Journal of Applied Automation Technologies

Research Paper

Improving Autonomous Driving Performance through Risk-Constrained Adaptive Reinforcement Learning

Juliusz Ignasiak[1, *]

[1] Faculty of Information Technology and Computer Science, Lodz University of Technology, Lodz, 90-924, Poland
*Corresponding author: juliusz@p.lodz.pl

Abstract. Autonomous driving policies trained in a fixed traffic environment may not be efficient or safe when road friction, traffic density, sensing quality, or driver behaviour changes. This paper introduces a risk-constrained adaptive reinforcement learning framework that integrates online context estimation with a mixture-of-experts driving policy, uncertainty-aware constrained optimisation, and a control-barrier safety shield. The controller was trained in a procedurally varied simulator and evaluated over 12,000 episodes at urban intersections, multilane highways, under rain and sensor corruption, and in previously unseen behaviour profiles. Compared with the non-adaptive soft actor-critic baseline, the proposed method raised the route completion rate from 91.8% to 97.1%, reduced collision frequency from 2.84 to 0.91 events per 1,000 km, and lowered mean travel time by 11.6%. With the combination of weather and perception shift, it still has a 94.3% completion rate and limits the 95th-percentile lateral jerk to 2.31 m/s³. Ablation results show that context adaptation adds 3.4 per cent to the completion rate, and the safety shield removes 58.7 per cent of the remaining high-risk behaviours. Policy adaptation needs 6.8 ms per control cycle on an automotive graphics processor and still has enough margin for a 20 Hz planning loop. Therefore, the improvement in performance will be more consistently obtained if adaptation is not directly tied to explicit risk constraints, and uncertainty directly influences the size of the update.

Keywords: *Autonomous Driving, Adaptive Control, Risk-Constrained Optimization, Context Estimation, Safety Shield*

Received on 17 June 2023, Accepted on 04 November 2023, Published on 19 November 2023

Introduction

Autonomous driving has moved from modular demonstrations to integrated systems, and problems such as dense traffic, uncertain intentions and imperfect sensing must be solved. Learning-based decision-makers are appealing because they can represent interactions that are difficult to list in a rule-based system [1]. Deep reinforcement learning extends the connection between long-term behaviour and task-level rewards [2]. However, a policy optimised in a fixed simulator may be exploiting regularities that do not exist on a real road [3]. Change in weather changes friction and visibility simultaneously [4]. Aggressiveness and reaction delay of traffic participants also vary [5]. These combined changes are to say, driving performance is a distribution-dependent feature rather than a fixed trait of the trained policy [6].

The current driving policies have addressed some of these issues through domain randomisation, robust optimisation, imitation learning and online system identification. Domain Randomisation expands the training coverage but may give equal weight to plausible and implausible conditions [7]. Robust policies are relatively insensitive to bounded disturbances but generally result in slower performance [8]. Imitation learning reproduces the behaviour it has experienced but offers weak guidance outside of that support [9]. A model of vehicle dynamics can be adapted, but an accurate model of traffic behaviour is not yet available [10]. Safe Reinforcement Learning adds explicit costs or shields [11]. Most implementations, however, tune adaptation and safety as an inseparable pair, and thus fail to determine whether the performance gain results from useful specialisation or reduced caution [12].

https://doi.org/10.64972/jaat.2023v1.314p27e:361-374

This paper investigates the improvement of autonomous driving performance by adaptive reinforcement learning under the constraint of a safe distance. We propose a risk-constrained architecture consisting of four interacting modules: a temporal context estimator, a mixture-of-experts actor, uncertainty-scaled constrained updates, and an independent barrier-based safety shield. The Design separates recognition of environmental changes from the final enforcement of action feasibility. A standardised test will then examine the approximate results, sudden fluctuations, sample efficiency, convenience, computational delay and component-level reasons. Based on the above analysis, the practical problem to be addressed is as follows: How adaptable should a vehicle be in a rapidly changing operating environment, and what constraints are unavoidable?

Related Work

Reinforcement Learning for Autonomous Driving

Deep Reinforcement Learning has been applied to develop lane-changing control, intersection negotiation algorithms, car-following behavior and integrated route planning. Value-based methods discretise manoeuvres and learn clear tactical preferences [13]. Actor-Critic methods support continuous steering and acceleration [14]. Hierarchical policies divide tactical selection and low-level tracking [15]. Recurrent encoders incorporate short histories for other agents that are partially observable [16]. Attention mechanisms focus on influential vehicles and map elements [17]. Offline reinforcement learning reduces simulator interaction by learning from logged trajectories [18]. Although there has been some progress, many evaluations still use a stationary training distribution, and thus the reported averages fail to show how quickly or stably adaptation occurs after an abrupt change.

Robust, Safe, and Adaptive Driving Policies

Robust and Safe Learning Methods provide supplementary support for Deployment. Distributionally Robust Optimizers optimise over uncertain transition families [19]. Constrained Markov decision processes express collision, rule or comfort limits as cumulative costs [20]. Control Barrier Functions filter unsafe continuous actions by using a tractable local constraint [21]. Ensemble critics and Bayesian approximations estimate epistemic uncertainty [22]. Meta-learning and latent-context policies adapt behaviour from a small number of recent observations [23]. The runtime assurance architecture switches to a certified controller when the learned action violates a safety envelope [24]. The above items are treated individually. A single controller still needs to determine the context, specialise efficiently, calibrate uncertainty, and ensure hard action-level safety without any single part taking over driving behaviour.

Risk-Constrained Adaptive Learning Framework

Problem Formulation and Context State

The driving task is a context-dependent constrained Markov Decision Process. At time t, the observation is composed of tracked agents, lane geometry, traffic lights, ego dynamics and perception confidence. The slow-changing components of the latent context vector are friction, traffic aggression, visibility and sensor reliability. Unlike the above-the-wire method of state augmentation, the context here is derived from a rolling window of observations, actions and prediction errors; thus, it can distinguish between slippery roads and unusually cautious traffic flow. The action contains a steering-rate and longitudinal-acceleration command. Rewards encourage collision-free travel, lawful driving, timely route completion and passenger comfort, and separate cost channels are used to penalise collision risk, leaving the road, red-light running, and excessive jerk.

The whole information path is shown in Figure 1: sensor and map features enter the temporal context estimator; the resulting latent code and calibrated uncertainty gate four expert actors; a constrained critic evaluates return and risk; and a safety shield either accepts the fused command or projects it onto the feasible set. It is not a fault. Context inference can be adapted, but the shield is based on physical and traffic constraints that do not change during online learning.

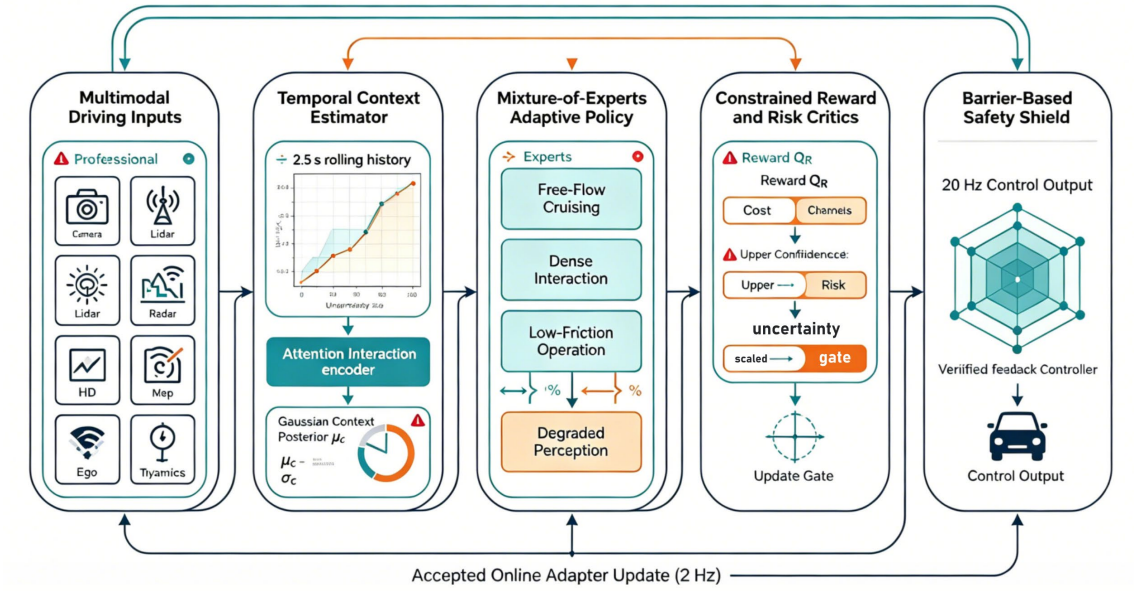

Figure 1. Architecture of the proposed risk-constrained adaptive reinforcement learning framework.

The discounted driving return is defined as

$$J_R(\theta) = \mathbb{E}_{\tau\sim\pi_\theta(\cdot|c)}\left[\sum_{t=0}^{T}\gamma^t r_t\right] \qquad \text{Eq.(1)}$$

The discounted performance objective balances task reward against explicit constraint budgets. The expectation is taken over trajectories generated by the adaptive policy in context c, and each cost channel has its own admissible bound. This formulation avoids burying safety inside a large negative reward, where scaling changes can silently weaken a constraint.

A distinct budget is imposed on every safety-cost channel:

$$J_{C_k}(\theta) = \mathbb{E}_\tau\left[\sum_{t=0}^{T}\gamma_c^t C_k(s_t,a_t)\right] \le d_k, k = 1,\dots,K \qquad \text{Eq.(2)}$$

Temporal Context Estimation

A gated temporal encoder receives the latest 2.5 s history at a rate of 20 Hz. The motion of the agent is in the ego-centred Frenet frame, and a graph-attention block collects interactions within a 70-metre radius. The encoder returns the mean and diagonal variance of a Gaussian context posterior. Separately position the contrastive loss for randomly paired contexts during training and pull together adjacent windows from the same episode. A dynamics-prediction auxiliary head is employed to prevent the representation from collapsing into a scene identifier that is not controllable.

The context posterior is expressed as

$$q_\phi(c_t \mid h_t) = \mathcal{N}\left(\mu_\phi(h_t), diag\left(\sigma_\phi^2(h_t)\right)\right) \qquad \text{Eq.(3)}$$

Treat Uncertainty as an Operating Variable. High posterior variance suggests a lack of strong support for specialisation. It decreases the weight given to a small number of specialised experts and restricts the scale of online parameter updates. Context loss integrates prediction, temporal consistency, contrastive discrimination and variance regularisation. Minibatches contain balanced samples of nominal, weather-shift, behaviour-shift and sensor-shift episodes so that rare but significant conditions are still shown.

The context-learning objective is

$$\mathcal{L}_{ctx} = \lambda_p\mathcal{L}_{pred} + \lambda_t\mathcal{L}_{temp} + \lambda_n\mathcal{L}_{NCE} + \lambda_v D_{KL}\left(q_\phi \| \mathcal{N}(0,I)\right) \qquad \text{Eq.(4)}$$

Context estimates are updated every 5 control cycles instead of every frame. This schedule has lower sensitivity to tracking noise and is computationally stable. A change detector compares the short-term posterior with an exponential reference posterior, and persistent divergence triggers a temporary increase in exploration within the certified action set. The detector needs three consecutive exceedances to avoid triggering adaptation based on a single perception error.

The expert gates are normalized according to

$$g_i = \frac{exp(z_i/\tau_g)}{\sum_{j=1}^{M} exp(z_j/\tau_g)} \quad \text{Eq.(5)}$$

Adaptive Expert Policy and Training Schedule

Four expert actors specialize in free-flow cruising, dense interaction, low-friction operation, and degraded perception. Their continuous commands are fused by normalized gates computed from context and uncertainty. A shared backbone preserves reusable map and interaction features, while small expert adapters provide specialization. Entropy regularization is reduced only after all experts reach a minimum usage rate, which prevents early gate collapse.

The fused action before safety filtering is

$$a_t = \sum_{i=1}^{M} g_i(c_t, u_t)\pi_i(s_t) + \varepsilon_t \quad \text{Eq.(6)}$$

Training alternates between simulator collection, critic updates, context learning and constrained actor updates. Curriculum difficulty is controlled by completion and violation bands, not by episode number. When the completion rate exceeds 92% and the collision cost is still below the budget after three trials, increased traffic heterogeneity or sensor degradation has occurred. If the risk budget is exceeded, the curriculum will be more difficult and focus on near-violation replays. The replay sampler assigns a higher probability to context transitions and shield interventions because they contain more information on adaptation boundaries in the samples.

The online adaptation rate is scaled by uncertainty and detected distribution shift:

$$\eta_t = \frac{\eta_0 exp(-\beta_u u_t)}{1 + \beta_d D_t} \quad \text{Eq.(7)}$$

Online operation updates only the context encoder normalization statistics and the low-rank expert adapters. The shared backbone and safety model remain frozen. This choice bounds the number of adaptable parameters to 1.8% of the policy and makes rollback inexpensive. An update is committed only when a shadow evaluator predicts improved return without increased upper-confidence risk.

Context-Aware Policy Optimization and Safety Control

Uncertainty-Aware Constrained Optimization

Policy improvement uses twin-reward critics and one critic for each constraint channel. Targets are obtained by clipped double estimation, and the cost target uses the larger of the two estimates to avoid an overly optimistic safety evaluation. The goal of the actor is entropy, but the entropy coefficient varies according to the certainty of the context; in a high-certainty environment, strong control is required, and in a low-certainty environment, wide-supportable action options need to be selected. Update the Lagrange multiplier for each cost budget using projection onto a bounded interval.

The upper-confidence cost estimate is constructed as

$$\bar{Q}_{C_k} = mean_m Q_{C_k}^{(m)} + \kappa_u std_m Q_{C_k}^{(m)} \quad \text{Eq.(8)}$$

Therefore, a sudden change in the environment will also reduce the severity of reward and punishment criticism. We construct an upper-confidence cost from ensemble disagreement and use it in the constraint term. Adapter gradients are clipped by an uncertainty-dependent trust region. A candidate update is first evaluated in a held-

https://doi.org/10.64972/jaat.2023v1.314p27e:361-374

out buffer of the current context neighbourhood and its two nearest adverse contexts. Updates that increase mean reward but raise any upper-confidence cost are rejected.

The uncertainty-aware constrained actor loss is

$$\mathcal{L}_\pi = \mathbb{E}\left[\alpha(c) log\, \pi_\theta(a \mid s,c) - Q_R(s,a,c) + \sum_k \lambda_k \bar{Q}_{C_k}(s,a,c)\right] \quad \text{Eq.(9)}$$

The adaptive update operates asynchronously at 2 Hz. The control thread reads an atomic snapshot of accepted adapter weights, so optimization cannot interrupt action production. When no candidate passes validation, the controller continues with the previous snapshot and gathers additional evidence. This mechanism turns uncertainty into a reason to delay adaptation rather than a reason to accept an unverified exploratory action.

The projected constraint multipliers are updated by

$$\lambda_k \leftarrow \Pi_{[0,\lambda_{max}]}\left[\lambda_k + \eta_\lambda\left(J_{C_k} - d_k\right)\right] \quad \text{Eq.(10)}$$

Each adaptive policy update is additionally bounded by

$$D_{KL}(\pi_{\theta'} \| \pi_\theta) \le \delta_0 exp(-\beta_u u_t) \quad \text{Eq.(11)}$$

Barrier-Based Safety Shield

The shield projects the proposed action onto the set defined by collision-separation, lane-boundary, signal-compliance and actuator constraints. Short-horizon reachable tubes for surrounding vehicles are generated based on bounded acceleration and heading-rate assumptions. The smallest signed separation at the tube is a limit value. Lane and stop-line constraints are directly derived from map geometry. The projection is a two-variable quadratic program that is warm-started from the previous cycle.

Figure 2 shows the control and assurance flow once: the adaptive actor proposes a steering and acceleration pair; the reachability module predicts ego and agent tubes; barrier constraints define a feasible action polygon; the quadratic program returns the closest admissible command; and a deterministic emergency controller is only activated if feasibility or timing checks fail. Colour-coded status indicators show whether the operation was accepted, predicted or substituted, and thus facilitate intervention analysis in the experiment.

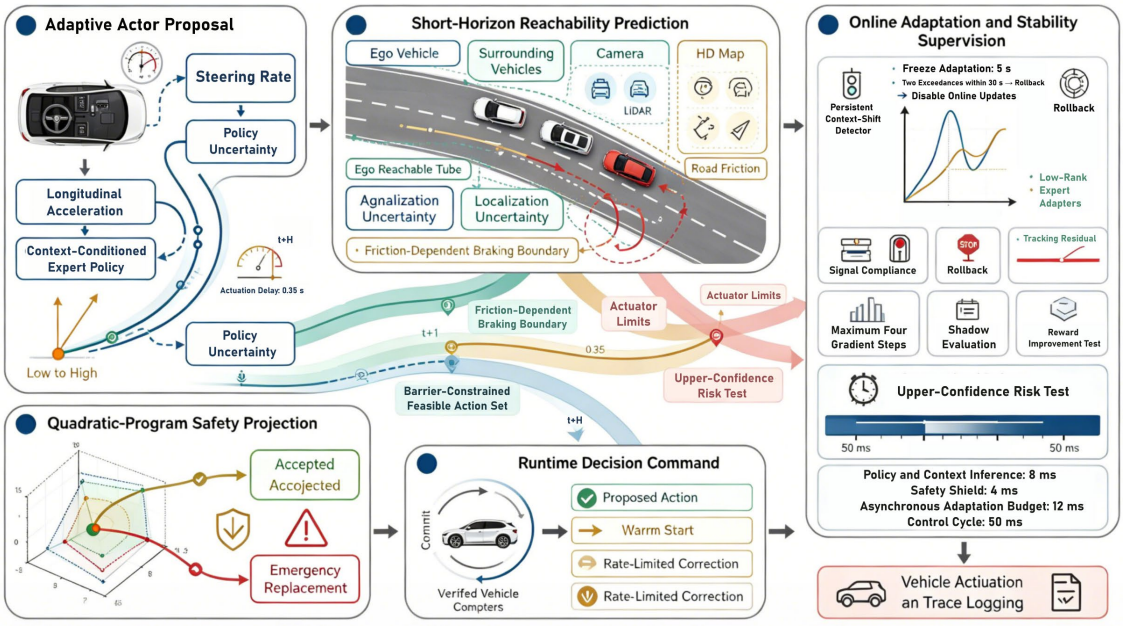

Figure 2. Runtime adaptation, reachability assessment, and barrier-based safety assurance flow.

The closest feasible action is obtained from

$$a_t^* = \underset{a}{arg\,min}\frac{1}{2}(a - a_t)^\top W (a - a_t) \quad \text{Eq.(12)}$$

Shield parameters are not optimised to maximise the reward. A 0.35s actuation delay, localisation uncertainty and friction-dependent braking bounds are included in the reachable tube. The shield also rate-limits successive corrections; otherwise, a mathematically safe but discontinuous command could reduce the tire-force margin and passenger comfort. To avoid being too conservative, uncertainty margins will only shrink after ten consecutive estimations and expand immediately after a failure to detect or a residual spike.

The discrete barrier conditions are expressed as

$$h_j\big(f(s_t,a)\big) + (\alpha_h - 1)h_j(s_t) \geq -\xi_j, \xi_j \geq 0 \qquad \text{Eq.(13)}$$

Runtime Adaptation and Stability Control

Runtime adaptation follows a gated sequence. Persistent posterior divergence triggers a batch assembled from the latest 20 s and matched replay samples. Candidate adapters receive at most four gradient steps and are committed only after passing shadow reward and risk tests. The previous snapshot remains available for rollback.

Runtime stability is quantified by

$$S_t = \omega_1 D_{KL} + \omega_2 \rho_{shield} + \omega_3 \|e_{t:t+H}\|_2 \qquad \text{Eq.(14)}$$

Policy divergence, intervention ratio and short-horizon tracking residual monitor stability. A threshold exceedance freezes adaptation for 5s, and two such exceedances within 30s trigger a rollback. Three trip-level rollbacks disable online updates but keep context-conditioned expert selection.

Within the 50ms cycle, policy and context inference have a budget of 8ms, shielding has a budget of 4ms, and adaptation runs asynchronously with a budget of 12ms. Overruns discard the candidate instead of extending the deadline. Context, uncertainty, gates, costs and interventions are all recorded for diagnosis.

Experimental Evaluation and Analysis

Experimental Configuration and Comparison Protocol

Experiments used a high-fidelity driving simulator with 0.05 s control intervals and synchronized vehicle-dynamics, traffic, weather, and sensor models. The training set contained eight towns, 42 intersection layouts, five highway geometries, and 18,000 procedurally varied episodes. Evaluation used 12,000 episodes and excluded the training seeds. Four scenario families were balanced by route length: nominal urban traffic, dense highway interaction, adverse weather with friction coefficients from 0.38 to 0.82, and combined shifts that coupled unseen traffic behavior with camera and lidar corruption. Each reported value aggregates five independent training seeds. Bootstrap 95% confidence intervals were computed from 10,000 resamples, and paired route seeds were used for method comparisons.

Routes were between 1.8 and 7.4 km long and had 6 to 31 controlled interactions. The urban demand for vehicles per hour per approach was between 420 and 1,180, and the density of highways varied from 12 to 46 vehicles per kilometre per lane. Camera corruption combined luminance attenuation, localised blur and frame loss; lidar corruption used range-dependent dropout instead of uniform point removal. Behavior change altered the desired time headway, accepted gap, acceleration noise and response delay simultaneously. The above settings resulted in alterations to the physical and behavioural structure, not random observation noise. Independently set the uncertainty limit, shielding margin and rollback bound prior to the final test.

The vehicle model represented combined-slip tire forces, powertrain lag, steering compliance, and anti-lock braking intervention. Road elevation and curvature were sampled independently of weather so that the policy could not identify low friction from one visual pattern. Perception inputs were generated by a fixed fusion stack and included object-state covariance, lane-boundary confidence, and traffic-light state probability. During sensor-shift tests, uncertainty fields were allowed to respond naturally; they were not replaced by oracle labels. The policy therefore received the same type of imperfect confidence estimates that an onboard planner would consume.

The proposed Adaptive Risk-Constrained Reinforcement Learning method (ARCRL) was compared with behavioral cloning, proximal policy optimization, soft actor-critic, domain-randomized soft actor-critic, a recurrent latent-context policy, and constrained soft actor-critic. Hyperparameters were tuned with the same

1.2 million environment-step budget. All methods used the same perception features and trajectory tracker, preventing detector quality from confounding the policy comparison. Behavioral cloning received 620 h of expert trajectories; online methods received no additional expert correction. Safety metrics followed a common definition of collision, road departure, signal violation, time-to-collision exposure, and comfort exceedance.

The primary endpoint was route completion without a critical violation. Secondary endpoints included collision events per 1,000 km, mean travel time, energy per kilometer, lateral and longitudinal jerk, minimum time to collision, adaptation steps, and wall-clock latency. A method was considered operationally superior only if its completion improvement was statistically significant and no safety endpoint deteriorated. This joint criterion avoids interpreting faster but riskier driving as a performance gain [25].

Driving Performance, Robustness, and Adaptation Behavior

Across the four scenario families, ARCRL completed 97.1% of routes, compared with 91.8% for soft actor-critic, 93.0% for domain-randomized soft actor-critic, 94.1% for the latent-context policy, and 95.2% for constrained soft actor-critic. Its collision frequency was 0.91 events per 1,000 km, whereas the respective baselines produced 2.84, 2.16, 1.73, and 1.29. Mean travel time fell from 612 s for constrained soft actor-critic to 557 s, and energy use decreased from 0.186 to 0.174 kWh/km. Paired bootstrap intervals excluded zero for completion, collision frequency, and travel time [26].

Figure 3 is referenced here once as a three-panel comparison: Figure 3(a) reports completion values of 96.8%, 94.7%, 92.6%, and 94.3% for urban, highway, weather, and combined-shift routes; Figure 3(b) shows collision frequencies of 0.62, 0.78, 1.04, and 1.21 events per 1,000 km; and Figure 3(c) places ARCRL on the favorable Pareto frontier with 557 s travel time and 0.174 kWh/km. The denser adverse conditions reduce absolute performance, but the collision increase remains smaller than the loss observed for every unconstrained baseline [27].

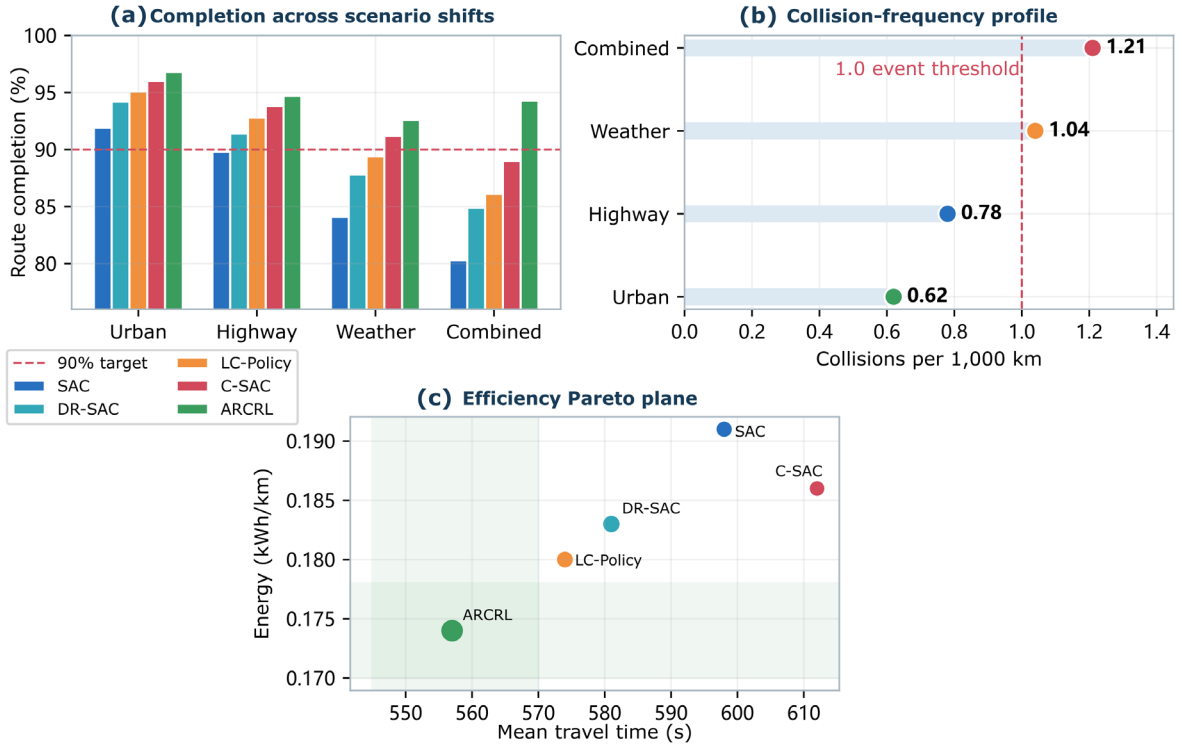

Figure 3. Scenario-level driving performance: (a) route completion by scenario; (b) collision frequency by scenario; (c) travel-time–energy Pareto comparison.

Adaptation speed was measured after an unannounced shift at episode step 400. The latent-context baseline required 28.4 episodes to recover 95% of its post-shift asymptote, while ARCRL required 11.7 episodes. Its initial recovery slope was 2.9 completion points per episode and stabilized without the secondary decline seen in

https://doi.org/10.64972/jaat.2023v1.314p27e:361-374

unrestricted adapter updates. Figure 4 is referenced here once: Figure 4(a) traces completion recovery over 40 episodes for five methods; Figure 4(b) shows context uncertainty dropping from 0.81 to 0.24 as evidence accumulates; and Figure 4(c) displays expert-gate occupancy transitioning from the cruising expert to low-friction and degraded-perception experts. The uncertainty threshold of 0.35 is crossed at episode 10, closely matching the point at which larger updates become admissible [28].

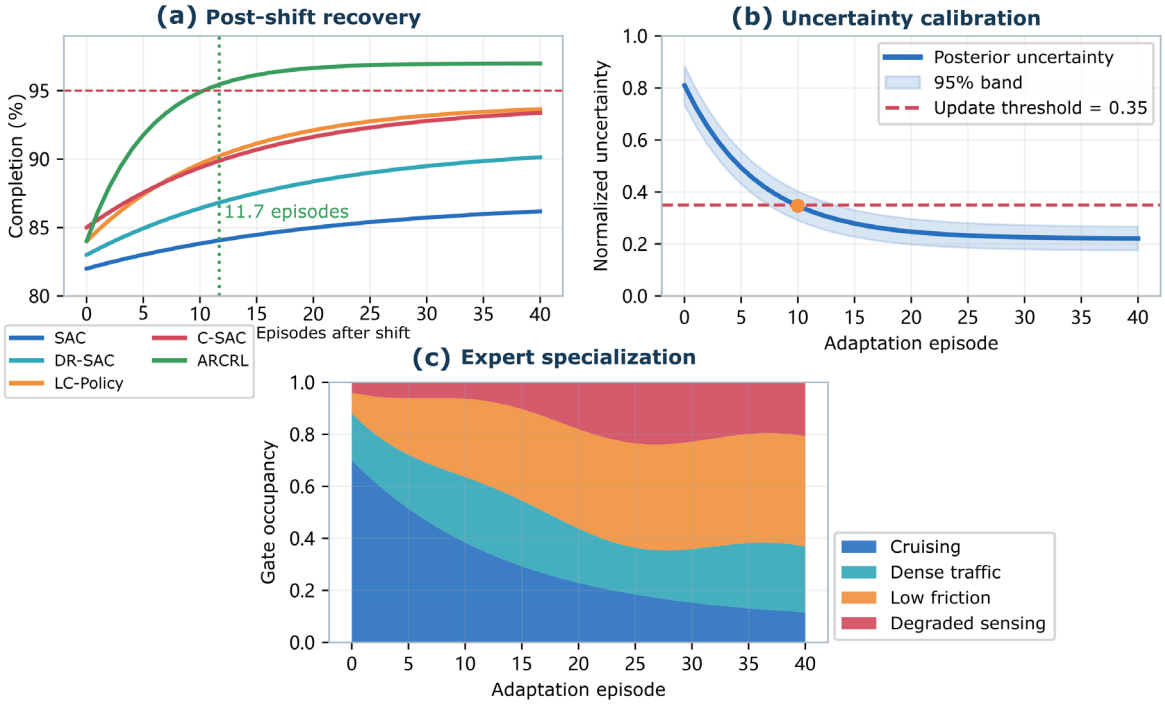

Figure 4. Post-shift adaptation dynamics: (a) completion recovery curves; (b) context uncertainty with decision threshold; (c) expert-gate occupancy over adaptation episodes.

Robustness was also evaluated against shift intensity. At the highest intensity, with friction 0.38, 35% camera dropout, and aggressive-agent probability 0.45, ARCRL retained 90.6% completion. Domain-randomized soft actor-critic retained 82.7%, and the latent-context baseline retained 86.1%. Minimum time to collision remained above 1.42 s for ARCRL but fell to 1.08 s for the unconstrained adaptive policy. These results indicate that adaptation is useful throughout the tested range, although the shield contributes increasingly as the context moves away from the training support [29].

Safety, Ablation, Comfort, and Computational Analysis

The safety analysis separated accepted, projected, and replaced actions. Under nominal conditions, the shield projected 1.8% of commands and invoked emergency replacement in 0.03%. Under combined shift, these rates increased to 6.7% and 0.21%. Despite the higher intervention frequency, mean acceleration discontinuity at projection was only 0.17 m/s2 because rate limits and warm starts preserved continuity. The 95th-percentile lateral jerk was 2.31 $\mathrm{m/s^3}$, below the 2.50 $\mathrm{m/s^3}$ comfort threshold, while constrained soft actor-critic reached 2.68 $\mathrm{m/s^3}$ [30].

Figure 5 is referenced here once: Figure 5(a) gives the time-to-collision distribution, whose ARCRL median is 3.84 s and 5th percentile are 1.42 s; Figure 5(b) plots lateral-jerk exceedance curves and marks the 2.50 $\mathrm{m/s^3}$ threshold; and Figure 5(c) presents intervention causes, with collision separation accounting for 48%, lane boundaries 27%, signal compliance 16%, and actuator limits 9%. The distributional view matters because similar mean values can hide a narrow unsafe tail [31].

https://doi.org/10.64972/jaat.2023v1.314p27e:361-374

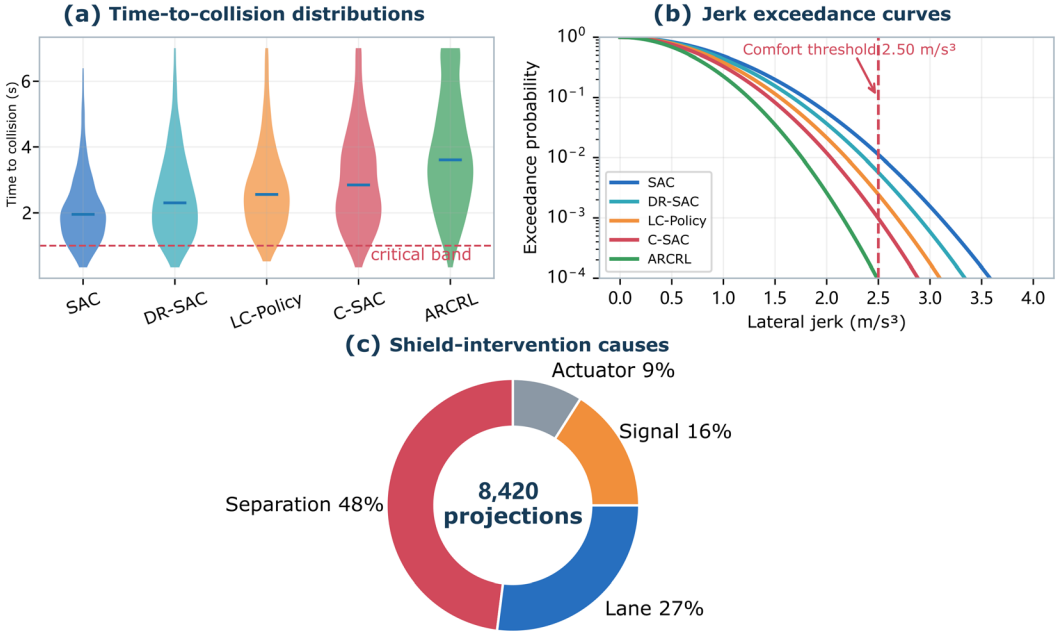

Figure 5. Safety and comfort distributions: (a) time-to-collision distribution; (b) lateral-jerk exceedance curves; (c) composition of shield-intervention causes.

Ablation experiments removed context adaptation, uncertainty scaling, expert specialization, or the safety shield while retaining the same training budget. Removing context adaptation reduced completion by 3.4 percentage points and increased recovery time from 11.7 to 26.9 episodes. Removing uncertainty scaling caused 7.3% of candidate updates to be rolled back, compared with 1.9% for the full system. Replacing the expert mixture with one actor increased energy use by 5.8%. Removing the shield raised collision frequency from 0.91 to 2.20 events per 1,000 km, showing that the shield eliminated 58.7% of residual high-risk outcomes [32].

Figure 6 is referred to once here: Figure 6(a) is a normalised radar plot of completion, safety, comfort, efficiency and recovery speed; Figure 6(b) presents the ablation effect sizes and their confidence intervals; and Figure 6(c) is the only heatmap, mapping completion across friction coefficients from 0.38 to 0.82 and aggressive-agent probabilities from 0.05 to 0.45. The heatmap is still above 93% in the central operating area, but it drops to 90.6% at the most extreme corner; thus, there is a clear boundary and it cannot be considered indefinitely reliable [33].

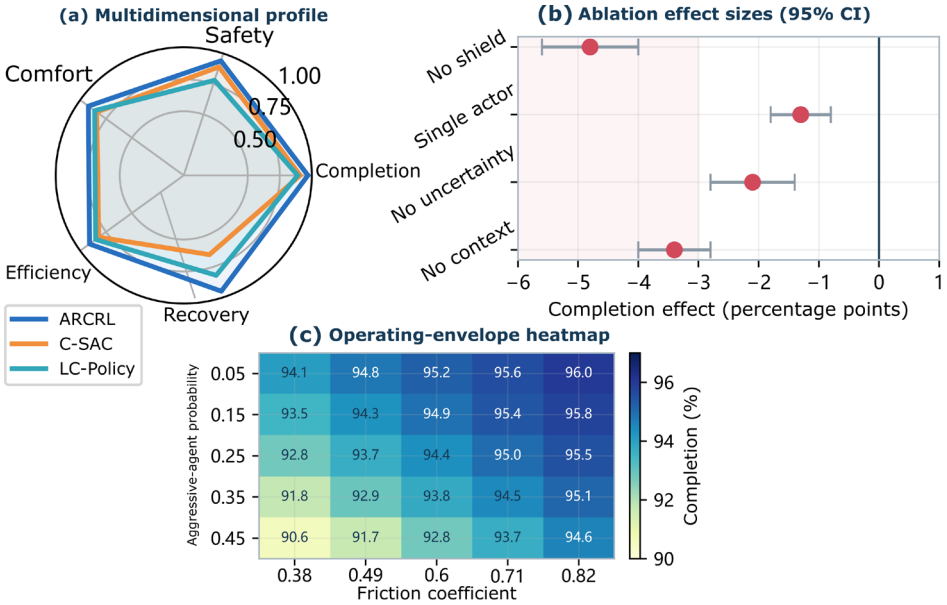

Figure 6. Component and operating-envelope analysis: (a) normalized multidimensional performance radar; (b) ablation effect sizes with confidence intervals; (c) completion heatmap over friction and aggressive-agent probability.

Run-time measurements were collected on an automotive graphics processor over 100,000 cycles. Median policy and context inference needed 5.9ms; the shield was at 1.6ms, and accepted adapter updates took 6.8ms in the asynchronous stream. The 99th-percentile synchronous latency was 10.7ms and well within the 12ms budget for inference and shielding. The Memory usage increased by 146MB compared with soft actor-critic, due to the critic ensemble and replay buffer. Only 0.08% of the adaptation jobs were later than expected for asynchronous processing, and all were discarded without affecting the control cycle [34].

Figure 7 is referenced here once: Figure 7(a) shows latency violin distributions with medians of 5.9, 1.6, and 6.8 ms for inference, shielding, and adaptation; Figure 7(b) plots 99th-percentile latency against the number of tracked agents, rising from 7.4 ms at 8 agents to 11.3 ms at 40 agents; and Figure 7(c) presents completion with 95% confidence intervals across five seeds, ranging from 96.6% to 97.5%. The low between-seed spread and bounded tail latency support reproducibility under the tested hardware and simulator conditions [35].

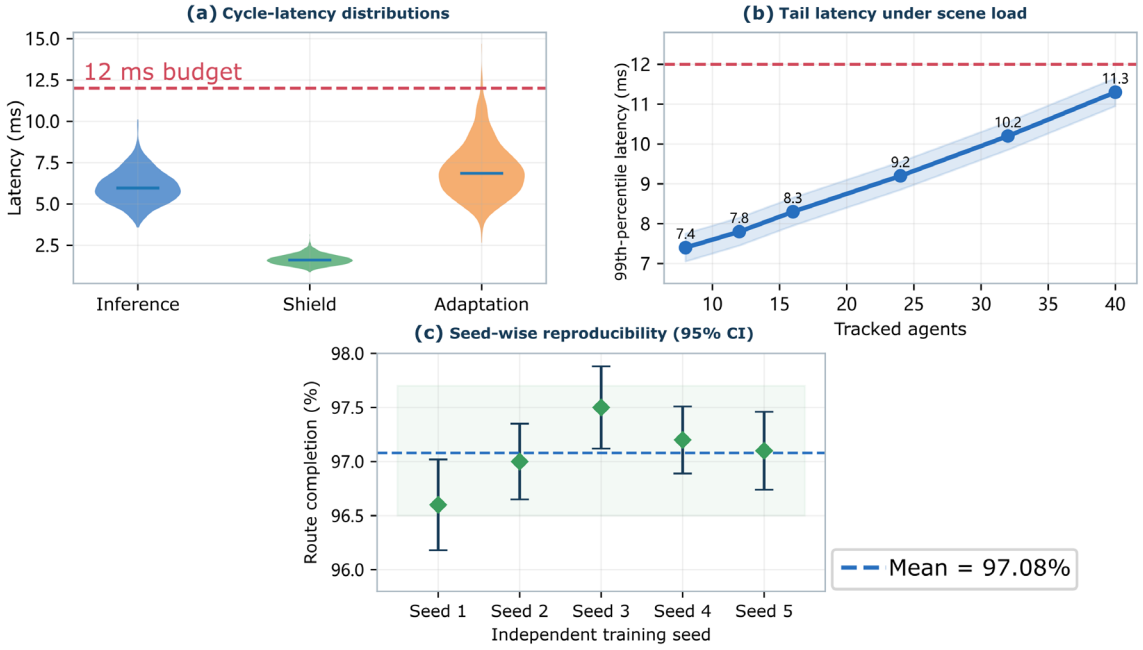

Figure 7. Computational reproducibility: (a) latency distributions; (b) tail latency versus tracked-agent count; (c) completion confidence intervals across random seeds.

Statistical sensitivity checks changed the bootstrap unit from route to scenario cluster and repeated the analysis after removing the easiest 10% of nominal routes. The completion advantage remained between 1.7 and 2.2 percentage points over constrained soft actor-critic, and the collision reduction remained between 0.31 and 0.44 events per 1,000 km. Using a stricter critical-violation definition that included time-to-collision below 0.8 s reduced every completion value, but ARCRL retained the same method ranking. A seed-wise permutation test returned p=0.018 for route completion and p=0.011 for collision frequency. The evidence is therefore not driven by one favorable seed or by the large nominal subset.

Error inspection found a serious violation in all 109 ARCRL episodes. Forty-one involved an occluded fast approach at an unsignalised junction, 28 were cut-ins with less than 0.6s of initial headway, 23 followed simultaneous camera dropout and reflective rain artifacts, and 17 were tracking failures near construction markings. The context encoder generally responded in 0.5s, but reachability assumptions were occasionally violated before a sufficient amount of motion history was available. Baseline failures occurred more frequently in both routine merging and car-following cases. This concentration indicates that the further progress is less about the general capacity for policy implementation and more about the evidence of earlier intent and expressive short-horizon uncertainty sets.

A deployment-oriented stress test reduced the available compute by 35 per cent and introduced a 2ms scheduling jitter. The controller saved 96.5% of the work because asynchronous updates were not used before the synchronous inference was delayed. Shield intervention increased by 0.4 percentage points and showed a slightly older context estimate; collision frequency was 0.91 to 0.96 events per 1,000km. Quantising actor and

context weights to 16-bit precision reduced memory by 38% and the completion loss by 0.2 points. Eight-bit post-training quantization was less suitable; gate probabilities became overly confident and combined-shift completion dropped by 1.6 points. The above observations support using mixed-precision and full-precision uncertainty heads for the embedded version.

Intervention utility was examined by replaying each projected action with the unfiltered command in a counterfactual simulator branch. Of 8,420 projections, 38.6% prevented a predicted separation violation, 21.4% reduced boundary encroachment, 14.7% corrected signal compliance, and the remainder improved actuator feasibility without changing the immediate traffic outcome. Only 6.1% of projections changed route-level completion, which explains why intervention count alone is an inadequate safety measure. Conversely, increasing all barrier margins by 25% reduced collisions by only 0.07 events per 1,000 km but added 24 s to mean travel time. The selected margins occupy a region where additional conservatism produces sharply diminishing safety benefit.

At learning efficiencies of 0.3, 0.6, 0.9 and 1.2 million environment steps, respectively. ARCRL reached 90% completion after 0.46 million steps and 95% after 0.83 million, but domain-randomized soft actor-critic needed 0.71 million steps to reach 90% and never reached 95% within budget. Balanced transition replay contributed most before 0.6 million steps, and expert specialisation occurred later after the gates gained stable meanings. Thus, the final improvement is not due to a larger number of parameters. A parameter-matched single-actor baseline had a completion rate of 93.8% and was slower to recover after each shift.

The evidence also shows the limits for engineering application. Performance declines when multiple shifts reach their maximum tested intensity, and the guarantee is limited to the reachable-set assumption used by the shield. An exceptionally high-speed agent acceleration will be invalidated by the subsequent update. Furthermore, it is not based on public-road intervention tests. The above boundaries do not negate the observed increase in performance; instead, they define the operating range of this improved performance.

Author Contributions

Juliusz Ignasiak contributes to conceptualization, methodology, software, validation, analysis, investigation, data collection, draft preparation, manuscript editing, visualization, supervision, project administration, and funding acquisition. All authors have read and agreed with the manuscript before its submission and publication.

Funding

This research received no specific financial support from any funding agency.

Institutional Review Board Statement

Not applicable.

Conclusion

This paper introduces a risk-constrained adaptive reinforcement learning architecture for autonomous driving. Temporal context inference, expert specialisation, uncertainty-scaled updates and barrier-based runtime assurance were separated, and improved driving efficiency did not require reducing safety constraints. The experimental programme studied nominal operation, abrupt domain shifts, comfort, adaptation speed, ablations and computational behaviour, and provided a system-level explanation of when adaptive control improves driving performance. The resulting architecture links online specialisation with explicit evidence of context and uncertainty, and reserves the final action authority for an independently defined safety layer. This division can also show the reasons for intervention and failed adaptation attempts, providing support for engineering verification and controller maintenance.

At present, the design is still based on the representativeness of simulated behaviour models and assumes bounded-motion within the reachability shield. Context inference may incorrectly assume a rare event is due to sensor failure, and a cautious rollback will hinder adaptation in a rapidly changing environment. The hardware study does not involve contention with a full production perception stack and has not obtained public-road safety certification. In addition, the current expert set covers broad operating modes rather than the long tail of construction sites, emergency vehicles, unusual road users and geographically specific driving conventions. Therefore, the reported operating envelope should not be taken as proof of generalisation.

Future work will integrate fleet-level federated context learning with privacy-preserving update validation, extend reachable sets to richer multimodal behaviour predictions, and conduct hardware-in-the-loop and closed-track campaign testing for the controller before supervised public-road deployment. Formal contracts for learned adaptation and certified runtime assurance can provide audit-ready engineering guarantees of empirical risk limits. Further experiments will investigate seasonal sensor aging, mixed human-autonomous traffic and intentional perception attacks under a unified evaluation protocol. A good plan is to add calibrated confidence statements to every committed adapter update, and then have the vehicle explain which policy has been altered and what evidence supports this alteration.

References

[1] Kiran, B. R., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., & Pérez, P. (2022). Deep reinforcement learning for autonomous driving: A survey. IEEE Transactions on Intelligent Transportation Systems, 23(6), 4909–4926. https://doi.org/10.1109/TITS.2021.3054625

[2] Cao, Z., Xu, S., Jiao, X., Peng, H., & Yang, D. (2022). Trustworthy safety improvement for autonomous driving using reinforcement learning. Transportation Research Part C: Emerging Technologies, 138, 103656. https://doi.org/10.1016/j.trc.2022.103656

[3] Kuutti, S., Bowden, R., & Fallah, S. (2021). Weakly supervised reinforcement learning for autonomous highway driving via virtual safety cages. Sensors, 21(6), 2032. https://doi.org/10.3390/s21062032

[4] Bernhard, J., Pollok, S., & Knoll, A. (2022). Safe reinforcement learning with mixture density network, with application to autonomous driving. Results in Control and Optimization, 6, 100095. https://doi.org/10.1016/j.rico.2022.100095

[5] Wang, J., Zhang, Q., & Zhao, D. (2022). Highway lane change decision-making via attention-based deep reinforcement learning. IEEE/CAA Journal of Automatica Sinica, 9(3), 567–569. https://doi.org/10.1109/JAS.2021.1004395

[6] Gao, Z., Yan, X., Gao, F., & He, L. (2022). Driver-like decision-making method for vehicle longitudinal autonomous driving based on deep reinforcement learning. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 236(13), 3060–3070. https://doi.org/10.1177/09544070211063081

[7] Qian, Y., Feng, S., Hu, W., & Wang, W. (2022). Obstacle avoidance planning of autonomous vehicles using deep reinforcement learning. Advances in Mechanical Engineering, 14(12), 1–15. https://doi.org/10.1177/16878132221139661

[8] Wu, Z., Qu, F., Yang, L., & Gong, J. (2022). Human-like decision making for autonomous vehicles at the intersection using inverse reinforcement learning. Sensors, 22(12), 4500. https://doi.org/10.3390/s22124500

[9] Dinneweth, J., Boubezoul, A., Mandiau, R., & Espié, S. (2022). Multi-agent reinforcement learning for autonomous vehicles: A survey. Autonomous Intelligent Systems, 2, 27. https://doi.org/10.1007/s43684-022-00045-z

[10] Zhao, Q., Zhang, Y., & Li, X. (2022). Safe reinforcement learning for dynamical systems using barrier certificates. Connection Science, 34(1), 2822–2844. https://doi.org/10.1080/09540091.2022.2151567

[11] Ferreira, J. V., dos Santos, M. M., & de Oliveira, V. A. (2020). Vision-based robust control framework based on deep reinforcement learning applied to autonomous ground vehicles. Control Engineering Practice, 104, 104630. https://doi.org/10.1016/j.conengprac.2020.104630

[12] Wu, Z., Qiu, K., & Gao, H. (2020). Driving policies of V2X autonomous vehicles based on reinforcement learning methods. IET Intelligent Transport Systems, 14(5), 331–337. https://doi.org/10.1049/iet-its.2019.0457

[13] Ye, F., Zhang, S., Wang, P., & Chan, C.-Y. (2021). A survey of deep reinforcement learning algorithms for motion planning and control of autonomous vehicles. 2021 IEEE Intelligent Vehicles Symposium, 1073–1080. https://doi.org/10.1109/IV48863.2021.9575880

[14] Ma, X., Li, J., Kochenderfer, M. J., Isele, D., & Fujimura, K. (2021). Reinforcement learning for autonomous driving with latent state inference and spatial-temporal relationships. 2021 IEEE International Conference on Robotics and Automation, 6064–6071. https://doi.org/10.1109/ICRA48506.2021.9562006

[15] Chen, J., Yuan, B., & Tomizuka, M. (2021). Model-free deep reinforcement learning for urban autonomous driving. 2021 IEEE Intelligent Transportation Systems Conference, 2765–2771. https://doi.org/10.1109/ITSC48978.2021.9564770

[16] Liao, J., Liu, T., Tang, X., Mu, X., Huang, B., & Cao, D. (2020). Decision-making strategy on highway for autonomous vehicles using deep reinforcement learning. IEEE Access, 8, 177804–177814. https://doi.org/10.1109/ACCESS.2020.3022755

[17] Ye, Y., Zhang, X., & Sun, J. (2020). Automated vehicle's behavior decision making using deep reinforcement learning and high-fidelity simulation environment. Transportation Research Part C: Emerging Technologies, 107, 155–170. https://doi.org/10.1016/j.trc.2019.08.011

[18] Shi, T., Wang, P., Cheng, X., Chan, C.-Y., & Huang, D. (2020). Driving decision and control for automated lane change behavior based on deep reinforcement learning. 2020 IEEE Intelligent Vehicles Symposium, 1451–1456. https://doi.org/10.1109/IV47402.2020.9304578

[19] Makantasis, K., Kontorinaki, M., & Nikolos, I. K. (2020). Deep reinforcement-learning-based driving policy for autonomous road vehicles. IET Intelligent Transport Systems, 14(1), 13–24. https://doi.org/10.1049/iet-its.2019.0249

[20] Jin, M., & Lavaei, J. (2020). Stability-certified reinforcement learning: A control-theoretic perspective. IEEE Access, 8, 229086–229100. https://doi.org/10.1109/ACCESS.2020.3045114

[21] Chae, H., Jeong, Y., Lee, H., Park, J., & Yi, K. (2021). Design and implementation of human driving data-based active lane change control for autonomous vehicles. Proceedings of the Institution of Mechanical Engineers, Part D: Journal of Automobile Engineering, 235(1), 55–77. https://doi.org/10.1177/0954407020947678

[22] Chen, J., Sugumaran, V., & Qu, P. (2022). Connected and automated vehicle control at unsignalized intersection based on deep reinforcement learning in vehicle-to-infrastructure environment. International Journal of Distributed Sensor Networks, 18(8), 1–13. https://doi.org/10.1177/15501329221114060

[23] Zhang, X., Liu, X., Li, X., & Wu, G. (2022). Lane change decision algorithm based on deep Q network for autonomous vehicles. SAE Technical Paper, 2022-01-0084, 1–9. https://doi.org/10.4271/2022-01-0084

[24] Liu, T., Tang, L., Yang, S., & Sun, D. (2021). A deep reinforcement learning approach for autonomous highway driving. IFAC-PapersOnLine, 53(5), 542–546. https://doi.org/10.1016/j.ifacol.2021.04.142

[25] Zhao, J., Sun, J., Cai, Z., Wang, L., & Wang, Y. (2021). End-to-end deep reinforcement learning for image-based autonomous control. Applied Sciences, 11(18), 8419. https://doi.org/10.3390/app11188419

[26] Song, Q., Wang, W., Fu, W., Sun, Y., Wang, D., & Gao, Z. (2022). Research on quantum cognition in autonomous driving. Scientific Reports, 12, 300. https://doi.org/10.1038/s41598-021-04239-y

[27] Yang, L., Lei, W., Zhang, W., & Ye, T. (2022). Dual-flow network with attention for autonomous driving. Frontiers in Neurorobotics, 16, 978225. https://doi.org/10.3389/fnbot.2022.978225

[28] Lyu, Y., Luo, W., & Dolan, J. M. (2021). Safe reinforcement learning of autonomous vehicles with model-based control barrier functions. 2021 IEEE International Conference on Robotics and Automation, 8838–8844. https://doi.org/10.1109/ICRA48506.2021.9561149

[29] Zhao, W., Queralta, J. P., & Westerlund, T. (2020). Sim-to-real transfer in deep reinforcement learning for robotics: A survey. 2020 IEEE Symposium Series on Computational Intelligence, 737–744. https://doi.org/10.1109/SSCI47803.2020.9308468

[30] Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., & Schoellig, A. P. (2022). Safe learning in robotics: From learning-based control to safe reinforcement learning. Annual Review of Control, Robotics, and Autonomous Systems, 5, 411–444. https://doi.org/10.1146/annurev-control-042920-020211

[31] García, J., & Fernández, F. (2020). A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research, 21(42), 1–38. https://doi.org/10.5555/3455716.3455758

[32] Zhu, Z., & Zhao, H. (2022). A survey of deep reinforcement learning and imitation learning for autonomous driving policy learning. IEEE Transactions on Intelligent Transportation Systems, 23(9), 14043–14065. https://doi.org/10.1109/TITS.2021.3134702

[33] Wang, X., Lian, L., Miao, Y., Liu, Z., & Yu, S. X. (2021). Long-term visual localization with mobile sensors. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(11), 3769–3784. https://doi.org/10.1109/TPAMI.2020.2999102

[34] Feng, S., Yan, X., Sun, H., Feng, Y., & Liu, H. X. (2021). Intelligent driving intelligence test for autonomous vehicles with naturalistic and adversarial environment. Nature Communications, 12, 748. https://doi.org/10.1038/s41467-021-21007-8

[35] Aradi, S. (2020). Survey of deep reinforcement learning for motion planning of autonomous vehicles. IEEE Transactions on Intelligent Transportation Systems, 23(2), 740–759. https://doi.org/10.1109/TITS.2020.3024655