

# Driving Decision-making Method for Autonomous Vehicles based on Anterior Cingulate Cortex Neural Regulation Mechanism

Jialin Liu

1. School of Mechanical and  
Automotive Engineering  
Shanghai University Of Engineering  
Science

Shanghai, China

2. Key Laboratory of Operation Safety  
Technology on Transport Vehicles

Ministry of Transport

Beijing, China

m315123403@sues.edu.cn

Xiaolan Wang\*

School of Mechanical and Automotive  
Engineering  
Shanghai University Of Engineering  
Science

Shanghai, China

jlu\_wangxiaolan@aliyun.com

Yue Li

1. Key Laboratory of Operation Safety  
Technology on Transport Vehicles  
Ministry of Transport

Beijing, China

2. Research Institute of Highway  
Ministry of Transport

Beijing, China

li.yue@rioh.cn

**Abstract**—Reinforcement learning algorithms have been widely applied in the field of autonomous driving. However, the choice of policy during the reinforcement learning process significantly impacts learning outcomes. Existing stochastic policy methods lack flexibility, preventing autonomous vehicles from quickly learning optimal strategies in complex traffic environments. This paper proposes a model that simulates the regulatory mechanism of the Anterior Cingulate Cortex (ACC) to balance exploration and exploitation. The model observes through the Medial Prefrontal Cortex (MPFC) module, evaluates using graph convolutional neural networks, and transmits dopamine signals of reward errors to the ACC module. Feedback neurons compute vigilance, and the LPFC adjusts action probabilities based on vigilance and action values. Experimental results validate that this approach achieves a balance in action selection strategies for autonomous vehicles, thereby improving the quality of actions taken by the agent.

**Keywords**—autonomous driving, reinforcement learning(RL), Anterior cingulate cortex(ACC), learning strategies, exploration-exploitation tradeoff

## I. INTRODUCTION

In recent years, traffic accidents caused by human errors have become a serious global issue [1]. The emergence of autonomous driving vehicles is seen as an effective solution to mitigate driver burden and has garnered widespread attention worldwide [2]. The decision-making system of autonomous driving systems is responsible for considering interactions with other traffic participants to select optimal actions, ensuring safe driving while completing tasks efficiently. The quality of these decisions directly impacts the safety and efficiency of vehicle operations.

The advancements in reinforcement learning (RL) have greatly propelled the development of autonomous driving decision-making technology. RL, based on Markov decision processes (MDP), employs closed-loop learning, utilizing reward functions as incentives and employing exploration-exploitation strategies for autonomous iterative learning [3], gradually improving decision-making capabilities. In complex traffic scenarios, RL can explore environments, extract feature information, and abstract hidden mappings of

optimal strategies, particularly showing immense potential in situations where complete environmental observation or explicit scene representation is not feasible. Therefore, RL plays a crucial role in the field of autonomous driving decision-making.

Reinforcement learning (RL) commonly utilizes stochastic policies such as greedy policy [4] and  $\epsilon$ -greedy policy [5]. The greedy policy is a deterministic policy where the agent selects the action with the highest action value, assigning a probability of 1 to the action with the maximum value and 0 to all others. On the other hand, the  $\epsilon$ -greedy policy is widely used to balance exploration and exploitation. With probability  $1-\epsilon$ , the agent selects the best-known action, while with probability  $\epsilon$ , it chooses a random action to explore potentially better options. This policy has shown effective results in deep reinforcement learning. However, during the exploration phase,  $\epsilon$ -greedy policy assigns equal probabilities to all actions, potentially leading to negative impacts on overall performance when the agent selects poorly performing actions.

To overcome these shortcomings, this paper proposes a novel neurophysiological motivation model to address the trade-off between exploration and exploitation in Q-learning. The model simulates the regulatory function of the Anterior Cingulate Cortex (ACC) between the Medial Prefrontal Cortex (MPFC) and Lateral Prefrontal Cortex (LPFC). It employs graph convolutional neural networks to extract scene topological information and uses an actor-critic framework to evaluate states. During feedback, the ACC calculates prediction errors to update internal models, driving learning in the LPFC. When making decisions, individuals continuously weigh the likelihood of utilizing existing options versus exploring other options, aiming to select the optimal action.

## II. METHODS AND PRINCIPLES

### A. Neurophysiological Basis

Reference [6] indicates that the Anterior Cingulate Cortex (ACC) plays a crucial role in goal-directed behavior. The ACC integrates information based on signals received from other modules, calculating the cognitive control costs and expected rewards of cognitive control to select actions that

---

\*Corresponding Author

This work was supported by the Opening Project of Key Laboratory of operation safety technology on transport vehicles, Ministry of Transport, PRC (KFKT2021-01), and partly supported by the Project of National Natural Science Foundation of China (no. 52172371).

maximize expected rewards. Substantial evidence suggests that ACC activity responds to increases in unsigned prediction errors and error detection. Hence, the ACC error likelihood theory has been developed.

The error likelihood model was proposed by Brown and Braver [7] in 2005. This model suggests that the Anterior Cingulate Cortex (ACC) is involved in calculating the likelihood of errors, even in situations without actual errors or response conflicts. According to this theory, the ACC responds to activity in other brain regions and sends warning signals to them. These signals prompt those regions to perform error detection, and their magnitude depends on the level of prediction error in the given environment. This determines whether cognitive control needs to be implemented.

The prediction error signals of the ACC are driven by the LPFC learning [8], while the LPFC regulates specific predictions [9] generated by the MPFC/ACC. The LPFC primarily maintains representations of stimuli that coincide with prediction errors in working memory, whereas the MPFC/ACC generates these prediction errors, similar to TD errors in computational temporal difference learning. As noted by Botvinick [10] et al., "the occurrence of pain and the indication of error feedback are signals of the same category, all of which indicate that the current distribution of attention has failed to prevent negative outcomes." In other words, as cognitive conflicts and erroneous outcomes increase, the ACC becomes active due to the need for behavioral adjustments. This increased activation of the ACC leads to increased activity in the dorsolateral prefrontal cortex (LPFC), which is primarily responsible for executing cognitive control to adjust behavior.

Rushworth and Kolling [11], among others, applied an optimal foraging model from ecology to humans, known as the foraging value theory (FVT). Inspired by behavioral ecology, FVT posits that many natural situations differ from common binary choice tasks in laboratories, as they do not involve two clearly defined options. In such contexts, FVT suggests that individuals continuously weigh the value of persisting in ongoing actions against the potential for switching to foraging opportunities. The ACC is involved in evaluating foraging values in the environment, rather than simply utilizing current resource areas. Electrophysiological recordings in non-human primates and neuroimaging studies in humans have shown that ACC activity increases when there is a need to shift from exploitation of the environment to exploration. These findings support the theory that the ACC monitors the value of alternative actions and compares them with current actions, indicating that switching to exploring other options may be more valuable than exploiting existing ones (such as default options).

These functions can be mapped to the learning processes of artificial agents. Therefore, in this study, we simulate ACC regulatory mechanisms to design a decision system for autonomous vehicles that balances between exploration and exploitation.

### B. Problem Formulation

We employ an undirected graph to model the interaction scenario, where each node represents a vehicle, and the edges

convey interaction information between pairs of vehicles. Specifically, this undirected graph can be denoted as  $G = \{N, E\}$ , which serves as our model input data, where  $N = \{n_1, n_2, \dots, n_n\}$ ,  $E = \{e_{ij}, i, j \in \{1, 2, \dots, n\}\}$ ,  $n$  denotes the number of vehicles. The output action space consists of discrete lane-change instructions.

$$a = [\text{left}, \text{straight}, \text{right}] \quad (1)$$

Markov Decision Processes are a classic method for modeling intelligent agents in complex environments. The experiments in this paper are conducted using the SUMO simulation platform, where the environment is fully observable. Therefore, the decision problem of autonomous vehicles in dynamic environments can be modeled as an MDP. An MDP is typically represented by a tuple  $(S, A, P, R, \gamma)$ , where  $S$  is the state space,  $A$  is the action space,  $P$  is the transition probability function,  $R$  is the reward function, and  $\gamma$  is the discount factor. In this environment, the agent selects actions based on the current state. When the agent performs an action in the environment, the environment updates its state in the next time step and provides a single-step reward. This process repeats until the task is completed or terminated. Modeling the lane-changing problem in simple scenarios and applying it to autonomous vehicles has shown promising results [12]. Moreover, this process exhibits similarities to the neural mechanisms introduced in Section 2.1.

### C. ACC Neural Modulation System

Based on the above principles, we designed an ACC neural modulation system that operates between the MPFC and LPFC, as illustrated in Fig. 1. Information from the environment is observed by the vehicle and input to the MPFC, which makes a state judgment. Subsequently, the ventral tegmental area (VTA) in the midbrain receives a reward  $r$ , and the VTA calculates a reward error  $\delta$  based on  $r$ . Meanwhile, the feedback neurons in the ACC module calculate vigilance  $\beta^*$  based on  $\delta$ . The LPFC then uses vigilance to determine the action probabilities  $P$ , which represent the probability of selecting each action.

Within the MPFC, environmental observation involves primarily graph convolution processes, where the constructed node matrix and adjacency matrix are input to the Graph Neural Network (GNN) module. This module comprises fully connected layers and graph convolution layers.

This process can be described as:

$$G_t = \phi^{GCN}(N_t, E_t) \quad (2)$$

Where  $N_t$  represents the node matrix,  $E_t$  denotes the adjacency matrix, and  $\phi^{GCN}$  represents a neural network with graph convolutional layers.

The graph convolutional features are fed into the DRL (Deep Reinforcement Learning) module to obtain the  $Q$ -values for lane-changing actions. The derivation process of the  $Q$ -values is as follows:

$$Q(s_t, a_t) = \phi^{DRL}(G_t) \quad (3)$$

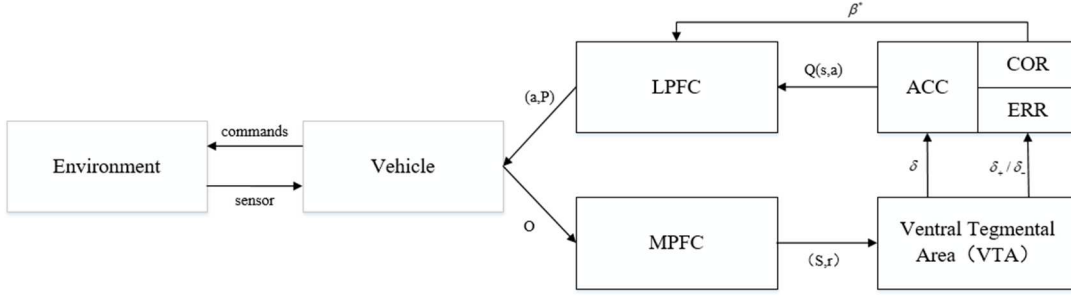


Fig. 1. Framework of ACC Neural Modulation System

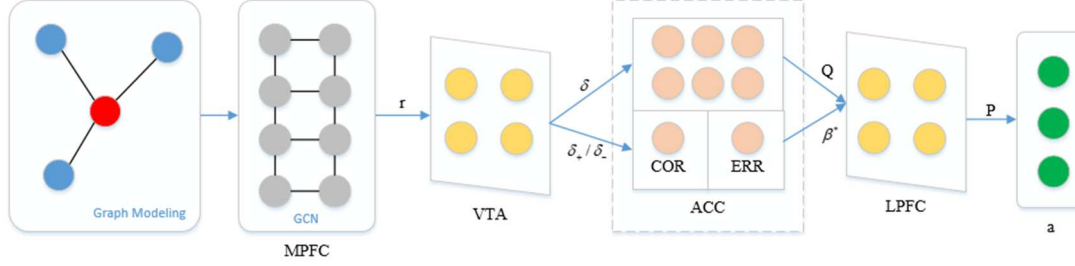


Fig. 2. The neural modulation system network model of the ACC

Where  $\phi^{\text{DRL}}$  represents the policy neural network of Deep Reinforcement Learning.

After the agent selects an action, the Q-values are adjusted based on the reward, calculating the reward prediction error  $\delta(t)$ :

$$\delta(t) = r_t + \gamma \max_a Q(s_t, a_t) - Q(s_{t-1}, a_{t-1}) \quad (4)$$

Where  $r_t$  represents the real-time reward obtained from interacting with the environment, and  $\gamma$  represents the discount factor.

The reward prediction error  $\delta$  not only affects the Q-values of actions but also influences a group of feedback-like neurons in the ACC (Anterior Cingulate Cortex) module. When the reward prediction error  $\delta$  is less than 0, the ERR neurons respond, whereas the COR neurons respond when  $\delta$  is greater than 0. This is specifically represented as:

$$\begin{cases} \delta_+(t) = \delta(t), & \text{if } \delta(t) \geq 0 \\ \delta_-(t) = \delta(t), & \text{if } \delta(t) < 0 \end{cases} \quad (5)$$

After receiving feedback, this module transmits a vigilance signal  $\beta^*$  to the LPFC (Lateral Prefrontal Cortex), simulating how strongly the human brain reacts to current actions. This sensitivity is crucial for the module's responsiveness to behavior. The computation process of vigilance  $\beta^*$  is as follows:

$$\beta^*(t) \leftarrow \beta^*(t) + \mu_+ \delta_+(t) + \mu_- \delta_-(t) \quad (6)$$

Where  $0 < \beta^* < 1$ ,  $\mu_+$  and  $\mu_-$  represent the update rates.

After receiving vigilance, the LPFC computes the exploration rate  $\beta$  as follows:

$$\beta = \frac{\omega_1}{1 + \exp(\omega_2 \cdot [1 - \beta^*] + \omega_3)} \quad (7)$$

Here  $\beta \geq 0$ ,  $\omega_1$ ,  $\omega_2$  and  $\omega_3$  are constants.

LPFC adjusts the probability of selecting an action based on the size of the exploration rate:

$$P(a_i) = \frac{\exp(\beta \cdot Q(s, a_i))}{\sum_{j=1}^n \exp(\beta \cdot Q(s, a_j))} \quad (8)$$

where  $n$  represents the total number of actions available in state  $s$ , and  $P(a_i)$  is the probability of selecting action  $a_i$ .

The neural modulation system network model of the ACC is illustrated in Fig. 2. This model achieves a balance between exploration and exploitation for the agent during policy selection. When the exploration rate  $\beta$  is low, it significantly influences the Q-values of each action, resulting in similar overall action values and nearly equal probabilities for each action to be chosen. In this scenario, the agent tends to explore the environment more. Conversely, when the exploration rate  $\beta$  is high, the influence of each action's intrinsic value on its overall value increases. This leads the agent to favor selecting the optimal action, thus exploiting the environment effectively.

### III. EXPERIMENTAL RESULTS

In this study, experiments were conducted using greedy policy,  $\epsilon$ -greedy policy, and ACC-adjusted policy. Initially, the agent employed a completely exploratory approach, and after training for a period, the agent was trained using the three strategies. The training rewards during the training process in the highway scenario are shown in Fig. 3. It can be observed from the graph that the reward curve obtained using the ACC-adjusted policy is slightly higher than that of the greedy policy and policy. The average rewards during training were 3289.64 for the greedy policy, 3302.55 for the  $\epsilon$ -greedy policy, and 3429.51 for the ACC-adjusted policy. During testing, the average rewards were 3277.84 for the greedy policy, 3298.17 for the  $\epsilon$ -greedy policy, and 3436.82 for the ACC-adjusted policy. Both during training and testing, the ACC-adjusted policy yielded the highest rewards, indicating that it can effectively achieve a balance between exploration and exploitation.

Given that the distribution of vehicles in the highway scenario is relatively sparse and the environment changes are relatively smooth, this paper further verifies the robustness and generalization of the model by conducting new training and testing in a dense intersection scenario. The performance is then compared with the better-performing strategies. The two strategies are combined with the widely used A3C, PPO, and DDPG algorithms. The training rewards in the roundabout intersection scenario are shown in Fig. 4. From the reward curves, it can be observed that for the same algorithm, the reward values of algorithms using the ACC-adjusted policy

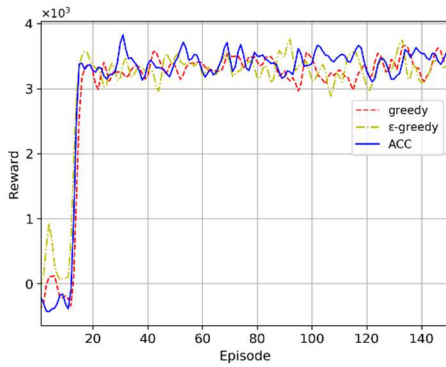


Fig. 3. The training rewards in the highway scenario

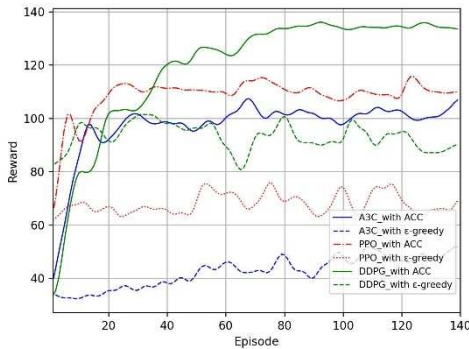


Fig. 4 The training rewards in the roundabout intersection scenario

are higher than those using the regular policy. Among the algorithms using the same policy, the DDPG algorithm performs the best, followed by the PPO algorithm, with the A3C algorithm performing the worst.

The experiments in the two scenarios demonstrate that the ACC-adjusted policy effectively accomplishes the action selection tasks for autonomous vehicle decision-making. Unlike traditional strategies that adjust action selection probabilities in a fixed manner, the ACC-adjusted policy is closer to human thinking, as it continuously and dynamically adjusts the selection policy to respond to changing environments. Compared to traditional strategies, this approach is less likely to fall into local optima and exhibits better adaptability in complex environments. However, the policy shows significant fluctuations in the early stages of training, which may be due to the high sensitivity of the policy to the numerous uncertainties present in the initial phase.

### IV. CONCLUSION

Based on the neural modulation mechanism of ACC, this paper proposes a method to balance exploration and exploitation. This method enables the agent to dynamically transition between exploring and exploiting the environment in autonomous driving scenarios. Compared to conventional greedy and  $\epsilon$ -greedy policy approaches, training and testing results demonstrate that this method achieves higher rewards and better learning effectiveness. It can enhance data quality to a certain extent, thereby enabling autonomous vehicles to make more accurate decisions. This policy can not only be widely applied in the field of autonomous driving but also combined with other deep reinforcement learning models, potentially extending its use to other areas of artificial intelligence beyond autonomous driving. However, the model's training is not very stable in the early stages, necessitating further refinement of the policy. Future research will focus on improving the policy and exploring the integration with studies on the prefrontal cortex, which is closely related to this module.

### REFERENCES

- [1] X. Xu, L. Zuo, X. Li, L. Qian, J. Ren and Z. Sun, "A reinforcement learning approach to autonomous decision making of intelligent vehicles on highways", *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 50, no. 10, pp. 3884-3897, Oct. 2020.
- [2] D. Zhao, H. Lam, H. Peng, S. Bao, D. J. LeBlanc, K. Nobukawa, et al., "Accelerated evaluation of automated vehicles safety in lane-change scenarios based on importance sampling techniques", *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 3, pp. 595-607, Mar. 2017.
- [3] H. Kurniawati, "Partially observable Markov decision processes and robotics," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 5, no. 1, pp. 253-277, May 2022.
- [4] B. H. Abed-alguni, "Action-selection method for reinforcement learning based on cuckoo search algorithm", *Arab. J. Sci. Eng.*, vol. 43, no. 12, pp. 6771-6785, 2018.
- [5] H. Hassen, S. Meherzi and Z. B. Jemaa, " $\epsilon$ -QLMR:  $\epsilon$ -greedy based Q-Learning algorithm for Multipath Routing in SDN networks," 2023 International Wireless Communications and Mobile Computing (IWCMC), Marrakesh, Morocco, 2023, pp. 234-239, doi: 10.1109/IWCMC58020.2023.10183270.
- [6] C. Amiez, J. P. Joseph, and E. Procyk, "Reward encoding in the monkey anterior cingulate cortex," *Cereb. Cortex*, vol. 16, pp. 1040-1055, Jul. 2006.
- [7] J. W. Brown and T. S. Braver, "Learned predictions of error likelihood in the anterior cingulate cortex," *Science*, vol. 307, no. 5712, pp. 1118-1121, Feb. 2005.

- [8] M. Medalla, J. P. Gilman, J. Y. Wang, and J. I. Luebke, "Strength and diversity of inhibitory signaling differentiates primate anterior cingulate from lateral prefrontal cortex," *Journal of Neuroscience*, vol. 37, no. 18, pp. 4717-4734, May 2017.
- [9] J. Sallet, R. Quilodran, M. Rothé, J. Vezoli, J. P. Joseph, and E. Procyk, "Expectations, gains, and losses in the anterior cingulate cortex," *Cogn. Affect. Behav. Neurosci.*, vol. 7, pp. 327-336, Dec. 2007.
- [10] M. M. Botvinick, J. D. Cohen, and C. S. Carter, "Conflict monitoring and anterior cingulate cortex: an update," *Trends in Cognitive Sciences*, vol. 8, no. 12, pp. 539-546, Dec. 2004.
- [11] M. F. Rushworth, N. Kolling, J. Sallet, and R. B. Mars, "Valuation and decision-making in frontal cortex: one or many serial or parallel systems?," *Current Opinion in Neurobiology*, vol. 22, no. 6, pp. 946-955, Dec. 2012.
- [12] R. Schubert, K. Schulze, and G. Wanielik, "Situation assessment for automatic lane-change maneuvers," *IEEE Transactions on Intelligent Transportation Systems*, vol. 11, no. 3, pp. 607-616, Sept. 2010.