The Evolution of Criticality in Deep Reinforcement Learning

Chidvilas Karpenahalli Ramakrishna[a], Adithya Mohan[b], Zahra Zeinaly[c] and Lenz Belzner[d]
AImotion Bavaria, Technische Hochschule Ingolstadt, Esplanade 10, 85049 Ingolstadt, Germany
{Chidvilas.Karpenahalli, Adithya.Mohan, Zahra.Zeinaly, Lenz.Belzner}@thi.de

Keywords: Criticality, Deep Reinforcement Learning, Agents, Autonomous Driving, Deep Q-Learning (DQN), Trust.

Abstract: In Reinforcement Learning (RL), certain states demand special attention due to their significant influence on outcomes; these are identified as critical states. The concept of criticality is essential for the development of effective and robust policies and to improve overall trust in RL agents in real-world applications like autonomous driving. The current paper takes a deep dive into criticality and studies the evolution of criticality throughout training. The experiments are conducted on a new, simple yet intuitive continuous cliff maze environment and the Highway-env autonomous driving environment. Here, a novel finding is reported that criticality is not only learnt by the agent but can also be unlearned. We hypothesize that diversity in experiences is necessary for effective criticality quantification which is majorly driven by the chosen exploration strategy. This close relationship between exploration and criticality is studied utilizing two different strategies namely the exponential ε-decay and the adaptive ε-decay. The study supports the idea that effective exploration plays a crucial role in accurately identifying and understanding critical states.

1 INTRODUCTION

Reinforcement Learning (RL) derives its name from the process of optimizing policy through a reward mechanism, which utilizes both positive and negative reinforcements to guide decision-making. Deep reinforcement learning (DRL) combines the approximation and generalization capabilities of neural networks with RL to allow agents to operate in complex, high-dimensional state and action spaces. Apart from enjoying incredible success in complex games (Mnih, 2013; Silver et al., 2016; Silver et al., 2017), DRL has also demonstrated remarkable success in addressing challenges related to autonomous driving (Ravi Kiran et al., 2022; Li et al., 2020), recommendation systems (Afsar et al., 2022; Chen et al., 2021), robotics (Gu et al., 2016), supply chain management and production (Panzer and Bender, 2022; Hubbs et al., 2020; Boute et al., 2022), energy management (Santorsola et al., 2023) and other real-world applications. Although significant advancements have been made in the field of DRL, some challenges exist and one such key concept in DRL that requires attention is that of critical states (Spielberg and Azaria, 2019). Critical states in the context of a Markov Decision Process (MDP) and RL are states in which the choice of action significantly influences the outcome. In other words, these are the states where the agent strongly prefers certain actions over others. The ability to detect and handle critical states is essential for building trust in RL systems, especially in real-world applications like Autonomous Driving (AD) (Huang et al., 2018). Monitoring the performance alone is insufficient as a trustworthy agent would also retain awareness of the consequences of incorrect actions. Hence, trust in the system may diminish if the agent's understanding of criticality degrades during learning. Studying the evolution of criticality ensures safe decision-making, a topic that, to our knowledge, has not been explored in prior work. Our contribution in the current research is threefold,

- First, we study the evolution of criticality during the learning process.
- We report a novel finding of unlearning criticality, which compromises safety and trust in RL systems, as it leads to policies that perform well but ignore criticality in decision-making.
- We hypothesize that effective criticality quantification requires sufficient visits and diverse experiences in critical states. This is validated through a study of two exploration strategies, showing that enhanced exploration can help retain criticality.

[a] https://orcid.org/0009-0001-3091-9523
[b] https://orcid.org/0009-0004-3572-9982
[c] https://orcid.org/0009-0006-8575-9033
[d] https://orcid.org/0009-0002-4683-5460

Karpenahalli Ramakrishna, C., Mohan, A., Zeinaly, Z. and Belzner, L.
The Evolution of Criticality in Deep Reinforcement Learning.
DOI: 10.5220/0013114200003890

In *Proceedings of the 17th International Conference on Agents and Artificial Intelligence (ICAART 2025) - Volume 3*, pages 217-224
ISBN: 978-989-758-737-5; ISSN: 2184-433X

2 BACKGROUND

2.1 Markov Decision Process (MDP)

An MDP models sequential decision-making as a tuple (S, A, P, R, γ). Here, S is the state space, A is the action space, $P(s'|s,a)$ is the *transition probability*, $R(s,a,s')$ is the *reward function* and γ is the discount factor which controls future rewards. MDPs satisfy the Markov property, where the next state s' depends only on the current state s and action a. When P and R are unknown, RL methods are used to learn optimal policies through environmental interactions.

2.2 *Q-value*

The action-value function or the *Q-value function* $Q^{\pi}(s,a)$ represents the expected cumulative reward an agent receives by starting from a given state s and taking an action a and following a policy $\pi(a|s)$. $Q^{\pi}(s,a)$ is shown in equation (1). Here, s_0 and a_0 are the initial state and action respectively, γ is the *discount factor*, t represents the time step and R is the reward function. $Q^{\pi}(s,a)$ contains encoded information regarding the long-term effects of choosing an action a in state s.

$$Q^{\pi}(s,a) = \mathbb{E}_{\pi}[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t, s_{t+1})|s_0 = s, a_0 = a] \quad (1)$$

2.3 Criticality in Reinforcement Learning

A critical state is one where the chosen action significantly impacts the outcome. Such states exhibit high variability in the expected return, which corresponds to the variance of the Q-function (Spielberg and Azaria, 2019; Karino et al., 2020; Spielberg and Azaria, 2022). Based on this, the current study uses the variance of the Q-function across all actions as the criticality metric C, as shown in equation (2).

$$C = Var[Q^{\pi}(s,a)] \quad (2)$$

2.4 Policy-Dependent Criticality

As shown in equation (2), the criticality metric depends on the Q-function which is policy-dependent, i.e., $Q^{\pi}(s,a)$. Consequently, the criticality of a state evolves during training as Q-values are updated (Spielberg and Azaria, 2019). This paper studies this evolution to understand the agent's perspective of criticality as an agent's ability to detect, handle and retain critical states, alongside its performance, is essential for building trust in RL systems (Huang et al., 2018).

2.5 Exploration and Criticality

As discussed in sub-section 2.4, criticality is policy-dependent since the Q-function $Q^{\pi}(s,a)$ evolves with policy updates. We hypothesize that for effective criticality quantification, the agent has to satisfy the following two conditions,

1. Sufficiently visit critical states.
2. Understand the effect of different actions in critical states, including the consequences of incorrect actions. So, diversity in experience is crucial for effective criticality quantification. Here, diversity of experience refers to the targeted exploration using the actions that give us a better understanding of the critical states.

The above conditions are primarily governed by the chosen exploration strategy. To study this relationship, we compare two strategies namely fixed exponential ε-decay (ε_{exp}) and adaptive ε-decay (ε_{ad}). The ε_{exp} strategy applies a fixed exponential decay to ε reducing it to a minimum value over time. When progress remaining p^i is explicitly available from the environment such as Highway-env (Leurent, 2018), ε_{exp} is decayed as shown in equation (3). Here, i is the episode, ε_{min} and ε_{max} are the minimum and maximum exploration rates and λ is a decay factor controlling the rate of decrease.

$$\varepsilon^i_{exp} = max(\varepsilon_{min}, \varepsilon_{max} \cdot e^{-\lambda\cdot(1-p^i)}), \quad (3)$$

$$\varepsilon^i_{ad} = \begin{cases} max(\varepsilon_{min}, \varepsilon^{i-1}_{ad}.\lambda), & R^i_{avg} > R^{i-1}_{avg_{best}} \\ min(\varepsilon^{i-1}_{ad}, \varepsilon^{i-1}_{ad}/\lambda), & R^i_{avg} \leq R^{i-1}_{avg_{best}} \end{cases} \quad (4)$$

In contrast, ε_{ad} adjusts ε based on performance as shown in equation (4). Here, R^i_{avg} is the average reward until the i^{th} episode, and $R^{i-1}_{avg_{best}}$ is the best average reward up to the $(i-1)^{th}$ episode. By adjusting exploration based on performance, ε_{ad} is expected to encourage further exploration in critical states, improving the diversity of experiences and aiding in better criticality quantification.

3 RELATED WORK

3.1 Fundamental Research

Criticality in RL was first introduced as the variability in the expected return across actions (Spielberg and Azaria, 2019). The paper introduces the *Criticality-Based Varying Stepnumber (CVS)* algorithm that utilizes criticality to adapt the step number in n-step algorithms like *n-step SARSA*. State Importance (SI), introduced in (Karino et al., 2020), uses Q-value variance to identify critical states, promoting exploitation in critical states and exploration in non-critical ones. Here, results in Atari and Walker2D showed faster learning compared to ε-greedy. In (Liu et al., 2023), a *Deep State Identifier (DSI)* method is introduced that detects critical states from video trajectories using return prediction and masking, validated on grid-world and Atari environments.

3.2 Autonomous Driving and Trust

In (Huang et al., 2018), the authors show that identifying and acting safely in critical states improves trust in black-box policies. In (Hwang et al., 2022), the authors introduce *Critical Feature Extraction (CFE)* which improves Inverse Reinforcement Learning (IRL) efficiency by identifying critical states from both positive and negative demonstrations, reducing computation while maintaining quality.

3.3 Adversarial Attacks

Adversarial strategies like *strategically-timed attacks* disrupt RL by targeting critical states, achieving similar performance degradation as continuous attacks with minimal intervention (Lin et al., 2017). Statistical metrics in (Kumar et al., 2021) showed that targeting critical states which make up about 1% of the states, reduced agent's performance by 40%.

3.4 Human-in-the-Loop RL

The studies (Ju, 2019), (Ju et al., 2020) and (Ju et al., 2021) use criticality in pedagogy to enhance learning in interactive learning environments, such as *Intelligent Tutoring Systems (ITS)*. *Criticality-Based Advice (CBA)* (Spielberg and Azaria, 2022) integrates human advice for critical states, improving learning efficiency. Here, Plain CBA requests advice when criticality exceeds a threshold, while Meta CBA combines criticality with existing strategies, outperforming traditional advice in grid world and Atari environments.

3.5 Literature Gap

Despite significant work on criticality, no study explores its evolution during training. We believe that studying this evolution will further enhance our understanding of what factors contribute to effective criticality quantification. In the current paper, we wish to address this gap by taking a deep dive into the evolution of criticality by closely studying the relationship between exploration and criticality.

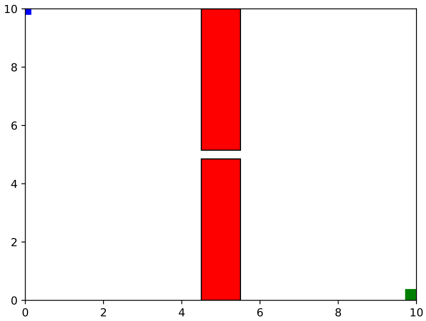


Figure 1: The continuous cliff maze environment where the agent is marked blue, the goal is green and the danger zones (cliffs) are red. Here, the agent starts from the top left corner and must navigate through the cliffs in the middle to reach the goal in the bottom right corner. The agent, when passing through the narrow passage is restricted from taking actions in other directions making this region highly critical. The narrow gap is kept at 0.3 units of vertical width and an action step size of 0.5 is used.

4 EXPERIMENTAL SETUP

4.1 Environments

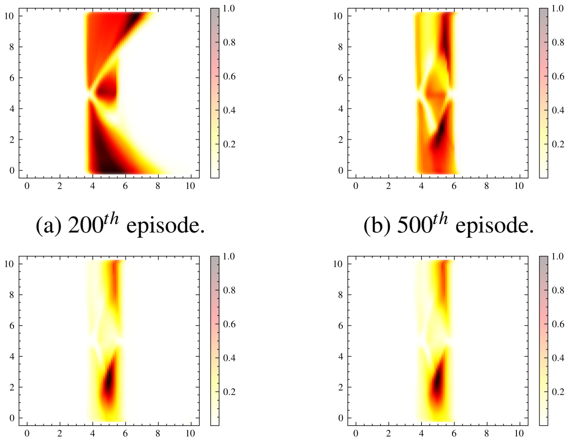
To study the evolution of criticality and the effect of exploration strategies, we use two environments namely the Continuous Cliff Maze and Highway-env (Leurent, 2018). The lightweight and interpretable Continuous Cliff Maze tests our hypothesis on exploration and criticality, while Highway-env extends the study to autonomous driving scenarios.

4.1.1 Continuous Cliff Maze

The Continuous Cliff Maze as shown in figure 1, is a modified version of the discrete maze in (Karino et al., 2020), with a continuous state space and discrete action space. It provides an intuitive, static environment to study criticality in a continuous state space using DRL. The central narrow gap and surrounding cliffs represent highly critical regions where action choices are restricted. The agent receives -1 reward for entering cliffs and $+10$ for reaching the goal.

4.1.2 Highway

The Highway-env (Leurent, 2018) is a collection of environments to train and test DRL agents in autonomous driving scenarios. It offers multiple environments like *Merge*, *Intersection* and *Roundabout*. In the current paper, we choose the *Highway* environment to study criticality quantification in highway autonomous driving scenarios. In the Highway environment, the state space is continuous and we choose discrete meta-actions namely $a = \{0 : Lane\_left, 1 :$



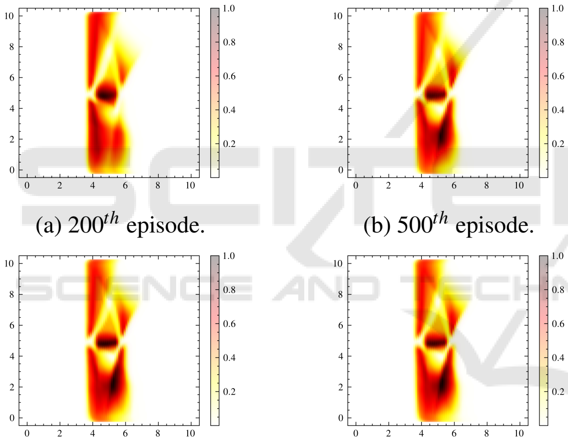
(a) 200^{th} episode.

(b) 500^{th} episode.

(c) 1200^{th} episode.

(d) 1900^{th} episode.

Figure 2: Normalized heatmaps of the evolution of criticality in the continuous cliff maze environment for four model checkpoints of one of the trials of the DQN$_{\text{exp}}$ model. The images show a clear unlearning of the criticality of the central narrow cliff and the surrounding regions.



(a) 200^{th} episode.

(b) 500^{th} episode.

(c) 1200^{th} episode.

(d) 1900^{th} episode.

Figure 3: Normalized heatmaps of the evolution of criticality in the continuous cliff maze environment for four model checkpoints of one of the trials of the DQN$_{\text{ad}}$ model. The images show the retention of critical information about the central narrow cliff and surroundings.

$Idle, 2 : Lane\_right, 3 : Faster, 4 : Slower\}$. Once we train the agent, we test it on four hand-crafted critical scenarios as shown in figure 4, to study the evolution of criticality.

4.2 Algorithm

Given the two environments in sub-section 4.1, which both have a continuous state space and a discrete action space, we train DRL agents using a *Deep Q-Network (DQN)* algorithm (Mnih, 2013). The output Q-values are used to quantify the criticality of a state s using equation (2). For exploration, we employ ε_{exp} and ε_{ad}, denoting the resulting models as DQN$_{\text{exp}}$ and DQN$_{\text{ad}}$, respectively. These models are used to study the effect of exploration strategies on criticality quantification.

5 RESULTS AND DISCUSSION

5.1 Continuous Cliff Maze

We train five DQN$_{\text{exp}}$ and DQN$_{\text{ad}}$ models for $2,000$ episodes, clipping ε between 0.9 and 0.01, with a step limit of $5,000$ and a replay buffer of $50,000$. Model checkpoints are saved every 100^{th} episode to study the evolution of criticality.

5.1.1 Evolution of Criticality

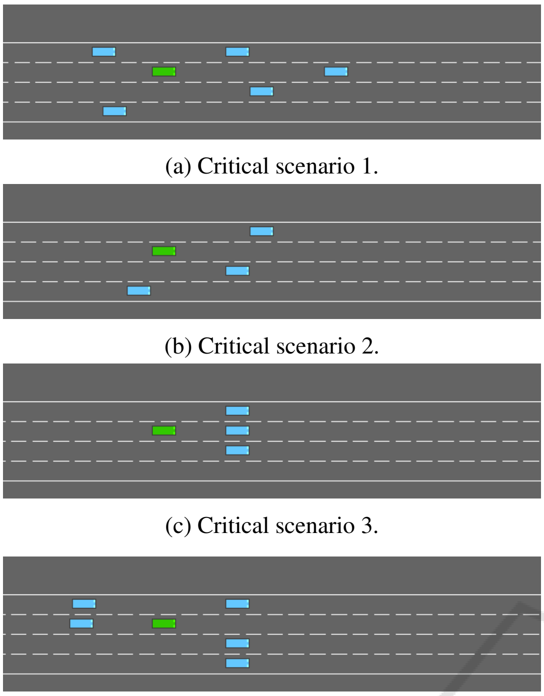
The evolution of criticality is analyzed using criticality heatmaps. Figure 2 shows that DQN$_{\text{exp}}$ exhibits unlearning of criticality in the given environment, while figure 3 demonstrates that DQN$_{\text{ad}}$ appears to retain criticality throughout training. To investigate this phenomenon, we analyze performance, critical state visitations and action diversity.

5.1.2 Performance Study

Figure 5 shows the epsilon decay curves. The *Simple Moving Average (SMA)* reward curves in figure 6 converge around $1,200$ episodes. Despite differences in epsilon decay, no significant performance difference is observed, ruling out performance as the cause of criticality unlearning in DQN$_{\text{exp}}$.

5.1.3 Critical State Visitations and Action Diversity

An agent must substantially visit critical states to gain knowledge of them. Figure 7 shows that DQN$_{\text{exp}}$ and DQN$_{\text{ad}}$ visit the central narrow gap, the Region of Interest (ROI), about $6,000$ times each thus ruling out the number of visitations as the reason for the unlearning. Figures 8 and 9 illustrate action selection strategies. DQN$_{\text{ad}}$ greatly prefers *Right* and *Left* actions, showing a preference for those actions that facilitate extended exploration of critical states. While DQN$_{\text{exp}}$ selects actions more uniformly, including Up and *Down*, which terminate the episode. This difference in action selection appears to contribute to DQN$_{\text{ad}}$'s ability to retain criticality, whereas DQN$_{\text{exp}}$ shows a tendency to lose it. This suggests that a targeted diversity in experiences may contribute to effective crit-



(a) Critical scenario 1.

(b) Critical scenario 2.

(c) Critical scenario 3.

(d) Critical scenario 4.

Figure 4: The four hand-crafted critical scenarios in the Highway environment with the ego vehicle in green and the surrounding vehicles in blue. Here the ego (agent) is restricted in its actions and has to carefully navigate through the surrounding vehicles without crashing. (a) The ego can either overtake on the right or slowly pass through the vehicles in front. (b) The ego has to pass through the vehicles in front (c) The ego has to overtake on the right or slow down. (d) The ego has to pass through the vehicles in front without slowing down to prevent a collision with the vehicles at the back.

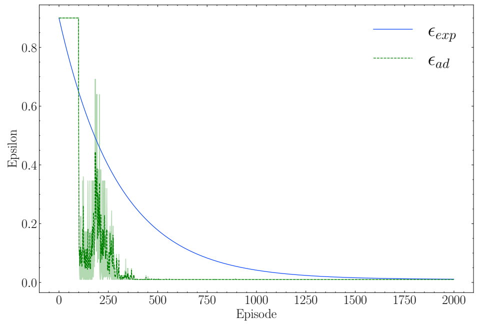


Figure 5: The epsilon decay curves for five trials of DQN_{exp} and DQN_{ad} models respectively. The plot is represented using mean and $20-80$ percentile bands.

icality quantification. Another important thing to note is that although figure 5 shows that ε_{ad} decays more rapidly than ε_{exp}, figure 9 indicates longer episodes experienced by DQN_{ad} resulting in enhanced exploration.

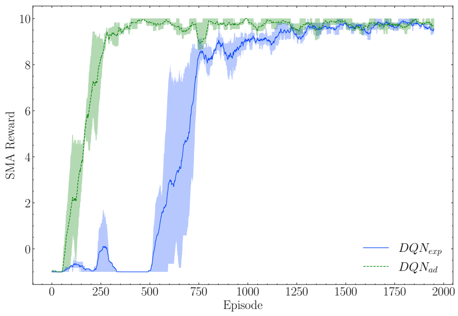


Figure 6: The SMA reward curves for five trials of DQN_{exp} and DQN_{ad} models respectively, where the window size to calculate the average is set to 50 episodes. The plot is represented using mean and $20-80$ percentile bands.

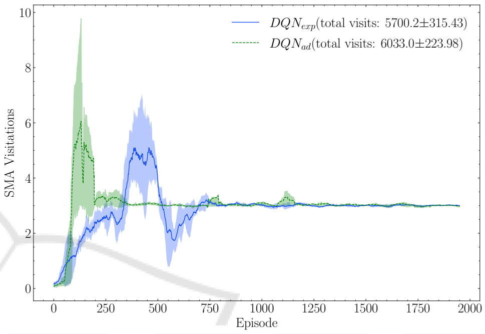


Figure 7: Critical states visitation during training presented as an SMA curve. Here, the ROI is the central narrow gap between the two cliffs. The window size for SMA is fixed to 50 episodes.

Given the similar performance of both models, the retention of criticality by DQN_{ad} suggests that it may be a more reliable and trustworthy choice under the given conditions.

5.2 Highway

The findings from the cliff maze environment are extended to a more complex Highway environment. We train five DQN_{exp} and five DQN_{ad} models using the standard DQN implementation from *stable baselines3* (Raffin et al., 2021). Training is conducted for $50,000$ steps with ε clipped between 1.0 and 0.01. Model checkpoints are saved every 100^{th} episode and criticality is calculated for four hand-crafted scenarios given in figure 4.

The evolution of criticality is analyzed as mean and variance curves using equations (5) and (6). For each scenario, criticality $C_m^j = Var[Q^{\pi}(s,a)]_m^j$ is computed at the j^{th} checkpoint across all m trials. The mean μ_C^j reflects overall trends, while variance $Var[C]^j$ captures variability. These results are illustrated in the *Evolution of Criticality plot (EC-plot)*, which conveys curve trends, with the Y-scale being

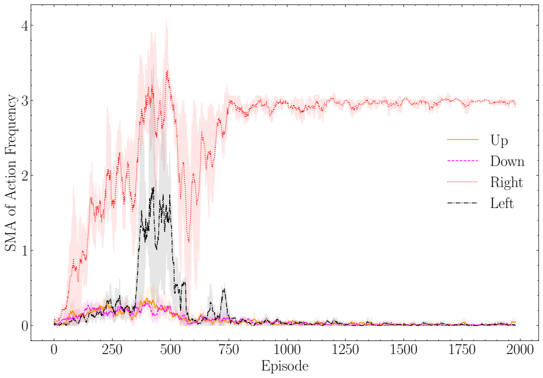


Figure 8: The SMA curve of action frequency of DQN$_{\text{exp}}$ agents during training, with a window size of 20. The ROI is the central narrow gap between the two cliffs. The plot is presented as mean and $20-80$ percentile bands for each action. The Y-axis shows the average number of times each action was chosen by the DQN$_{\text{exp}}$ agents per episode. The plots show the DQN$_{\text{exp}}$ agents actively choosing Up and $Down$ actions until the end of 750 episodes.

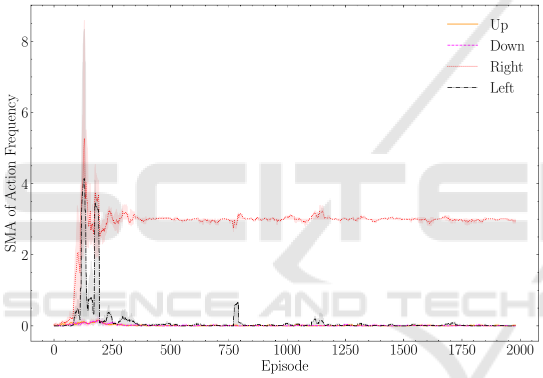


Figure 9: The SMA curve of action frequency of DQN$_{\text{ad}}$ agent during training, with a window size of 20. The ROI is the central narrow gap between the two cliffs. The plot is presented as mean and $20-80$ percentile bands for each action. The plots show that the DQN$_{\text{ad}}$ agents having a very high preference for $Right$ and $Left$ actions.

proportional to Q-values but not of significance.

$$\mu_C^j = \frac{1}{m} \sum_m Var[Q^\pi(s,a)]_m^j \tag{5}$$

$$Var[C]^j = \frac{1}{m} \sum_m (Var[Q^\pi(s,a)]_m^j - \mu_C^j)^2 \tag{6}$$

5.2.1 Performance Study

The ε_{exp} and ε_{ad} decay curves are shown in Figure 10 as mean and $20-80$ percentile bands. The ε_{ad} decay exhibits a step-like behaviour, reducing exploration only when the average reward improves as shown in equation (4)). This promotes extended exploration, enhancing action diversity in critical states in contrast to ε_{exp}.

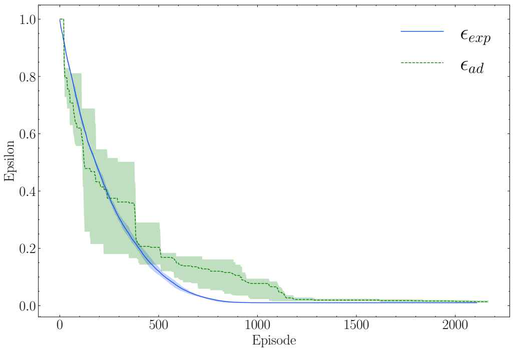


Figure 10: The ε_{exp} and ε_{ad} decay curves with mean and $20-80$ percentile bands.

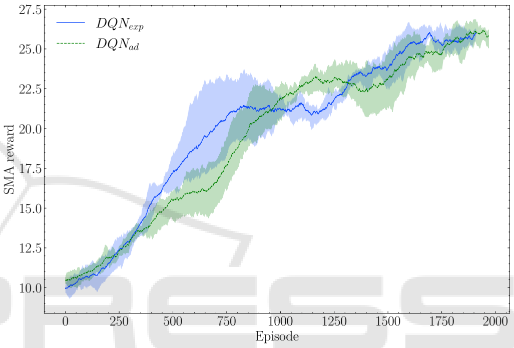


Figure 11: The SMA reward curves for DQN$_{\text{exp}}$ and DQN$_{\text{ad}}$ as mean and $20-80$ percentile bands, with a window size of 200 episodes.

The SMA reward curves in figure 10 show higher variability for DQN$_{\text{exp}}$ due to identical decay schedules across trials, leading to differences in experience diversity. For DQN$_{\text{ad}}$, the reward curves are more stable despite varying decay behaviour, indicating consistent learning. By the end of $2,000$ episodes, both models achieve similar performance enabling a fair comparison.

5.2.2 EC-plots

The EC-plot for DQN$_{\text{exp}}$ in figure 12 shows a sharp increase in criticality from episode 1 to $1,000$, aligning with the exploration phase as shown in figures 10 and 11. The criticality peaks around episode 900 in scenario 1, followed by a gradual decrease with low variance. For scenarios 2 and 4, criticality drops sharply after episode $1,000$ with fluctuations, while scenario 3 shows a gradual decline. These trends mirror the unlearning behaviour observed in the continuous cliff maze environment.

Comparing the epsilon decay, rewards, and EC-plot reveals that criticality unlearning occurs after the reduced exploration phase around episode 1,000, even as model performance continues to improve. This highlights that criticality can be unlearned, a crucial consideration for real-world applications like autonomous driving, where retaining criticality is essential for overall safety and trust.

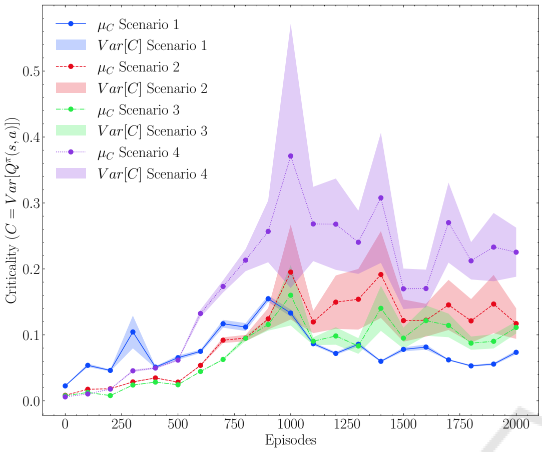


Figure 12: The EC-plot for the DQN$_{\text{exp}}$ model for the four hand-crafted critical scenarios. A general unlearning of criticality can be observed for all four scenarios during training.

The EC-plot for DQN$_{\text{ad}}$ depicted in figure 13 shows that the ε_{ad} strategy retains awareness of criticality throughout training, with criticality increasing gradually in scenarios 2 and 3 and sharply in scenario 4, though with high variance. However, unlearning persists for scenario 1 after episode 1,500, indicating room for improvement in the ε_{ad} schedule and developing more advanced exploration strategies that guarantee criticality retention. Given the similar performance of DQN$_{\text{exp}}$ and DQN$_{\text{ad}}$, the latter is preferable as it retains criticality. For real-world applications like autonomous driving, agents must not only perform well but also retain awareness of criticality to ensure safe decision-making which makes DQN$_{\text{ad}}$ the suitable choice.

6 CONCLUSIONS

This study provides insights into the evolution of criticality during training, emphasizing the importance of sufficient state visitations and diverse experiences for effective criticality quantification. By comparing two exploration strategies, ε_{exp} and ε_{ad}, we observe that the ε_{ad} strategy tends to retain criticality throughout training, while ε_{exp} models exhibit a tendency for unlearning, despite achieving comparable performance. This behaviour is observed consistently across both a static continuous cliff maze environment and a more dynamic, complex Highway environment suggesting that ε_{ad} may be more reliable for safety-critical applications such as autonomous driving. While our findings suggest that the ε_{ad} strategy retains criticality better than ε_{exp}, we note that the effectiveness of an exploration strategy can depend on the environment and training dynamics. A more advanced exploration strategy that proactively ensures targeted diversity is needed. To support further research into the topic, the code used in the current research is available at our GitHub repository [https://github.com/aimotion-autonomous-driving-cluster/The-Evolution-of-Criticality-in-Deep-Reinforcement-Learning.git].

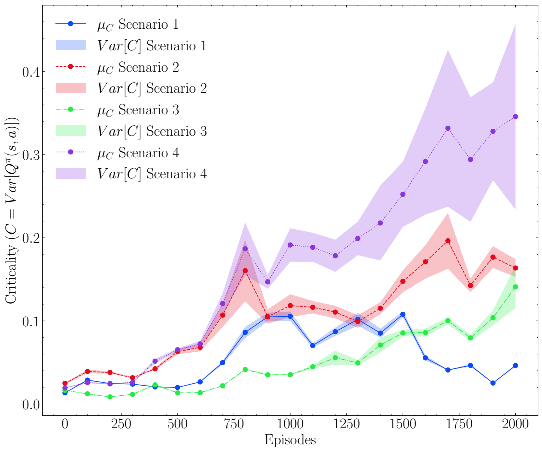


Figure 13: The EC-plot of the DQN$_{\text{ad}}$ model for four hand-crafted critical scenarios. We observe sharp sustained and increasing criticality curves for critical scenarios 2, 3, and 4 while unlearning is still observed for critical scenario 1.

In the future, we aim to use more robust criticality metrics for scenario generation (Westhofen et al., 2023) and study criticality in entropy-based RL methods like Soft Actor-Critic (SAC). Additionally, we will investigate the interplay between criticality and model uncertainty, as higher $Var[Q^{\pi}(s,a)]$ values may reflect uncertainty rather than criticality and high uncertainty need not correspond to higher criticality.

REFERENCES

Afsar, M. M., Crump, T., and Far, B. (2022). Reinforcement learning based recommender systems: A survey. *ACM Computing Surveys*, 55(7):1–38.

Boute, R. N., Gijsbrechts, J., Van Jaarsveld, W., and Vanvuchelen, N. (2022). Deep reinforcement learning for inventory control: A roadmap. *European Journal of Operational Research*, 298(2):401–412.

Chen, X., Yao, L., McAuley, J., Zhou, G., and Wang, X. (2021). A survey of deep reinforcement learning in recommender systems: A systematic review and future directions. *arXiv preprint arXiv:2109.03540*.

Gu, S., Holly, E., Lillicrap, T. P., and Levine, S. (2016). Deep reinforcement learning for robotic manipulation. *arXiv preprint arXiv:1610.00633*, 1:1.

Huang, S. H., Bhatia, K., Abbeel, P., and Dragan, A. D. (2018). Establishing appropriate trust via critical states. In *2018 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 3929–3936. IEEE.

Hubbs, C. D., Li, C., Sahinidis, N. V., Grossmann, I. E., and Wassick, J. M. (2020). A deep reinforcement learning approach for chemical production scheduling. *Computers & Chemical Engineering*, 141:106982.

Hwang, M., Jiang, W.-C., and Chen, Y.-J. (2022). A critical state identification approach to inverse reinforcement learning for autonomous systems. *International Journal of Machine Learning and Cybernetics*, 13(5):1409–1423.

Ju, S. (2019). Identify critical pedagogical decisions through adversarial deep reinforcement learning. In *In: Proceedings of the 12th International Conference on Educational Data Mining (EDM 2019)*.

Ju, S., Zhou, G., Abdelshiheed, M., Barnes, T., and Chi, M. (2021). Evaluating critical reinforcement learning framework in the field. In *International conference on artificial intelligence in education*, pages 215–227. Springer.

Ju, S., Zhou, G., Barnes, T., and Chi, M. (2020). Pick the moment: Identifying critical pedagogical decisions using long-short term rewards. *International Educational Data Mining Society*.

Karino, I., Ohmura, Y., and Kuniyoshi, Y. (2020). Identifying critical states by the action-based variance of expected return. In *International Conference on Artificial Neural Networks*, pages 366–378. Springer.

Kumar, R. P., Kumar, I. N., Sivasankaran, S., Vamsi, A. M., and Vijayaraghavan, V. (2021). Critical state detection for adversarial attacks in deep reinforcement learning. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 1761–1766. IEEE.

Leurent, E. (2018). An environment for autonomous driving decision-making. https://github.com/eleurent/highway-env.

Li, G., Li, S., Li, S., Qin, Y., Cao, D., Qu, X., and Cheng, B. (2020). Deep reinforcement learning enabled Decision-Making for autonomous driving at intersections. *Automotive Innovation*, 3(4):374–385.

Lin, Y.-C., Hong, Z.-W., Liao, Y.-H., Shih, M.-L., Liu, M.-Y., and Sun, M. (2017). Tactics of adversarial attack on deep reinforcement learning agents. *arXiv preprint arXiv:1703.06748*.

Liu, H., Zhuge, M., Li, B., Wang, Y., Faccio, F., Ghanem, B., and Schmidhuber, J. (2023). Learning to identify critical states for reinforcement learning from videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1955–1965.

Mnih, V. (2013). Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Panzer, M. and Bender, B. (2022). Deep reinforcement learning in production systems: a systematic literature review. *International Journal of Production Research*, 60(13):4316–4341.

Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. *Journal of Machine Learning Research*, 22(268):1–8.

Ravi Kiran, B., Sobh, I., Talpaert, V., Mannion, P., Al Sallab, A. A., Yogamani, S., and Pérez, P. (2022). Deep reinforcement learning for autonomous driving: A survey. *IEEE Trans. Intell. Transp. Syst.*, 23(6):4909–4926.

Santorsola, A., Maci, A., Delvecchio, P., and Coscia, A. (2023). A reinforcement-learning-based agent to discover safety-critical states in smart grid environments. In *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*. IEEE.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., et al. (2016). Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484–489.

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2017). Mastering chess and shogi by self-play with a general reinforcement learning algorithm. *arXiv preprint arXiv:1712.01815*.

Spielberg, Y. and Azaria, A. (2019). The concept of criticality in reinforcement learning. In *2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 251–258. IEEE.

Spielberg, Y. and Azaria, A. (2022). Criticality-based advice in reinforcement learning. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Westhofen, L., Neurohr, C., Koopmann, T., Butz, M., Schütt, B., Utesch, F., Neurohr, B., Gutenkunst, C., and Böde, E. (2023). Criticality metrics for automated driving: A review and suitability analysis of the state of the art. *Archives of Computational Methods in Engineering*, 30(1):1–35.