Proceedings of the International Conference on Industrial Engineering and Operations Management

Publisher: IEOM Society International, USA DOI: 10.46254/BA08.20250300
Published: December 20, 2025

Adaptive Reinforcement Learning with Control Barrier Functions for Safe State Avoidance in Discrete Event Systems

Md Nur-A-Adam Dony
Department of Electrical and Electronic Engineering
Rajshahi University of Engineering & Technology, Bangladesh
mdnuraadamdony@gmail.com

Md Ebrahim Khallil
Department of Electrical and Computer Engineering
Tennessee Technological University, Cookeville, TN 38505, USA
ebrahim140232@gmail.com

Md. Suruz Ali
Department of Electrical and Electronic Engineering
Pabna University of Science and Technology
Pabna, Dhaka, Bangladesh
suruz.eee.pust36@gmail.com

Abstract

Safety-critical autonomous systems require controllers that achieve performance objectives while guaranteeing constraint satisfaction. This paper presents a framework integrating Discrete Event System (DES) supervisory control with reinforcement learning (RL) through Control Barrier Function (CBF) principles. We establish a formal connection between DES state avoidance and CBF-based safety by defining discrete states as approximations of CBF level sets. The proposed DES-RL framework employs barrier-inspired reward shaping that encodes safety requirements into the learning process, enabling agents to discover optimal policies while maintaining forward invariance of safe sets. We develop a Q-learning algorithm with CBF-shaped rewards that learns safe policies without requiring explicit system dynamics models. The framework is validated through an autonomous vehicle lane-keeping application, demonstrating zero lane departure violations across 100 evaluation episodes. Comparative analysis with traditional CBF-QP controllers shows that DES-RL achieves equivalent safety under nominal to strong disturbance conditions.

Keywords

Reinforcement Learning, Control Barrier Functions, Discrete Event Systems, Autonomous Vehicles, Safe Control.

1. Introduction

The deployment of autonomous systems in safety-critical applications has created an urgent need for control strategies that guarantee both high performance and strict safety constraints (Ames et al., 2019). Reinforcement learning (RL) has emerged as a powerful paradigm for learning optimal control policies (Sutton and Barto, 2018), achieving remarkable success in complex tasks (Mnih et al., 2015). However, standard RL algorithms optimize cumulative rewards without explicit consideration of safety constraints. Control Barrier Functions (CBFs) provide a mathematically rigorous framework for ensuring safety in dynamical systems (Ames et al., 2017). Discrete Event

Systems (DES) theory (Ramadge and Wonham, 1987) offers supervisory control where systems are modeled as finite automata. The primary objectives are: (1) establish formal connection between DES state avoidance and CBF-based safety; (2) develop a model-free RL algorithm with CBF-inspired reward shaping; (3) validate through autonomous vehicle lane-keeping with zero safety violations; (4) compare with traditional CBF-QP controllers. The remainder of this paper is organized as follows: Section 2 reviews related work in RL, CBFs, and DES. Section 3 presents the methodology including system model, theoretical framework, and proposed algorithm. Section 4 describes the lane-keeping application. Section 5 presents results and discussion. Section 6 concludes the paper.

2. Literature Review

The foundations of value-based reinforcement learning were established with Q-learning (Watkins and Dayan, 1992), which learns optimal action-value functions through temporal difference up-dates without requiring environment models. This approach was later extended to high-dimensional state spaces through Deep Q-Networks (Mnih et al., 2015), enabling direct learning from raw sensory inputs. For continuous control tasks, policy gradient methods such as Proximal Policy Optimization (Schulman et al., 2017) have demonstrated remarkable success by directly optimizing parameterized policies.

The theoretical foundations for safety in control systems are rooted in set invariance theory (Blanchini, 1999). Control Barrier Functions formalize these concepts for control-affine systems, where a barrier function defines the safe set and the CBF-QP formulation provides a mechanism to minimally modify nominal controllers while ensuring safety (Ames et al., 2017). The robust properties of CBFs under model uncertainty were analyzed by Xu et al. (2015), establishing input-to-state safety conditions.

In the discrete domain, DES theory (Ramadge and Wonham, 1987) models systems as finite automata with supervisory control preventing transitions to forbidden states. The integration of safety into RL has been approached through Constrained MDPs, where CPO (Achiam et al., 2017) uses trust region methods with cost constraints. In parallel, a large body of work has studied planning and acting in partially observable stochastic domains using POMDP formulations (Kaelbling et al., 1998; Cassandra et al., 1994; Shani et al., 2013), which provide a principled framework for decision-making under uncertainty but remain computationally demanding for high-dimensional control problems. Cheng et al. (2019) combined model-free RL with CBF-based safety filtering using Gaussian processes for dynamics learning, while Taylor et al. (2020) proposed episodic updates of CBF parameters. However, existing approaches face fundamental limitations: CBF methods require accurate dynamics knowledge, constrained optimization provides only soft guarantees, and safety shielding can interfere with the learning process. Our framework addresses these challenges by embedding CBF-inspired constraints directly into the DES state structure and reward function, enabling model-free learning with interpretable safety guarantees.

3. Methodology

This section presents the theoretical framework for integrating Discrete Event Systems with Con-trol Barrier Functions through reinforcement learning. We begin with the system model, establish the formal connection between DES and CBF, and develop the proposed learning algorithm with theoretical guarantees.

3.1 System Model

Consider a continuous-time control-affine dynamical system:

$$\dot{x} = f(x) + g(x)u + w \tag{1}$$

where $x \in \mathcal{X} \subseteq \mathbb{R}^n$ is the state vector, $u \in \mathcal{U} \subseteq \mathbb{R}^m$ is the control input, $w \in \mathcal{W} \subseteq \mathbb{R}^n$ represents bounded disturbances with $\|w\| \leq w_{\max}$, and $f: \mathbb{R}^n \to \mathbb{R}^n, g: \mathbb{R}^n \to \mathbb{R}^{n\times m}$ are locally Lipschitz continuous functions.
We assume the existence of a continuously differentiable barrier function $h: \mathbb{R}^n \to \mathbb{R}$ that defines the safe set:

$$\mathcal{C} = \{x \in \mathbb{R}^n : h(x) \geq 0\} \tag{2}$$

The boundary of the safe set is $\partial\mathcal{C} = \{x: h(x) = 0\}$, and we require $\nabla h(x) \neq 0$ for all $x \in \partial\mathcal{C}$.

3.2 Discrete Event System Abstraction

We construct a discrete abstraction of the continuous state space by partitioning it according to the level sets of the barrier function. Define threshold values $0 < \epsilon_2 < \epsilon_1$ and construct the following discrete state space:

$$S_{\text{CENTERED}} = \{x \in \mathbb{R}^n : h(x) \geq \epsilon_1\} \quad (3)$$
$$S_{\text{WARNING}} = \{x \in \mathbb{R}^n : \epsilon_2 \leq h(x) < \epsilon_1\} \quad (4)$$
$$S_{\text{RECOVERY}} = \{x \in \mathbb{R}^n : 0 < h(x) < \epsilon_2\} \quad (5)$$
$$S_{\text{UNSAFE}} = \{x \in \mathbb{R}^n : h(x) \leq 0\} \quad (6)$$

The discrete state space is $\mathcal{S} = \{S_{\text{Centered}}, S_{\text{WARNING}}, S_{\text{RECOVERY}}, S_{\text{UNSAFE}}\}$. The state mapping function $\sigma: \mathcal{X} \to \mathcal{S}$ determines the discrete state from the continuous state:

$$\sigma(x) = \begin{cases} S_{\text{CENTERED}} & \text{if } h(x) \geq \epsilon_1 \\ S_{\text{WARNING}} & \text{if } \epsilon_2 \leq h(x) < \epsilon_1 \\ S_{\text{RECOVERY}} & \text{if } 0 < h(x) < \epsilon_2 \\ S_{\text{UNSAFE}} & \text{if } h(x) \leq 0 \end{cases} \quad (7)$$

The DES abstraction $G = (\mathcal{S}, \mathcal{A}, \delta, s_0)$ is defined where $\mathcal{A}$ is the finite action set, $\delta: \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ is the probabilistic transition function induced by the continuous dynamics, and $s_0 = \sigma(x_0)$ is the initial discrete state.

3.3 Problem Formulation

We formulate the safe control problem as a Constrained Markov Decision Process (CMDP). The objective is to find a policy $\pi: \mathcal{S} \to \mathcal{A}$ that maximizes expected cumulative reward while ensuring the system never enters the unsafe state.

Definition 1 (Safe Policy). A policy π is said to be safe if, for all initial states $s_0 \in \mathcal{S} \setminus \{S_{\text{UNSAFE}}\}$ and for all time steps $t \geq 0$:

$$\mathbb{P}[s_t = S_{UNSAFE} \mid s_0, \pi] = 0 \quad (8)$$

Definition 2 (Optimal Safe Policy). The optimal safe policy π^* is defined as:

$$\pi^* = \text{argmax}_{\pi \in \Pi_{\text{safe}}} \mathbb{E}_\pi \left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t) \right] \quad (9)$$

subject to the constraint in (8), where Π_{safe} denotes the set of all safe policies, $\gamma \in (0,1)$ is the discount factor, and $R: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function.

3.4 Connection Between DES State Avoidance and CBF Safety

We now establish the formal connection between DES state avoidance in the discrete abstraction and CBF-based safety in the continuous domain.

Lemma 1 (DES-CBF Correspondence). Let $h: \mathbb{R}^n \to \mathbb{R}$ be a Control Barrier Function for the system (1) with safe set $\mathcal{C}$ defined by (2). If a policy π ensures that the discrete state $s_t \neq S_{\text{UNSAFE}}$ for all $t \geq 0$, then the continuous state satisfies $h(x(t)) > 0$ for all $t \geq 0$, i.e., the system remains in the safe set $\mathcal{C}$.

Proof. By construction of the state mapping (7), we have:

$$s_t \neq S_{\text{UNSAFE}} \Leftrightarrow \sigma(x(t)) \neq S_{\text{UNSAFE}} \Leftrightarrow h(x(t)) > 0 \quad (10)$$

Since $\mathcal{C} = \{x: h(x) \geq 0\}$ and $h(x(t)) > 0$ implies $x(t) \in \text{Int}(\mathcal{C}) \subset \mathcal{C}$, the system remains strictly within the safe set.

3.5 CBF-Shaped Reward Structure

To encode CBF safety requirements into the reinforcement learning framework, we designed a reward function that incorporates the barrier function value. The reward structure consists of three components:

$$R(s, a, s') = R_{\text{task}}(s, a) + R_{\text{barrier}}(s') + R_{\text{violation}}(s') \quad (11)$$

Task Reward: Encourages progress toward the control objective:

$$R_{\text{task}}(s, a) = \begin{cases} r_{\text{nominal}} & \text{if } s = S_{\text{CENTERED}} \\ r_{\text{warning}} & \text{if } s = S_{\text{WARNING}} \\ r_{\text{recovery}} & \text{if } s = S_{\text{RECOVERY}} \end{cases} \quad (12)$$

with $r_{\text{nominal}} > r_{\text{warning}} > r_{\text{recovery}} > 0$.

Barrier Reward: Proportional to the barrier function value, encouraging states with larger safety margins:

$$R_{\text{barrier}}(s') = \beta \cdot \bar{h}(s') \quad (13)$$

where $\beta > 0$ is a weighting coefficient and $\bar{h}(s') = \mathbb{E}_{x \in s'}[h(x)]$ is the expected barrier value in state s'.

Violation Penalty: A large negative reward for entering the unsafe state:

$$R_{\text{violation}}(s') = \begin{cases} -\lambda & \text{if } s' = S_{\text{UNSAFE}} \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

where $\lambda \gg 0$ is a large penalty coefficient.

Proposition 2 (Reward-CBF Alignment). As $\lambda \to \infty$, the optimal policy under the reward structure (11) converges to a safe policy that avoids S_{UNSAFE}, provided such a policy exists.

Proof. Let π_λ^* denote the optimal policy for penalty coefficient λ. For any policy π that enters S_{UNSAFE} with positive probability $p > 0$, the expected return satisfies:

$$V^\pi(s_0) \leq V_{\max} - p \cdot \gamma^{t_{\text{unsafe}}} \cdot \lambda \tag{15}$$

where $V_{\max}$ is the maximum possible return without violations and t_{unsafe} is the expected time to violation.

For any safe policy π_{safe} :

$$V^{\pi_{\text{safe}}}(s_0) \geq V_{\min} > -\infty \tag{16}$$

As $\lambda \to \infty, V^\pi(s_0) \to -\infty$ for any unsafe policy, while $V^{\pi_{\text{safe}}}(s_0)$ remains bounded. Therefore, π_λ^* must be safe for sufficiently large λ.

3.6 Proposed Algorithm: DES-RL with CBF-Shaped Rewards

The action-value function $Q: \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ represents the expected cumulative reward for taking action a in state s and following the optimal policy thereafter:

$$Q^*(s,a) = \mathbb{E}\left[R(s,a,s') + \gamma \max_{a'} Q^*(s',a') \mid s,a\right] \tag{17}$$

The Q-learning update rule with CBF-shaped rewards is:

$$Q(s_t,a_t) \leftarrow Q(s_t,a_t) + \alpha\left[R_t + \gamma \max_{a'} Q(s_{t+1},a') - Q(s_t,a_t)\right] \tag{18}$$

where $\alpha \in (0,1)$ is the learning rate.

We employ an ε-greedy exploration strategy with decay:

$$a_t = \begin{cases} \text{argmax}_{a \in \mathcal{A}} Q(s_t,a) & \text{with probability } 1-\varepsilon_t \\ \text{uniform}(\mathcal{A}) & \text{with probability } \varepsilon_t \end{cases} \tag{19}$$

where $\varepsilon_t = \max(\varepsilon_{\min}, \varepsilon_0 \cdot \rho^t)$ with decay rate $\rho \in (0,1)$.

3.7 Theoretical Analysis

Theorem 3 (Safety and Convergence of DES-RL).

Consider the DES abstraction $G = (\mathcal{S}, \mathcal{A}, \delta, s_0)$ constructed from the continuous system (1) with barrier function h and state mapping (7). Let the reward function be defined by (11) & violation penalty λ. Under the following conditions:

(I) The learning rate satisfies $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$

(II) Every state-action pair is visited infinitely often

(III) There exists at least one safe policy $\pi_{\text{safe}} \in \Pi_{\text{safe}}$

(IV) The penalty coefficient satisfies $\lambda > \frac{V_{\max}}{(1-\gamma)\cdot p_{\min}}$

Then the following hold:

Part I (Necessity): If the learned policy π^* is optimal under the reward structure (11), then π^* is safe, i.e., $\pi^* \in \Pi_{\text{safe}}$.

Part II (Sufficiency): The Q-learning algorithm with CBF-shaped rewards (Algorithm 1) converges with the optimal safe policy π^* that maximizes expected return among all safe policies.

Proof.

Proof of Part I (Necessity): We prove by contradiction. Assume π^* is optimal but not safe. Then there exists a time $T > 0$ such that:

$$\mathbb{P}[s_T = S_{\text{UNSAFE}} \mid s_0, \pi^*] = p > 0 \tag{20}$$

Step 1: Bound the expected return of π^*. The value function decomposes as:

$$V^{\pi^*}(s_0) = \mathbb{E}_{\pi^*}\left[\sum_{t=0}^{T-1} \gamma^t R_t\right] + \mathbb{E}_{\pi^*}[\gamma^T R_T] + \mathbb{E}_{\pi^*}\left[\sum_{t=T+1}^{\infty} \gamma^t R_t\right] \tag{21}$$

Since $R_t \leq R_{\max}$ for non-violation rewards and $R_T = -\lambda$ upon violation with probability p :

$$V^{\pi^*}(s_0) \leq \frac{R_{\max}}{1-\gamma} - \gamma^T \cdot p \cdot \lambda \tag{22}$$

Step 2: Bound the expected return of any safe policy. For $\pi_{\text{safe}} \in \Pi_{\text{safe}}$, no violations occur, so $R_{\text{violation}} = 0$ always. Thus:

$$V^{\pi_{\text{safe}}}(s_0) \geq \frac{R_{\min}}{1-\gamma} \tag{23}$$

Step 3: Derive contradiction. For π^* to be optimal, we require $V^{\pi^*}(s_0) \geq V^{\pi_{\text{safe}}}(s_0)$. Substituting (22) and (23):

$$\frac{R_{\max}}{1-\gamma} - \gamma^T \cdot p \cdot \lambda \geq \frac{R_{\min}}{1-\gamma} \tag{24}$$

Rearranging for λ :

$$\lambda \leq \frac{R_{\max} - R_{\min}}{(1-\gamma)\cdot\gamma^T \cdot p} \leq \frac{V_{\max}}{(1-\gamma)\cdot p_{\min}} \tag{25}$$

This contradicts condition (iv). Therefore, π^* must be safe.

Proof of Part II (Sufficiency): The proof proceeds in two steps.

Step 1: Convergence of Q-values. Under conditions (i) and (ii), the standard Q-learning convergence theorem (Watkins and Dayan, 1992) guarantees:

$$Q(s,a) \to Q^*(s,a) \text{ w.p. } 1, \forall (s,a) \in \mathcal{S} \times \mathcal{A} \tag{26}$$

The convergence holds because the CBF-shaped reward is bounded:

$$|R(s,a,s')| \leq |R_{\text{task}}| + \beta|h_{\max}| + \lambda < \infty \tag{27}$$

Step 2: Optimality among safe policies. From Part I, the optimal policy π^* under (11) is safe. For any safe policy, $R_{\text{violation}} = 0$ since S_{UNSAFE} is never visited. The greedy policy:

$$\pi^*(s) = \text{argmax}_{a\in\mathcal{A}} Q^*(s,a) \tag{28}$$

achieves maximum expected return by the Bellman optimality principle. Since π^* is both safe and return-maximizing, it is optimal among all safe policies.

Corollary 4 (Forward Invariance). Under the conditions of Theorem 3, the policy learned by Algorithm 1 ensures forward invariance of the safe set $\mathcal{C}$: if $x(0) \in \mathcal{C}$, then $x(t) \in \mathcal{C}$ for all $t \geq 0$.

Proof. The proof follows by chain of implications:

1. From Theorem 3, the learned policy π^* is safe
2. By Definition 1, π^* safe $\Rightarrow s_t \neq S_{\text{UNSAFE}}$ for all $t \geq 0$
3. By Lemma 1, $s_t \neq S_{\text{UNSAFE}} \Rightarrow h(x(t)) > 0$ for all $t \geq 0$
4. Since $h(x(t)) > 0 \Rightarrow x(t) \in \text{Int}(\mathcal{C}) \subset \mathcal{C}$

Therefore, forward invariance of $\mathcal{C}$ is established.

Algorithm 1 DES-RL with CBF-Shaped Rewards

Require: Barrier function h, thresholds ϵ_1, ϵ_2, learning rate α, discount factor γ, exploration parameters $\varepsilon_0, \varepsilon_{\min}, \rho$, penalty λ, barrier weight β, episodes N

Ensure: Learned Q-table Q

```
Initialize Q(s, a) ← 0 for all s ∈ S, a ∈ A
Set ε ← ε0
for episode = 1, 2, . . . , N do
    Initialize continuous state x0, compute s0 ← σ(x0)
    while episode not terminated do
        Select action at using ε-greedy policy (19)
```

Execute a_t, observe x_{t+1}, compute $s_{t+1} \leftarrow \sigma(x_{t+1})$
Compute $R_t \leftarrow R_{\text{task}}(s_t, a_t) + \beta \cdot h(x_{t+1}) - \lambda \cdot \mathbf{1}[s_{t+1} = S_{\text{UNSAFE}}]$9: Update $Q(s_t, a_t)$ using (18)
if $s_{t+1} = S_{\text{UNSAFE}}$ **or** $t > T_{\max}$ **then**
Terminate episode
end if
end while
Update $\varepsilon \leftarrow \max(\varepsilon_{\min}, \varepsilon \cdot \rho)$
end for
return Q

4. Application: Autonomous Vehicle Lane-Keeping

4.1 Vehicle Model and Safety Constraints

Consider a vehicle with lateral dynamics $\dot{y} = u + w$, where y is lateral position, u is control input, and w is disturbance. The state vector is $x = [y, \dot{y}]^{\top}$. Discretized dynamics ($\Delta t = 0.1$ s):

$$y_{k+1} = y_k + \dot{y}_k \cdot \Delta t + \frac{1}{2} u_k \cdot \Delta t^2 \tag{29}$$

$$\dot{y}_{k+1} = \dot{y}_k + u_k \cdot \Delta t + w_k \cdot \Delta t \tag{30}$$

Lane parameters: width $W_{\text{lane}} = 3.7$ m, vehicle width $W_{\text{vehicle}} = 1.8$ m, safe margin $d_{\text{safe}} = 0.95$ m.
The barrier function is:

$$h(y) = d_{\text{safe}}^2 - y^2 = 0.9025 - y^2 \tag{31}$$

with safe set $\mathcal{C} = \{x: |y| \leq 0.95 \text{ m}\}$

4.2 DES Abstraction and Actions

Thresholds: $d_{\text{warning}} = 0.57$ m, $d_{\text{recovery}} = 0.81$ m, giving $\epsilon_1 = 0.578$ and $\epsilon_2 = 0.247$. The discrete states and actions are defined in Table 1and Table 2

Table 1. DES States

State	Condition
S_{CENTERED}	$\lvert y \rvert \leq 0.57$ m
S_{WARNING}	$0.57 < \lvert y \rvert \leq 0.81$ m
S_{RECOVERY}	$0.81 < \lvert y \rvert < 0.95$ m
S_{DEPARTED}	$\lvert y \rvert \geq 0.95$ m

Table 2. DES Action Space

Action	Control Input
MAINTAIN	$u = 0$
GENTLE_CORRECT	$u = -0.5\text{sign}(y)$
AGGRESSIVE_CORRECT	$u = -1.5\text{sign}(y)$
RISKY_ACCELERATE	$u = 0.3\text{sign}(y)$

4.3 Reward Structure and Training

The reward structure follows (11) and training hyperparameters are shown in Table 3 and Table 4.

Table 3. Reward Structure

Reward Component	Value
R_{task} (CENTERED + MAINTAIN)	5.0
R_{task} (CENTERED + other)	4.0
R_{task} (WARNING)	2.0
R_{task} (RECOVERY)	0.5
$R_{\text{barrier}} = \beta(d_{\text{safe}}^2 - y^2)$	$\beta = 1.0$
$R_{\text{violation}}$ (DEPARTED)	-1000

Table 4. Reward Structure and Training Hyperparameters

Parameter	Value
Episodes N	5000
Max steps $T_{\max}$	200
Learning rate α	0.1
Discount factor γ	0.95
Initial exploration ε_0	0.3
Min exploration $\varepsilon_{\min}$	0.01
Decay rate ρ	0.999

5. Results and Discussion

5.1 Training and Convergence

Figure **1** shows training performance. The learning curve stabilizes around $850 - 880$ reward after 1000 episodes with zero safety violations throughout. The average barrier value remains above 0.8, well above the safety boundary. Q-values reveal learned preferences: AGGRESSIVE_CORRECT achieves highest Q-value (≈ 80) in WARNING state, while MAINTAIN has highest Q-value in CENTERED state.

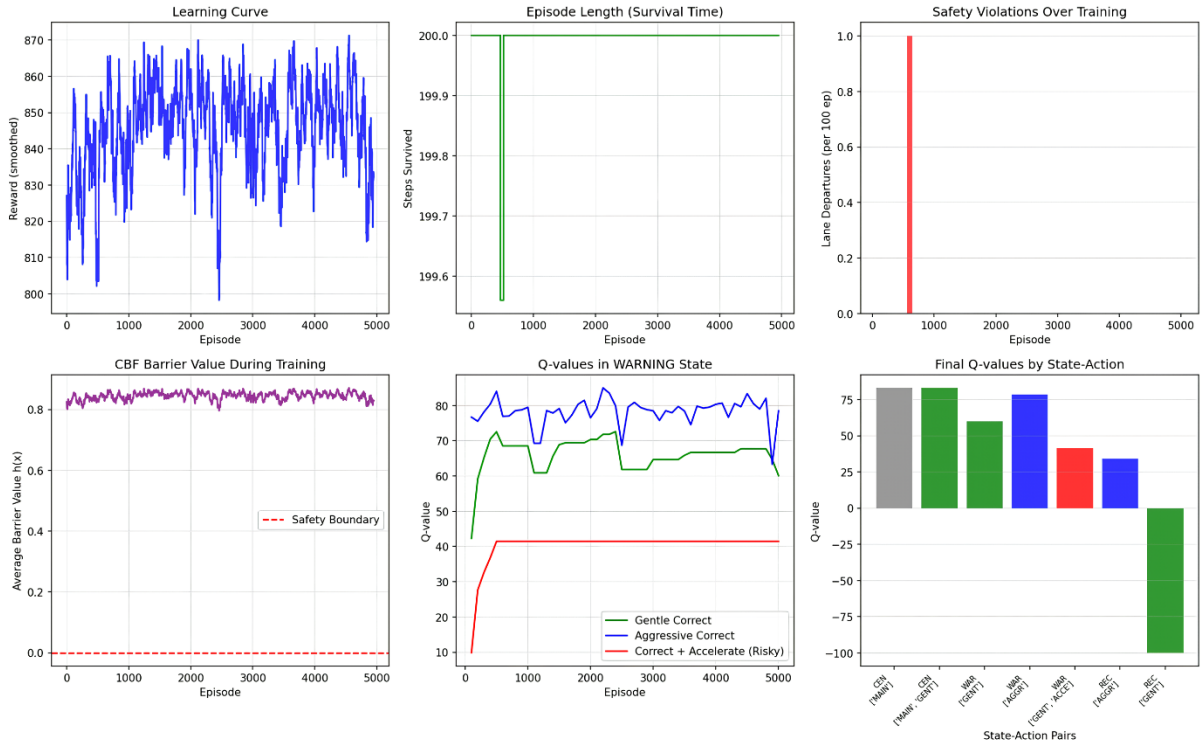

Figure 1. Training performance: (a) Learning curve; (b) Episode length; (c) Safety violations; (d) Barrier value; (e) Q-value evolution in WARNING state; (f) Final Q-values.

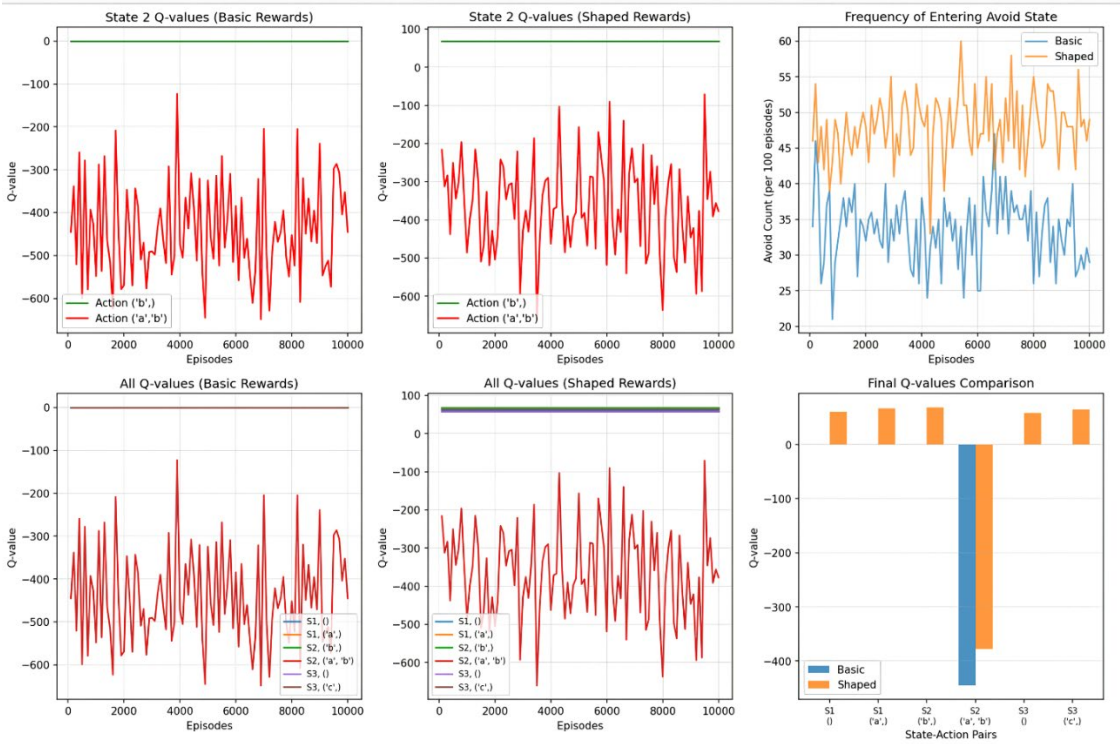

Figure 2. Convergence analysis: (a-b) State 2 Q-values under basic and shaped rewards; (c) Avoid state frequency; (d-e) All Q-values over training; (f) Final Q-value comparison.

Figure **2** compares basic vs. CBF-shaped rewards. The dangerous action develops strongly negative Q -values (-400 to -500). Shaped rewards produce positive Q -values for safe actions, encouraging active task completion while maintaining safety.

5.2 Evaluation and Comparison

Over 100 evaluation episodes:

- 0% lane departure rate,
- average reward 857.92 ± 52.32,
- average barrier value 0.847 ± 0.031.
- Time distribution: 89.3% CENTERED,
- 9.8% WARNING, 0.9% RECOVERY.
- These results validate Theorem 3.

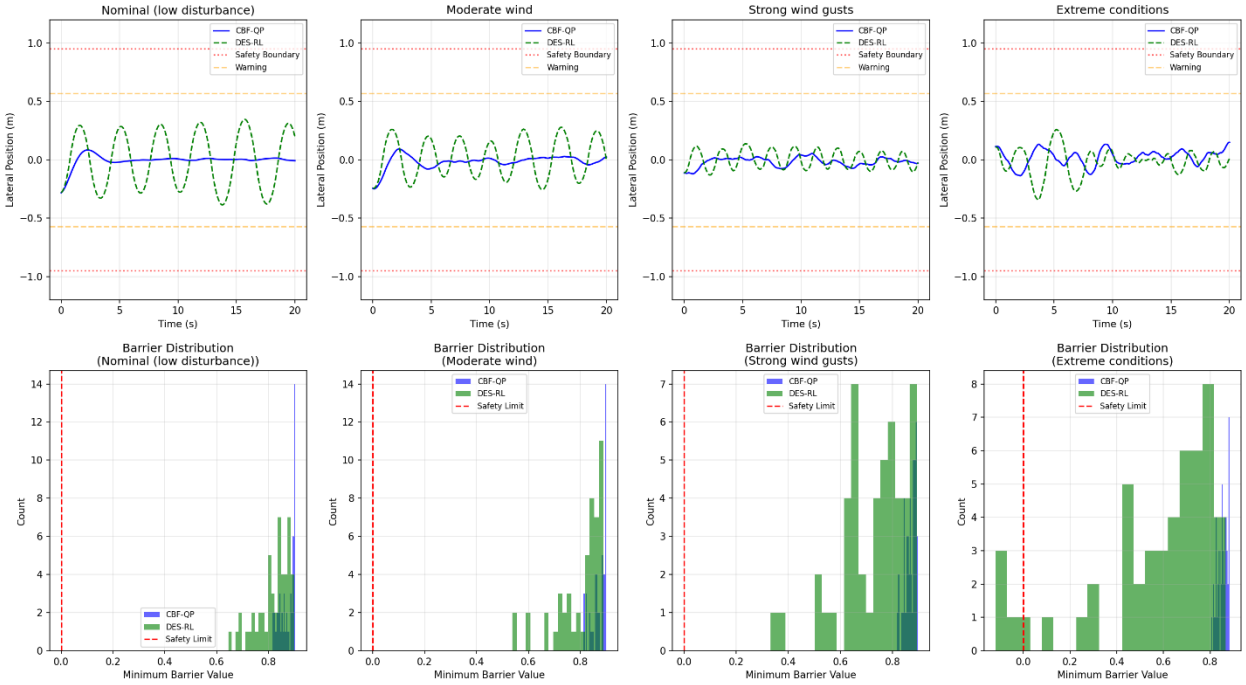

Figure 3. CBF-QP vs. DES-RL comparison: (Top) Trajectories under varying disturbances; (Bottom) Minimum barrier value distributions.

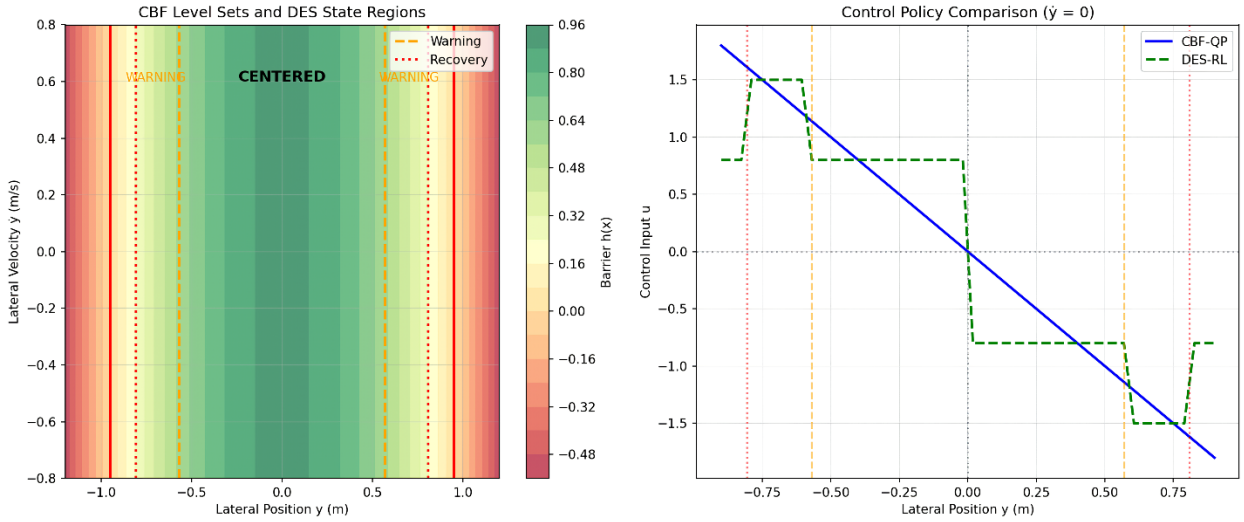

Figure 4. State space analysis: (Left) CBF level sets with DES state boundaries; (Right) Control policy comparison at $\dot{y}=0$.

Figure **3** compares controllers under varying disturbances. Both achieve 0% violations under nominal to strong conditions. Under extreme conditions ($\sigma_w = 0.50$), CBF-QP maintains 0% while DES-RL shows 8% violations. Figure **4** shows CBF level sets and DES regions. The discrete states effectively partition continuous safety regions.

Trade-offs: CBF-QP provides formal guarantees but requires dynamics knowledge and real-time QP solving; DES-RL is model-free with only table lookups after training but shows degradation under extreme disturbances (Table 5).

Table 5. Safety Performance Comparison

Controller	**Nominal**	**Moderate**	**Strong**	**Extreme**
CBF-QP Violation Rate	0%	0%	0%	0%
DES-RL Violation Rate	0%	0%	0%	8%

6. Conclusion

This paper presented a novel framework integrating Discrete Event System supervisory control with Control Barrier Function principles through reinforcement learning. The key insight is that DES states defined as discretization's of CBF level sets enable translation of continuous safety constrains into discrete state avoidance specifications, allowing model-free learning of safe policies without explicit dynamics knowledge.

The main contributions are: (1) a formal DES-CBF correspondence establishing that discrete state avoidance ensures continuous safe set invariance (Lemma 1); (2) a CBF-shaped reward structure proven to induce safe optimal policies (Proposition 2); (3) theoretical guarantees on safety and convergence (Theorem 3); and (4) a practical Q-learning algorithm requiring only table lookups after training. Experimental validation on autonomous vehicle lane-keeping demonstrated zero safety violations across 100 evaluation episodes, with performance equivalent to model-based CBF-QP controllers under nominal to strong disturbances.

Future work will extend the framework to continuous action spaces using actor-critic methods, develop adaptive state abstraction techniques, and investigate multi-agent coordination with collective safety constraints.

References

Achiam, J., Held, D., Tamar, A. and Abbeel, P., Constrained policy optimization, Proceedings of the 34th International Conference on Machine Learning, pp. 22–31, 2017.

Ames, A.D., Coogan, S., Egerstedt, M., Notomista, G., Sreenath, K. and Tabuada, P., Control barrier functions: Theory and applications, Proceedings of the 18th European Control Conference, pp. 3420–3431, 2019.

Ames, A.D., Xu, X., Grizzle, J.W. and Tabuada, P., Control barrier function based quadratic programs for safety critical systems, IEEE Transactions on Automatic Control, vol. 62, no. 8, pp. 3861–3876, 2017.

Blanchini, F., Set invariance in control, Automatica, vol. 35, no. 11, pp. 1747–1767, 1999.

Cassandra, A.R., Kaelbling, L.P. and Littman, M.L., Acting optimally in partially observable stochastic domains, Proceedings of the 12th National Conference on Artificial Intelligence (AAAI-94), pp. 1023–1028, 1994.

Cheng, R., Orosz, G., Murray, R.M. and Burdick, J.W., End-to-end safe reinforcement learning through barrier functions for safety-critical continuous control tasks, Proceedings of the AAAI Conference on Artificial Intelligence, vol. 33, no. 1, pp. 3387–3395, 2019.

Garcia, J. and Fernandez, F., A comprehensive survey on safe reinforcement learning, Journal of Machine Learning Research, vol. 16, no. 1, pp. 1437–1480, 2015.

Kaelbling, L.P., Littman, M.L. and Cassandra, A.R., Planning and acting in partially observable stochastic domains, Artificial Intelligence, vol. 101, no. 1–2, pp. 99–134, 1998.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A.A., Veness, J., Bellemare, M.G., Graves, A., Riedmiller, M., Fidjeland, A.K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S. and Hassabis, D., Human-level control through deep reinforcement learning, Nature, vol. 518, no. 7540, pp. 529–533, 2015.

Ramadge, P.J. and Wonham, W.M., Supervisory control of a class of discrete event processes, SIAM Journal on Control and Optimization, vol. 25, no. 1, pp. 206–230, 1987.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A. and Klimov, O., Proximal policy optimization algorithms, arXiv preprint arXiv:1707.06347, 2017.

Shani, G., Pineau, J. and Kaplow, R., A survey of point-based POMDP solvers, Autonomous Agents and Multi-Agent Systems, vol. 27, no. 1, pp. 1–51, 2013.

Sutton, R.S. and Barto, A.G., Reinforcement Learning: An Introduction, 2nd ed., MIT Press, Cambridge, MA, 2018.

Taylor, A.J., Singletary, A., Yue, Y. and Ames, A.D., Learning for safety-critical control with control barrier functions, Proceedings of the 2nd Conference on Learning for Dynamics and Control, pp. 708–717, 2020.

Watkins, C.J.C.H. and Dayan, P., Q-learning, Machine Learning, vol. 8, no. 3–4, pp. 279–292, 1992.

Wonham, W.M. and Ramadge, P.J., On the supremal controllable sublanguage of a given language, SIAM Journal on Control and Optimization, vol. 25, no. 3, pp. 637–659, 1987.

Xu, X., Tabuada, P., Grizzle, J.W. and Ames, A.D., Robustness of control barrier functions for safety critical control, IFAC-PapersOnLine, vol. 48, no. 27, pp. 54–61, 2015.

Biographies

Md Nur-A-Adam Dony is a Ph.D. candidate in Electrical and Computer Engineering at Tennessee Technological University, Cookeville, TN, USA. He received M.S. degrees in Electrical Engineering from The Pennsylvania State University (2025) and The University of Texas Rio Grande Valley, and a B.Sc. in Electrical and Electronic Engineering from Rajshahi University of Engineering and Technology, Bangladesh. His research interests include safe reinforcement learning, control barrier functions, and model predictive control for autonomous systems.

Md Ebrahim Khallil is an M.S. student in Electrical and Computer Engineering at Tennessee Technological University, Cookeville, TN, USA, with expected graduation in May 2026. He received his B.Sc. in Electrical and Electronics Engineering from Pabna University of Science and Technology, Bangladesh, and previously worked as an Assistant Engineer at Reverie Power and Automation Engineering Ltd. His research interests include optimal power flow, renewable energy integration, reinforcement Learning, smart grids, and machine learning for power systems