^\pi]$, $r_\pi(s,a)$ represents a fear augmented reward, f^π denotes an on-policy fear model.

Thus for any $Q^\pi(s,a), Q_i^\pi(s,a) : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$,

$$\begin{aligned}\|\mathcal{T}Q^\pi(s,a) - \mathcal{T}Q_i^\pi(s,a)\|_\infty &= \sup|\mathcal{T}Q^\pi(s,a) - \mathcal{T}Q_i^\pi(s,a)| \\ &= \gamma \sup\left| \mathop{\mathbb{E}}_{s'\sim p, a'\sim\pi(\cdot|s')}\big[Q^\pi(s',a') - Q_i^\pi(s',a')\big]\right| \\ &\le \gamma \sup \mathop{\mathbb{E}}_{s'\sim p, a'\sim\pi(\cdot|s')}\big[\,|Q^\pi(s',a') - Q_i^\pi(s',a')|\,\big] \\ &\le \gamma \sup|Q^\pi(s',a') - Q_i^\pi(s',a')| \\ &= \gamma\|Q^\pi(s',a') - Q_i^\pi(s',a')\|_\infty. \end{aligned} \tag{2}$$

Hence $\mathcal{T}$ is indeed a γ-contraction in ∞-norm. In other words, $\mathcal{T}$ has a unique fixed point which can be obtained by iteration.

Lemma 2 *(Fear-Constrained Policy Improvement) Let π_{new} be the optimal solution of the maximization problem defined in Eq. (7). Then $Q^{\pi_{\text{new}}}(s,a) \ge Q^{\pi_{\text{old}}}(s,a)$ for $\forall(s,a) \in \mathcal{S} \times \mathcal{A}$.*

Proof: Let π_{new} be defined as:

$$\pi_{\text{new}}(\cdot|s) = \arg\max_{\pi} \mathop{\mathbb{E}}_{a\sim\pi}[Q^{\pi_{\text{old}}}(s,a) - \lambda f^{\pi}], \quad \forall s \in \mathcal{S}. \tag{3}$$

Since we can always choose $\pi_{\text{new}} = \pi_{\text{old}}$, then it is obvious that:

$$\mathop{\mathbb{E}}_{a\sim\pi_{\text{new}}}[Q^{\pi_{\text{old}}}(s,a) - \lambda f^{\pi_{\text{new}}}] \geq \mathop{\mathbb{E}}_{a\sim\pi_{\text{old}}}[Q^{\pi_{\text{old}}}(s,a) - \lambda f^{\pi_{\text{old}}}], \quad \forall s \in \mathcal{S}. \tag{4}$$

Next, with Eq. (9), it follows that:

$$\begin{aligned}
&Q^{\pi_{\text{old}}}(s,a) \\
&= r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim p, a'\sim\pi_{\text{old}}}[Q^{\pi_{\text{old}}}(s',a') - \lambda f^{\pi_{\text{old}}}] \\
&\leq r(s,a) + \gamma \mathop{\mathbb{E}}_{s'\sim p, a'\sim\pi_{\text{new}}}[Q^{\pi_{\text{old}}}(s',a') - \lambda f^{\pi_{\text{new}}}] \\
&\vdots \\
&\leq Q^{\pi_{\text{new}}}(s,a), \quad \forall (s,a) \in \mathcal{S}\times\mathcal{A},
\end{aligned}$$

where we have repeatedly expanded $Q^{\pi_{\text{old}}}$ on the right-hand side by applying the Bellman equation.

Theorem 1 *(Fear-Constrained Policy Iteration). The fear-constrained policy iteration, which alternates between the fear-constrained policy evaluation and the fear-constrained policy improvement, can converge to a policy π^* such that $Q^{\pi^*}(s,a) \geq Q^{\pi}(s,a)$ for $\forall\pi$ and $\forall(s,a) \in \mathcal{S}\times\mathcal{A}$, assuming that $|\mathcal{A}| < \infty$ and reward is bounded.*

Proof: Let π_k denote the policy at iteration k. For $\forall\pi_k$, we can always find its associated Q^{π_k} via the fear-constrained policy evaluation process follows from Lemma 1. With Lemma 2, the sequence $Q^{\pi_k}(s,a)$ is monotonically increasing for $\forall(s,a) \in \mathcal{S}\times\mathcal{A}$. Since Q^{π} is bounded everywhere for $\forall\pi$ (both the reward and fear model are bounded), the policy sequence π_k converges to some $\pi^\dagger$ as $k \to \infty$. At convergence, it must follow that:

$$\mathop{\mathbb{E}}_{a'\sim\pi^\dagger}[Q^{\pi^\dagger}(s,a) - \lambda f^{\pi^\dagger}] \geq \mathop{\mathbb{E}}_{a'\sim\pi}[Q^{\pi^\dagger}(s,a) - \lambda f^{\pi}], \quad \forall\pi, \forall s \in \mathcal{S}. \tag{5}$$

With the same iterative argument as in Lemma 2, the following inequality can be derived:

$$Q^{\pi^\dagger}(s,a) \geq Q^{\pi}(s,a), \quad \forall \pi, \forall (s,a) \in \mathcal{S} \times \mathcal{A}.$$

That is, the value of any other policy is lower than that of the converged policy $\pi^\dagger$. Consequently, $\pi^\dagger$ is optimal, namely, $\pi^\dagger = \pi^*$.

Supplementary Note 3-Implementation of Fear-Neuro-Inspired Reinforcement Learning

Our code is available at https://github.com/TMIS-Turbo/FNI-RL.

Algorithm 3 Fear-neuro-inspired reinforcement learning

Initialize wold model network parameters ϕ_k.
Initialize adversarial policy network parameters $\bar{\theta}$, policy network parameters θ, action-value network parameters ϕ^1 and ϕ^2, target action-value network parameters $\bar{\phi^1} \leftarrow \phi^1$, and $\bar{\phi^2} \leftarrow \phi^2$.
Initialize an empty replay memory $\mathcal{M}$ and an empty virtual replay memory $\hat{\mathcal{M}}$, a dual variable λ.
for episode $n_e = 1, 2, \dots N_e$ **do**
  Reset state s_0.
  for time step in the environment $t = 1, 2, \dots T$ **do**
    Observe state s_t and sample action:
    $a_t \sim \pi(\cdot|s_t)$.
    Sample a transition from the environment:
    $s_{t+1}, r_t, c_t, d_t \sim p(\cdot|s_t, a_t)$.
    Store the transition in the experience replay memory $\mathcal{M}$:
    $\mathcal{M} \leftarrow \mathcal{M} \cup \{(s_t, a_t, r_t, c_t, s_{t+1}, d_t)\}$.
  end for
  for gradient step of the world model $g_w = 1, 2, \dots G_w$ **do**
    Sample mini-batch of transitions from the experience replay memory $\mathcal{M}$.
    Update the ensemble of diagonal Gaussian world models network parameters via Eq. (3):
    for the kth world model $k = 1, 2, \dots K$ **do**
      $\phi_k \leftarrow \nabla_{\phi_k} J_w(\phi_k)$.
    end for
  end for
  for gradient step of the agent $g_a = 1, 2, \dots G_a$ **do**
    Sample mini-batch of states s from the real experience replay memories $\mathcal{M}$.
    for rollout horizon in the world model $h = 1, 2, \dots H$ **do**
      Sample mixed actions using Eq. (1):
      $\tilde{a} \sim \pi^{\text{mix}}(\cdot|s)$.
      Sample virtual transitions from the world model $\hat{T}$:
      $\hat{s}', \hat{c} \sim \hat{T}(\cdot|s, \tilde{a})$.
      Store the transitions in the virtual experience replay memory $\hat{\mathcal{M}}$:
      $\hat{\mathcal{M}} \leftarrow \hat{\mathcal{M}} \cup \{(s, \tilde{a}, \hat{s}', \hat{c})\}$.
    end for
    Sample mini-batch of transitions from the real experience replay memories $\mathcal{M}$.
    Update the action-value network parameters using Eq. (11):
    $\phi^1 \leftarrow \nabla_{\phi^1} J_c(\phi^1)$, $\phi^2 \leftarrow \nabla_{\phi^2} J_c(\phi^2)$.
    Sample mini-batch of transitions from the real and virtual replay memories $\mathcal{M} \cup \hat{\mathcal{M}}$.
    Update the policy and adversarial policy network parameters using Eq. (19) and Eq. (6):
    $\theta \leftarrow \nabla_\theta J_a(\theta)$, $\bar{\theta} \leftarrow \nabla_{\bar{\theta}} J_{\bar{a}}(\bar{\theta})$.
    Update the dual variables using Eq. (20):
    $\lambda \leftarrow \nabla_\lambda J_d(\lambda)$.
    Update the target action-value network parameters using polyak averaging:
    $\bar{\phi^1} \leftarrow \rho\bar{\phi^1} + (1-\rho)\phi^1$, $\bar{\phi^2} \leftarrow \rho\bar{\phi^2} + (1-\rho)\phi^2$.
  end for
end for

Supplementary Note 4-Implementation of the Baseline Algorithms

The PPO, SAC, CPO, SAC-Lag, SMBPO, SMBPPO, GAIL and Roach baseline methods are implemented based on the following codebases:

https://github.com/DLR-RM/stable-baselines3,

https://github.com/rail-berkeley/softlearning,

https://github.com/jachiam/cpo,

https://github.com/liuzuxin/FSRL,

https://github.com/gwthomas/Safe-MBPO,

https://github.com/akjayant/mbppol,

https://github.com/openai/imitation,

https://github.com/zhejz/carla-roach).

We are very grateful to the relevant researchers for their contributions.

For the NGSIM-based initial policy models of the two IL methods, we process the raw data from the US 101 highway dataset to obtain the training dataset for highway scenario. The raw trajectory data contains a lot of noises due to sensory and processing errors, and thus smoothing the trajectories at first is necessary. The trajectory smoothing is done with two steps: first smoothing the x and y values using the Savitzky-Golay Filter and then recomputing velocities and accelerations with respect to the smoothed x and y values. Note that we have converted the unit in the dataset from feet to meters.

Additionally, we select the vehicles running on the mainline lanes (lane 1 to lane 5) and treat them as ego vehicles. For each ego vehicle, we select 20 timesteps evenly from its whole trajectory in the section. We adopt information from the 6 nearest vehicles within a 200-meter distance from the ego vehicle, encompassing the relative distance, orientation, speed, and velocity direction of the front, back, left-front, left-back, right-front, and right-back vehicles. Moreover, we incorporate the speed and velocity direction of the ego vehicle, resulting in a

state representation of the agent with a total of 26 dimensions. The ego vehicle's instantaneous acceleration is used as the human driver's action as we do not consider the lane change behavior. All the features will be normalized to $[0, 1]$ by dividing their respective maximum values. To balance the training data, we down-sample the data points with an acceleration between -1.0 and 1.0 m/s^2 and up-sample the data points with an acceleration lower than -3 m/s^2, in order to enable the policy to learn emergency brake. The amount of training data for the highway driving scenario is eventually 65623.

The NGSIM-driver builds a mapping from the feature vector to action, aiming to reproduce the human driver's actions under given states. The policy should address uncertainties, including the uncertainty of human behaviors and the uncertainty of model outputs. Therefore, we use the deep ensemble method to learn the human policy, which consists of an ensemble of M neural networks with the same structure but different random initializations. Each neural network is a two-layer MLP with 256 hidden neurons and ReLU non-linearity. The output of the neural network is the parameters of a Gaussian distribution, i.e., the mean $\hat{\mu}$ and variance $\hat{\sigma}^2$. The Gaussian distribution is utilized here to capture the uncertainty of human actions. On the other hand, to capture the model uncertainty (out-of-distribution uncertainty), we take all networks of the ensemble and combine their results into a Gaussian mixture distribution with mean $\mu_\pi(\mathbf{s})$ the variance $\sigma_\pi^2(\mathbf{s})$, shown as:

$$\begin{aligned} \mu_\pi(\mathbf{s}) &= \frac{1}{M}\sum_{i=1}^{M}\hat{\mu}_i(\mathbf{s}), \\ \sigma_\pi^2(\mathbf{s}) &= \frac{1}{M}\sum_{i=1}^{M}\hat{\sigma}_i^2(\mathbf{s}) + \left[\frac{1}{M}\sum_{i=1}^{M}\hat{\mu}_i^2(\mathbf{s}) - \mu_\pi^2(\mathbf{s})\right], \end{aligned} \tag{S1}$$

where $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ are the mean and variance of the i-th network in the ensemble. In this paper, we choose the number of networks $M = 5$. During the testing phase, we only take the mean value of the Gaussian mixture distribution as the action instead of sampling from it.

Essentially, we use imitation learning to train the human-like driving policy, which is to

minimize the discrepancy between human actions and policy output actions. To achieve uncertainty estimation, instead of mean squared error or mean absolute error, we use the negative log-likelihood to train the neural networks in the ensemble individually. The negative log-likelihood loss (NLL) for Gaussian distribution is defined as:

$$\mathcal{L}_{NLL}(\mathbf{s}, \mathbf{a}) = \frac{\log \hat{\sigma}_\theta^2(\mathbf{s})}{2} + \frac{\left(\mathbf{a} - \hat{\mu}_\theta(\mathbf{s})\right)^2}{2\hat{\sigma}_\theta^2(\mathbf{s})}, \tag{S2}$$

where $\mathbf{s}$, $\mathbf{a}$ are the state feature vector and ground-truth human action, respectively; θ is the parameter of the neural network.

In practice, for multiple samples from a mini-batch, we average over the log-likelihood of all samples to get the mean negative log-likelihood loss (MNLL):

$$\mathcal{L}_{MNLL} = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}_{NLL}(\mathbf{s}_i, \mathbf{a}_i). \tag{S3}$$

We use PyTorch to implement the neural networks and train them with the mean negative log-likelihood loss using the Adam optimizer. We initialize the parameters of networks in the ensemble with different random seeds, and the data is also randomly shuffled for training each network.

Supplementary Note 5-Neural Network and Simulation Details

The policy and adversary networks in this study are designed via two fully connected hidden layers, and the sizes of the both hidden layers are 256. Additionally, all activation functions in the hidden layers are rectified linear unit (ReLU) functions. Similar to the policy network, the critic networks also consist of two hidden layers with a width of 256. The world model leverages a branched architecture based on multi-layer perceptrons (MLPs) with ReLU activation and 200 hidden width. Moreover, we use 5 different initialized world models to construct the ensemble world model.

Simulation-based training and testing processes are implemented with the SUMO platform to test the performance of the proposed decision making method for autonomous vehicles. SUMO is leveraged to create three stochastic mixed traffic flows with different densities in the scenarios (a)-(e) and the three cut-in scenarios. The speed control and lane changing policies of the social vehicles are determined via the intelligent driving model (IDM) of SUMO. In the scenarios (a)-(d), the maximum traffic speed is 15 m/s. Moreover, in the scenarios (e) and the cut-in scenarios, the maximum traffic speed is set to 30 m/s. The longitudinal acceleration of the vehicle is limited between -7.6 m/s^2 and 7.6 m/s^2.

All model training and testing are performed on a single computer with a 2.90-GHz 12 core Intel i9-8950HK CPU.

Supplementary Note 6-Human-in-the-Loop Experimental Details

The experiment is conducted on a human-in-the-loop platform, as shown in Fig. 4(b) of the article. By using CARLA simulation software, the platform can specify sensor suites in a flexible way, model the environment with high fidelity, and freely control the autonomous driving modules. The hardware platform includes a workstation with an NVIDIA RTX 3080 GPU, three heads-up displays, a Logitech G29 steering wheel, a pedal, and a driver seat. With this system, human drivers can observe the in-vehicle view in real time, simulating real driving situations.

This experiment mainly aims to compare the performance of AI agents with that of human drivers. Additionally we attempt to examine whether the component used to simulate the amygdala in the proposed framework is important for achieving good performance.

These scenarios are categorized into three levels of aggression: normal (scene-0), aggressive (scene-1), and extremely aggressive (scene-2). The right cut-in vehicle always spawns from the rear and has a greater cruising velocity than the ego vehicle (by 20km/h). As long as the cut-in vehicle surpasses the ego vehicle by 2m (that is, its rear end exceeds the front end of the ego vehicle), a cut-in intention will be generated. The aggressiveness of the cut-in vehicle is manifested differently in its hesitation times and its longitudinal distances to the maneuver endpoint. The hesitation time is defined as maintaining the original velocity and not initiating any lane changes, and the maneuver endpoint is the longitudinal position at which the cut-in vehicle completes its lane change. Specifically, in the normal cut-in scenario (i.e., scene-0), the cut-in vehicle hesitates for 1s after the intention occurs, providing the ego vehicle with ample time to identify the intention. The maneuver endpoint is 9 meters ahead of the vehicle's present longitudinal position, which allows for a relatively smooth trajectory. Then, in the aggressive cut-in scenario (i.e., scene-1), the hesitation time after the cut-in intention occurs is decreased to 0.8s, and the longitudinal distance to the maneuver endpoint is 8m. This cut-in situation is highly challenging for the ego vehicle. In the extremely aggressive cut-in scenario

(i.e., scene-2), the hesitation time is further reduced to 0.6s, and the longitudinal distance to the maneuver endpoint is 5m, which leads to an extremely reckless cut-in trajectory and is extremely hazardous to the ego vehicle.

The study protocol and consent form were approved by the Nanyang Technological University Institutional Review Board, protocol number IRB-2018-11-025. All research was performed per relevant guidelines/regulations. Informed consent was obtained from all participants.

Supplementary Figures

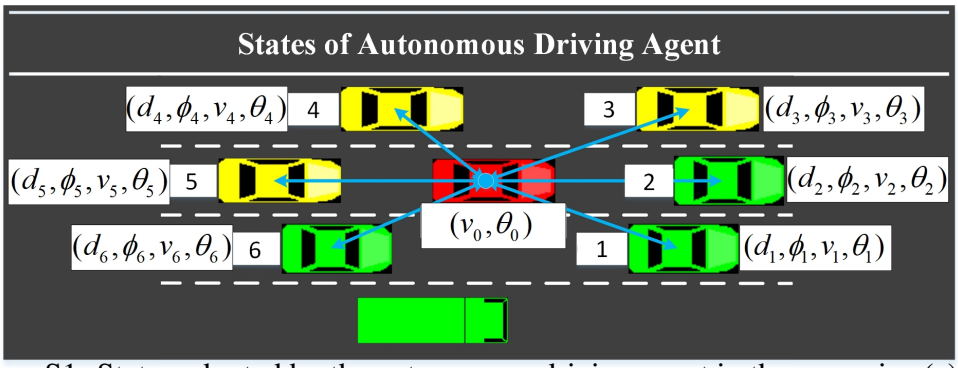


Figure S1: States adopted by the autonomous driving agent in the scenarios (a)-(b).

We adopt information from the 6 nearest vehicles within a 200-meter distance from the ego vehicle, encompassing the relative distance, orientation, speed, and velocity direction of the front, back, left-front, left-back, right-front, and right-back vehicles. Vehicle-1, vehicle-2, vehicle-3, vehicle-4, vehicle-5 and vehicle-6 denote the right-front, front, left-front, left-back, back and right-back vehicles, respectively. d_i, ϕ_i, v_i and θ_i represent the relative distance from the ego vehicle, orientation, speed and velocity direction of vehicle-i, respectively. Moreover, we incorporate the speed (v_0) and velocity direction (θ_0) of the ego vehicle, resulting in a state representation of the agent with a total of 26 dimensions.

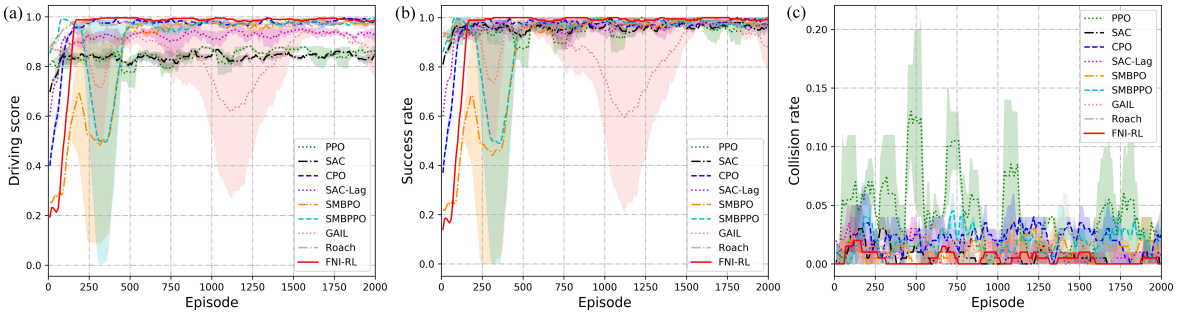


Figure S2: Training curves of different autonomous driving agents in the scenario (a). (a) Driving score; (b) Success rate; (c) Collision rate.

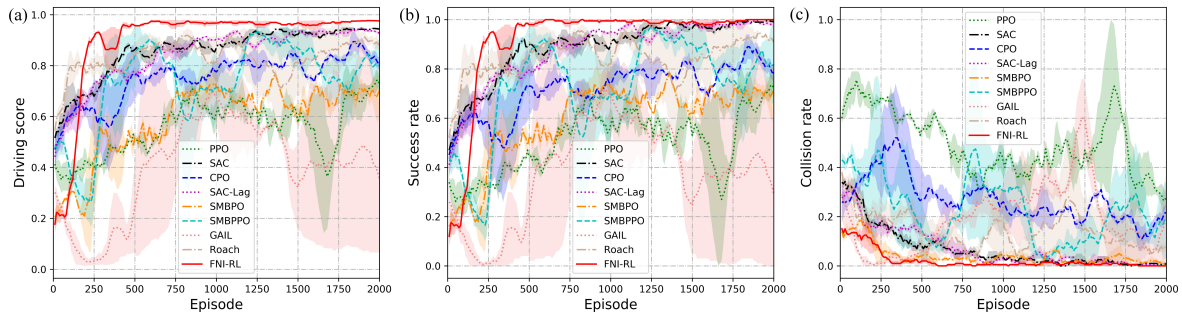


Figure S3: Training curves of different autonomous driving agents in the scenario (b). (a) Driving score; (b) Success rate; (c) Collision rate.

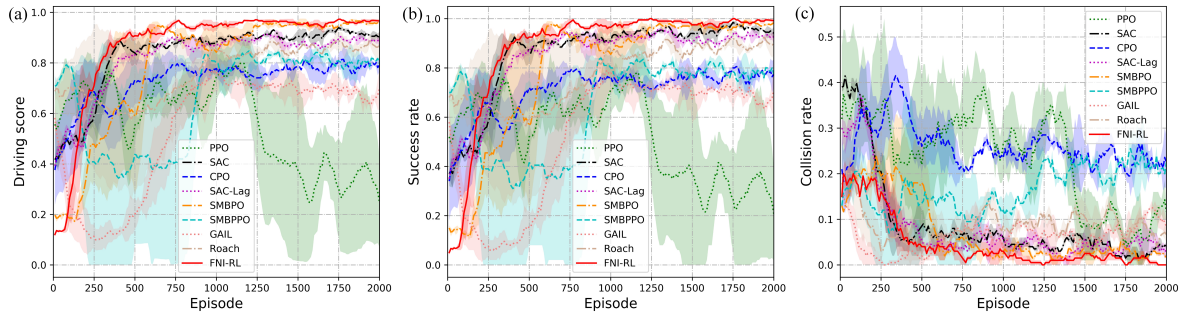


Figure S4: Training curves of different autonomous driving agents in the scenario (c). (a) Driving score; (b) Success rate; (c) Collision rate.

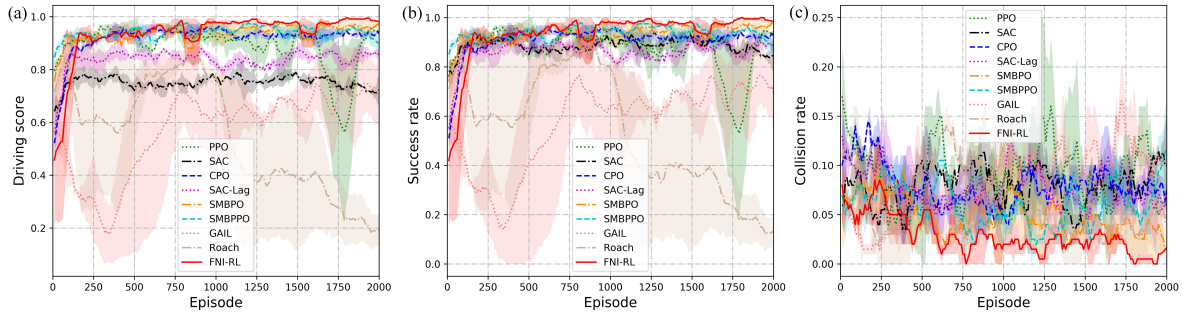


Figure S5: Training curves of different autonomous driving agents in the scenario (d). (a) Driving score; (b) Success rate; (c) Collision rate.

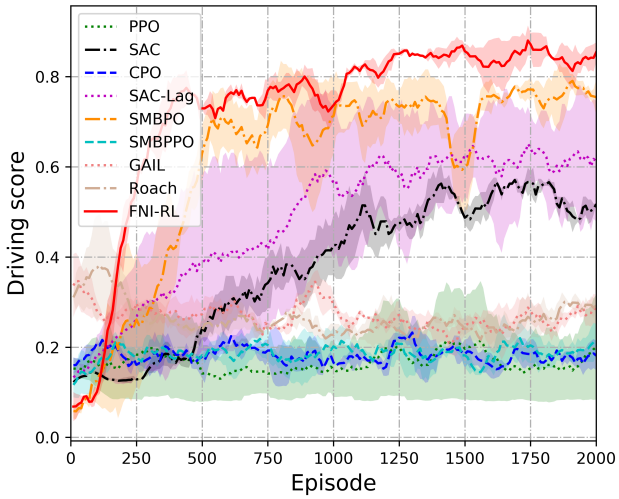


Figure S6: The training performance of the different autonomous driving agents on the long-term goal-driven navigation task based on the stochastic dynamic traffic flows.

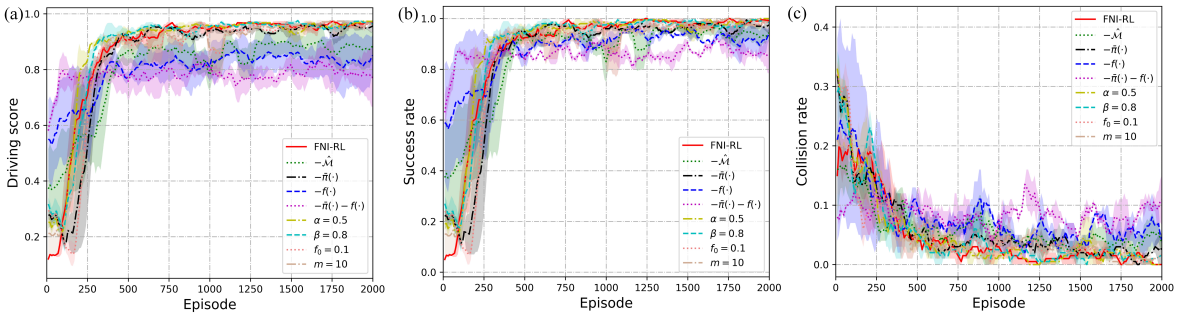


Figure S7: Training curves of different autonomous driving agents in the ablation study. (a) Driving score; (b) Success rate; (c) Collision rate.

Supplementary Tables

Table S1: Experimental configuration of the human-in-the-loop platform.

Hardware configuration	Simulation software	CARLA
	Steering wheel suit	Logitech G29
	CPU of the host computer	Intel i9-11900k
	GPU of the host computer	NVIDIA RTX 3080
	Monitoring device	Joint heads-up monitors $\times$ 3
	Other equipment	Driver seat suit
Software configuration	Control sample frequency	20Hz
	Render frequency	40Hz
	Spawned vehicle type	CARLA electric sedan
	Programming script	Python

Table S2: FNI-RL hyperparameters.

World model learning rate l_w	0.001
Adversary learning rate $l_{\bar{a}}$	0.001
Dual learning rate l_d	0.0001
Actor learning rate l_a	0.0003
Critic learning rate l_c	0.0003
Scale coefficient ρ	0.995
Discount factor γ	0.99
Constraint threshold f_0	0.5
Mixed policy's weight α	0.80
Fear model's weight β	0.90
Planning horizon m	5
Planning batch size	32
Actor-critic batch size	64
World model batch size	128