Interpretable Autonomous Driving Model Based on Cognitive Reinforcement Learning

Yijia Li[1], Hao Qi[2], Fenghua Zhu[1], Yisheng Lv[1] and Peijun Ye[1]†

***Abstract*— With the rapid development of autonomous driving technology, the safety of driving systems has increasingly become the focus of attention. However, although many existing autonomous driving decision-making algorithms, such as deep reinforcement learning, demonstrate excellent performance, their decision-making processes lack interpretability and are opaque to users. To address this problem, this paper constructs an interpretable driving model from the perspective of human cognition, which can not only imitate human driving behavior through cognitive reinforcement learning methods, but also show better performance in driving experiments. In addition, the paper also proposes an analysis method for abnormal driving behavior, which provides a new idea for discovering potential unsafe behaviors during driving and exploring the possible impact of this behavior pattern on driving tasks.**

I. INTRODUCTION

In the field of intelligent transportation, with the advancement of technologies such as high-definition maps, high-performance sensors, and artificial intelligence, autonomous driving has made significant progress in achieving a safer and more convenient transportation experience. Particularly, data-driven deep learning technologies have greatly enhanced the perception and decision-making capabilities of autonomous driving systems. However, most current autonomous driving systems focus more on safety and stability than on providing interpretability [1], and the decision-making logic behind them is often difficult to understand. Interpretability is crucial for autonomous driving systems. Especially in emergency traffic situations, when the system makes interpretable decisions, it not only assists technicians in identifying and addressing potential safety issues within the system but also enhances users' trust in the technology [2].

The core task of autonomous driving is to convert environmental perception information into vehicle control signals, This includes various technological processes such as sensing information processing, planning, decision-making, and motion control. This paper focuses on the interpretability of driving models, thus the existing methods will be summarized based on interpretable methods and uninterpretable methods.

*This work is supported in part by National Natural Science Foundation of China under Grant 62076237, Grant T2192933, and in part by Youth Innovation Promotion Association of the Chinese Academy of Sciences under Grant 2021130.

[1]Yijia Li, Fenghua Zhu, Yisheng Lv and Peijun Ye are all with the State Key Laboratory of Multimodal Artificial Intelligence Systems, Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China. {liyijia2023, fenghua.zhu, yisheng.lv, peijun.ye}@ia.ac.cn

[2]Hao Qi is with School of Rail Transportation, Shandong Jiaotong University, Jinan 250357, China. 21121015@stu.sdjtu.edu.cn

† Corresponding Author.

For interpretable driving models, the more typical ones are rule-based models. The idea is to establish a behavioral rule library based on traffic regulations, perception information and other prior knowledge, and guide the behavior of the vehicle based on logical rules and different states of the vehicle. For example, Montemerlo et al. use a finite state machine to realize the transformation of different vehicle states [3]. Vanholme et al. calculate the optimal trajectory of the vehicle based on traffic rules, human rules and system rules [4]. Zhao et al. use an ontology-based knowledge base. Rule reasoning and making decisions in real time [5]. In cognitive psychology, Anderson et al. proposed a cognitive architecture called adaptive control of rational thinking (ACT-R), which can simulate various cognitive tasks such as human memory, problem solving, and learning [6]. Salvucci developed a driver behavior calculation model based on the ACT-R cognitive architecture, which simulates the switching of human visual attention from the near point to the origin during driving, thereby calculating stable lateral and longitudinal control amounts [7]. Deng et al. integrated the queuing network with ACT-R and proposed the QN-ACTR architecture, which can simulate driving tasks in a variety of traffic environments [8]. The above-mentioned rule-based and cognitive system-based driving models have strong interpretability and can accurately simulate human decision-making behavior in response to different traffic conditions. However, they require a large amount of prior knowledge to build the rule base, and their adaptability in dynamic and uncertain environments is relatively weak.

For driving models that lack interpretability, the two representative paradigms are data-driven and reward-oriented learning methods. Data-driven methods are represented by methods such as deep learning (DL) and imitation learning (IL) [9]. They need to learn high-level feature representations and uncovering dependencies between data from large datasets. This process enables the mapping from perceptual information to decision-making. Wang et al. used generative adversarial networks (CGAN) to simulate human driving behavior [10], and Cui et al. used CNN to achieve end-to-end trajectory prediction [11]. Reward-driven methods mainly refer to deep reinforcement learning (DRL), where agents learn through trial and error by interacting with the environment to maximize their cumulative returns. This type of method does not rely on a large amount of data, but requires full interaction with the environment and a reasonable design of action space, state space, and reward function. Huang et al. introduced expert knowledge into DRL for training, which improved sample efficiency and also reduced the workload