Behavioral Alignment and Verifiable Explainability in Autonomous Driving

Ashish Pandian

Electrical Engineering and Computer Sciences
University of California, Berkeley

Technical Report No. UCB/EECS-2025-79
http://www2.eecs.berkeley.edu/Pubs/TechRpts/2025/EECS-2025-79.html

May 15, 2025

Acknowledgement

I would like to express my sincere gratitude to Professor Bayen for welcoming me into his lab and providing me with meaningful research opportunities. Your guidance and insights have shaped both this work and my approach to research. Thanks also to Professor Jiao for serving as my second reader. To my labmates, who patiently answered my questions and shared their expertise, thank you for making the lab a place of both learning and camaraderie. I owe special thanks to my parents and brother, whose support has been constant throughout my academic journey. Finally, to my friends who celebrated small victories with me and provided necessary distractions when needed, your presence made this journey meaningful. This achievement is, in many ways, a shared one.

Behavioral Alignment and Verifiable Explainability in Autonomous Driving

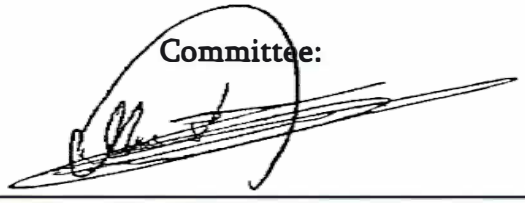
By Ashish Pandian

Research Project

Submitted to the Department of Electrical Engineering and Computer Sciences, University of California at Berkeley, in partial satisfaction of the requirements for the degree of **Master of Science, Plan II**.

Approval for the Report and Comprehensive Examination:

Committee:



Professor Alexandre M. Bayen
Research Advisor

May 13, 2025

(Date)

* * * * * * *

Professor Jiantao Jiao
Second Reader

May 12, 2025

(Date)

Behavioral Alignment and Verifiable Explainability in Autonomous Driving

by

Ashish Pandian

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Masters of Science

in

Electrical Engineering and Computer Science

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Alexandre Bayen, Chair
Professor Jiantao Jiao

Spring 2025

Behavioral Alignment and Verifiable Explainability in Autonomous Driving

Abstract

Behavioral Alignment and Verifiable Explainability in Autonomous Driving

by

Ashish Pandian

Masters of Science in Electrical Engineering and Computer Science

University of California, Berkeley

Professor Alexandre Bayen, Chair

The integration of AI systems into society represents a two-way road centered on human-AI alignment: AI systems must understand human intentions while humans must comprehend AI decision-making processes. Autonomous vehicles offer a compelling case study where this alignment is essential, as these systems must navigate complex social environments dominated by human expectations, implicit norms, and unpredictable behaviors. Despite remarkable technical advances in robotics and machine learning, widespread adoption of autonomous systems remains constrained. This thesis addresses this bidirectional challenge through two complementary research directions. First, we demonstrate that by learning from human demonstrations rather than engineering explicit rewards, autonomous systems can internalize the subtle social dynamics that govern human interaction. Second, by developing a framework for transparent reasoning, we enable humans to build appropriate trust in autonomous decisions through explanations that are both comprehensible and verifiably accurate. By addressing the reciprocal nature of human-AI alignment, this work contributes to the broader goal of creating AI systems that can be deployed not merely as optimization engines but as socially intelligent agents capable of harmonious integration with humans.

Contents

Acknowledgments

I would like to express my sincere gratitude to Professor Bayen for welcoming me into his lab and providing me with meaningful research opportunities. Your guidance and insights have shaped both this work and my approach to research. Thanks also to Professor Jiao for serving as my second reader; your thoughtful feedback substantially improved this thesis. To my labmates, who patiently answered my questions and shared their expertise, thank you for making the lab a place of both learning and camaraderie. I owe special thanks to my parents and brother, whose support has been constant throughout my academic journey. Finally, to my friends who celebrated small victories with me and provided necessary distractions when needed, your presence made this journey meaningful. This achievement is, in many ways, a shared one.

Chapter 1

Introduction: The Dual Challenge of Human-AI Alignment in Driving

1.1 Motivation

The field of robotics is about to enter a transformative new era. After decades of development primarily confined to controlled industrial and laboratory settings, autonomous systems are now poised for integration into real world environments. This shift from structured to unstructured domains marks not just a technical evolution, but a fundamental step in human–robot interaction. Among these innovations, autonomous vehicles (AVs) offer one of the most promising frontiers. They serve as intelligent agents capable of managing traffic, reducing accidents, and improving fuel efficiency. Yet realizing this vision requires more than technical competence. It demands the establishment of trustworthy interactions between AVs and human road users, including drivers, passengers, and pedestrians.

A central challenge lies in ensuring that autonomous systems operate not merely as optimization engines tuned for fuel savings or travel time, but as socially competent agents that understand, anticipate, and adapt to human behavior. This is particularly critical in mixed-autonomy settings, where autonomous vehicles must share the road with human drivers whose actions are governed by subtle behavioral norms, incomplete information, and occasionally irrational choices.

Every day, traffic congestion and stop-and-go waves disrupt commutes, contributing to increased emissions, wasted fuel, and significant economic loss. While AVs serve as mobile actuators in the flow of traffic with potential for traffic smoothing, deployments in real-world settings have highlighted a second dimension of concern: even when technically correct, autonomous vehicle actions can appear inexplicable or unsettling to nearby human drivers, prompting disengagement or manual override.

This thesis tackles these dual facets of the human-AI alignment problem in autonomous driving: (1) designing driving policies that are aligned with human social norms and expectations, and (2) ensuring that the decision-making processes of such policies are transparent

and comprehensible to human users.

1.2 The Two-Sided Problem of Human-AI Alignment

Achieving successful integration of AVs into our traffic systems requires addressing a fundamental problem often termed human-AI alignment. This concept encapsulates the two-sided challenge of ensuring mutual understanding and predictability between artificial agents and humans. This challenge encompasses both the AI's comprehension of human norms and the human's comprehension of the AI's processes.

AI Understanding Humans: Modeling Latent Behavioral Objectives

For AVs to operate smoothly and safely alongside human drivers, they must possess more than just the ability to bridge the simulation-to-reality gap; they require a nuanced understanding of inherent human driving behaviors, intentions, and expectations. Due to the complexity of the problem, reinforcement learning(RL) approaches, especially for real-world problems, are driven by engineered objectives commonly referred to as proxy rewards. While these objectives may optimize predefined metrics, they frequently produce policies that violate subtle social norms. In the domain of traffic smoothing, this means maintaining uncomfortably large headways that invite aggressive cut-ins or braking in ways that feel jarring to human drivers.

This lack of behavioral alignment stems from optimizing metrics that fail to capture the complexities of real-world human interaction and decision-making. Crafting explicit reward functions to encode these nuances is notoriously difficult; dense rewards require intricate reward engineering and may still miss subtle preferences, while sparse rewards often provide insufficient guidance for learning complex, long-horizon behaviors characteristic of human driving. Learning by demonstration is a natural way for agents to learn complex behaviors as a small set of demonstrations is often easy to obtain from a human expert. Demonstrations further alleviate the need for exploration, as they forego both the search problem and the exploration-exploitation trade-off by reducing the task to distribution matching.

Learning from demonstrations has shown tremendous success in robotics, enabling complex behaviors without exhaustive reward engineering. One form of learning from demonstrations, imitation learning, has shown remarkable efficacy by directly mimicking expert actions in similar states. Inverse Reinforcement Learning (IRL) builds on this by aiming to recover the underlying reward function that best explains observed expert behavior, thereby capturing the latent objectives driving human actions. This allows AVs to move beyond simple efficiency calculations towards socially intelligent agents grounded in inferred human preferences.

Humans Understanding AI: The Imperative for Verifiable Transparency

Conversely, even an optimally efficient and behaviorally considerate AV will struggle for acceptance if its decision-making processes are perceived as opaque or arbitrary. Complex AI systems, particularly those based on deep reinforcement learning, often function as "black boxes." Their operational logic, while potentially optimal according to learned objectives, can be inscrutable to human observers. When an AV takes an action that deviates from typical human behavior, even if beneficial for overall traffic flow, the lack of a clear, verifiable rationale can lead to confusion, mistrust, and eventual rejection by users.

Passengers might feel uncomfortable, other drivers might misinterpret the AV's intentions, and regulators may hesitate to approve widespread deployment without mechanisms for ensuring transparency and accountability. Therefore, enabling humans to understand the underlying rationale for an AV's decision processes is as critical as the effectiveness of the actions themselves. Explainability is not merely a desirable feature; it is a prerequisite for building trust and facilitating effective human-AI collaboration on the road. Furthermore, these explanations must strive for using logical correctness and grounding in the system's actual operational dynamics as verifiable accounts of the decision process.

1.3 A Synergistic Framework for Behavior and Explanation

To address the dual challenge of human-AI alignment in autonomous driving, this thesis proposes a synergistic framework that tackles both the AI's understanding of human behavior and the human's understanding of the AI's reasoning. We posit that true alignment requires progress on both fronts. Our approach comprises two complementary research thrusts: first, synthesizing AV control policies that intrinsically blend system efficiency with learned human behavioral objectives inferred via IRL; and second, developing a mechanism for generating transparent, verifiable, and human-understandable explanations for the resulting agent's actions, with a focus on ensuring logical soundness.

Synthesizing Human-Aligned Policies via Inverse Reinforcement Learning and Policy Composition

Addressing the need for AVs to understand and respect human norms, we leverage IRL to move beyond traditional control optimization and the difficulties of manual reward design. IRL seeks to infer the reward functions that rationalize observed actions, providing a pathway to capture successful and transferable definitions of tasks directly from human demonstrations.

Following advancements in adversarial learning, we utilize a discriminator-based IRL approach to learn a disentangled reward function implicitly capturing human driving prefer-

ences from noisy, real-world traffic data, such as that available from the Vandertest dataset. Extracting this reward function serves multiple purposes: it yields a stronger data-driven model of human driving behavior compared to traditional models, such as IDM, that may not capture implicit social norms, and it provides a component reward function that allows downstream traffic optimization tasks to explicitly account for human behavioral factors during deployment.

Rather than using this learned reward solely for imitation, we introduce a policy composition framework. This framework explicitly blends the learned human behavioral reward with an engineered energy efficient and traffic smoothing focused reward function. By training a reward function that balances different levels of human norms, we create a tunable framework that enables understanding behaviors across a spectrum between these two contrasting policies.

Enabling Verifiable Transparency via Grounded Explanations

Complementing the development of behaviorally nuanced policies, the second part addresses the need for transparency in RL policies through explanation. While Large Language Models (LLMs) offer powerful capabilities for generating fluent text, their application in safety-critical domains necessitates rigorous validation to mitigate risks such as hallucination or unfaithful reasoning.

Inspired by the concept of world models and the need for robust verification, our approach emphasizes mechanisms to ensure explanations are logically sound and reflect the true dynamics understood by the agent. A key aspect is the use of hypothetical scenarios and simulated rollouts; by tasking the explanation system with predicting and justifying behavior not just in the observed state but also under counterfactual conditions, we create a richer "playground" for probing and verifying its understanding.

Generating synthetic data or hypothetical variations allows the system to demonstrate its grasp of cause-and-effect and the policy's behavior across a wider range of situations than observed data alone might provide. This process of generating and evaluating behavior in hypothetical contexts serves as an important verification layer, ensuring the generated explanations are not merely superficial rationalizations, but are grounded in the agent's decision making process.

1.4 Overview

This thesis confronts the dual challenge of human-AI alignment in autonomous driving by developing and evaluating an integrated framework focused on both behavioral compatibility and verifiable operational transparency. Our primary contribution lies in a policy composition framework that uses IRL to extract a human behavioral reward model from real-world traffic data.

Complementing this, we address the critical need for transparency by providing humans with explanations focused on verifiable reasoning, ensuring the trustworthiness of natural language explanations for the decision-making of RL agents.

By addressing both sides of the alignment challenge, how AI understands human behavior and how humans understand AI reasoning, this work provides a more holistic pathway toward developing autonomous systems that are not only technically capable but also socially compatible and interpretable, making them suitable for safe and harmonious real world deployment.

Chapter 2

Human Behavioral Modeling via Inverse Reinforcement Learning

2.1 Introduction

Traffic congestion, particularly the formation of stop-and-go waves, remains a persistent challenge in modern transportation networks. These oscillatory phenomena result from *string instability*, where minor fluctuations in a vehicle's velocity or acceleration propagate downstream, amplifying into larger disturbances [30, 33]. The consequences include increased fuel consumption, elevated emissions, reduced throughput, and substantial economic costs.

The emergence of AVs presents new opportunities for mitigating such inefficiencies. Unlike infrastructure-based interventions, AVs can serve as mobile actuators embedded in the traffic stream [16]. However, most AV control policies are optimized using engineered reward functions that prioritize system-level objectives like energy efficiency or traffic stability. While effective in simulation, such policies often exhibit unnatural behaviors in mixed-autonomy settings, potentially leading to reduced acceptance and compliance from human drivers [25, 14].

A primary challenge lies in capturing the nuanced, often implicit, objectives that guide human driving decisions. Manually engineering reward functions within a standard RL framework to capture these subtleties proves remarkably difficult [1]. This challenge, often referred to as the reward specification problem, underlies many of the difficulties in developing AVs that can seamlessly integrate into mixed-autonomy traffic flows.

Imitation Learning (IL) attempts to address this by mimicking expert behavior through state-action mappings. However, IL fundamentally lacks transferability across scenarios and provides limited insight into underlying behavioral motivations [24]. Inverse Reinforcement Learning (IRL), by contrast, seeks to infer the latent reward functions that explain observed behavior, offering a more robust and interpretable solution [18, 42].

In this work, we leverage real-world human driving data to extract reward functions through IRL and integrate these IRL-derived human reward models into a dual-objective

policy optimization framework for traffic smoothing. By focusing on recovering the reward function, IRL aims to model the *why* behind human actions, offering a potentially more robust and interpretable foundation for developing behaviorally aligned AV policies. Our goal is to use the inferred reward structure to balance human-like behavior with system-level efficiency goals in AV control design.

We apply this approach to trajectories collected from the I-24 westbound highway during rush hour via the VanderTest project [40]. We train a discriminator network using Adversarial Inverse Reinforcement Learning (AIRL) to learn human driving behavior [6]. Our empirical study examines the trade-offs between energy optimality and human-likeness across varying levels of congestion and vehicle platoon configurations. The results show that our method retains 85–90% of the fuel efficiency gains achieved by fully optimized controllers while producing behavior that more closely aligns with that of human drivers.

2.2 Related Work

Mixed Autonomy Traffic and AV Control

Traffic congestion, especially the formation of stop-and-go waves due to string instability, has been extensively studied in transportation engineering [30, 33]. Work on string instability demonstrated how human driving behavior alone could generate traffic waves even in the absence of bottlenecks or lane changes [30]. These instabilities, originating from minor variations in human driving behavior and amplified through the traffic stream, significantly affect economic and environmental costs [2, 26].

Autonomous vehicles offer a promising avenue for addressing these issues by acting as distributed controllers within the traffic flow. Research has shown, both in simulation [29, 34] and increasingly in real-world experiments [14, 40], that even a small percentage of strategically controlled AVs can dampen traffic waves and improve overall efficiency. A closed-track experiment demonstrated that a single autonomous vehicle following an optimal control policy could stabilize a ring of 22 vehicles, effectively eliminating stop-and-go waves [29]. Building on this, researchers explored various AV penetration rates and control strategies in simulation, finding that as few as 5% of vehicles being autonomous could yield significant improvements in traffic flow [34].

The MegaVanderTest provides empirical evidence for these benefits through the deployment of 100 connected autonomous vehicles on the I-24, but also highlights significant adoption challenges [12]. These studies demonstrate that while the theoretical potential of AVs to improve traffic conditions is substantial, acheiving these benefits requires AV control strategies that can seamlessly integrate with human-driven vehicles.

Reinforcement Learning for AV Control

Reinforcement Learning provides a powerful paradigm for training AV control policies, enabling agents to learn complex strategies through trial-and-error interaction [31]. In the context of autonomous driving, RL has been successfully applied to various tasks, from lane-keeping and obstacle avoidance to urban traffic navigation scenarios [19].

A particularly relevant application of RL in traffic is the development of policies for mitigating "phantom traffic jams," where reinforcement learning has been used to train autonomous vehicle controllers that smooth traffic flow in simulation [39].

However, the effectiveness of RL hinges critically on the specification of the reward function. Misspecified rewards can lead to unintended or undesirable behaviors, a challenge often referred to as the alignment problem [1, 25]. This difficulty in manually crafting appropriate reward functions for complex, socially-situated tasks like driving motivates the exploration of learning objectives from data.

Imitation Learning

Imitation Learning offers a direct approach to learning from demonstrations. Behavior cloning [20] treats policy learning as a supervised learning problem, mapping observed states to expert actions. However, it suffers from the covariate shift problem (also known as the tightrope walking problem) where minor errors can lead the agent into states unrepresented in the training data, causing compounding failures [23]. DAgger [24] attempts to mitigate this by interactively querying the expert for labels on agent-visited states, but this requires an online expert, making it impractical for many real-world applications.

Generative Adversarial Imitation Learning (GAIL) [9] adopts an adversarial framework, training a policy to produce state-action distributions indistinguishable from expert demonstrations. While GAIL offers improved robustness over simple behavior cloning, a fundamental limitation persists across IL methods: they primarily learn what the expert did, not why they did it. This lack of insight into the underlying objectives limits interpretability and transferability to new scenarios or goals.

Inverse Reinforcement Learning

Inverse Reinforcement Learning directly addresses this limitation by aiming to recover the reward function that rationalizes the observed expert behavior [18]. Early IRL approaches faced challenges, notably the inherent ambiguity where multiple reward functions could explain the same optimal policy.

The Maximum Entropy Inverse Reinforcement Learning (MaxEnt IRL) framework proposed by Ziebart et al. [**ziebart2008maximumEntropy**] introduced a principled approach for resolving the ambiguity inherent in inverse reinforcement learning. It models the expert as acting stochastically, assigning probabilities to trajectories according to a Boltzmann distribution over cumulative rewards. This formulation selects the reward function that maximizes

the entropy of the trajectory distribution while matching observed feature expectations, thereby favoring explanations that make the fewest additional assumptions. Follow-up work by Ziebart [**ziebart2010modeling**] and others [**levine2018reinforcement**] established a formal connection between reward maximization and maximum likelihood estimation, framing IRL as a problem of probabilistic inference over trajectories.

A key insight in both forward and inverse RL is that certain transformations of the reward function preserve the optimal policy. Ng et al. formalized this through the concept of reward shaping [17], highlighting the non-uniqueness in the reward recovery problem. An IRL algorithm must account for this invariance, either by imposing additional structure on the reward function or by explicitly modeling the potential function component.

MaxEnt IRL often requires solving a forward RL problem within its optimization loop, which can be computationally demanding, especially in large or continuous state spaces or when dynamics are unknown. This limitation motivated the development of more scalable approaches.

Guided Cost Learning (GCL) [**finn2016guided**] introduces an importance sampling mechanism to estimate the partition function, which becomes intractable in high-dimensional continuous environments. Unlike traditional approaches that assume access to a known dynamics model, GCL jointly learns both the reward function and a sampling policy. The learned policy generates trajectories that are used to reweight the likelihood estimates during reward learning, allowing the algorithm to scale to settings where the transition dynamics are unknown or difficult to model.

The advent of Generative Adversarial Networks (GANs) [7] introduced a powerful adversarial training paradigm that has since influenced many areas of machine learning, including inverse reinforcement learning. In the GAN framework, a generator network produces synthetic data that aims to resemble samples from the true data distribution, while a discriminator network is trained to distinguish between real and generated data. The two networks are trained in tandem, with the generator improving its outputs to fool the discriminator. This adversarial setup provides a mechanism for learning complex data distributions without requiring explicit likelihoods.

Adversarial Inverse Reinforcement Learning (AIRL) [6] extends this adversarial approach to the IRL setting by replacing the generator with a policy and the real data with expert demonstrations. The discriminator is trained to distinguish expert transitions from those generated by the current policy. Crucially, AIRL structures the discriminator such that it implicitly recovers a reward function along with a dynamics-dependent shaping term. The learned reward function is disentangled from the environment's dynamics, making it more transferable across domains. By casting IRL as an adversarial game, AIRL enables learning reward functions that both explain expert behavior and generalize to new settings.

Compared to MaxEnt IRL, AIRL can be more scalable as it bypasses the need to repeatedly solve the full forward RL problem within the loop, instead leveraging the powerful optimization dynamics of adversarial training. This structure makes AIRL particularly well-suited for learning rewards from complex, high-dimensional data like real-world driving trajectories, motivating its selection for modeling human driving behavior in this work.

2.3 Methodology

Problem Formulation

We formulate the problem of mitigating stop-and-go waves as a reward learning and policy optimization challenge in a mixed-autonomy setting. Stop-and-go waves emerge due to the inherent limitations of human reaction times and decision-making processes in dense traffic. Autonomous vehicles have the potential to act as mobile actuators to smooth traffic flows and prevent these phantom jams.

To demonstrate these effects, we leverage a dataset of human driving trajectories collected from the VanderTest project [40] to infer a human reward function. This dataset contains real-world driving behavior on the I-24 highway under varying traffic conditions, including congestion, providing a rich source for modeling typical human responses.

Our objective is to learn a reward function that captures the implicit objectives guiding human driving behavior in various traffic conditions and develop a policy optimization framework that balances these learned human preferences with explicit energy efficiency goals.

Energy-Optimal Reinforcement Learning Objective

To establish a baseline for vehicle energy optimization, we utilize the control reward function developed by [14]. This function penalizes high fuel consumption and rewards smooth driving that maintains safe distances:

$$r_{energy}(s_t, a_t, s_{t+1}) = 1 - c_0 E_t - c_1 a_t^2 - c_2 P_t \tag{2.1}$$

where E_t is instantaneous fuel consumption, a_t is acceleration, P_t is a headway penalty, and c_0, c_1, c_2 are weighting coefficients set to $c_0 = 1.0\frac{1}{Gal}$, $c_1 = 0.002\frac{s^2}{m}$, and $c_2 = 2$.

Inverse Reinforcement Learning via AIRL

We employ IRL to infer the human reward function from real world dataset. IRL is chosen over Imitation Learning because our goal is to understand the underlying reward objective rather than just mimicking actions. This provides a basis for generalization, transferability, and interpretability.

We utilize Adversarial Inverse Reinforcement Learning (AIRL) [6] to extract the reward function. AIRL formulates the inverse reinforcement learning problem as a game between a generator policy π_G and a discriminator $D_{\theta,\phi}$. The discriminator aims to distinguish between state–action–next state tuples sampled from expert demonstrations D and those produced by the current policy π_G. Its structure supports reward disentanglement:

$$D_{\theta,\phi}(s, a, s') = \frac{\exp(f_{\theta,\phi}(s, a, s'))}{\exp(f_{\theta,\phi}(s, a, s')) + \pi_G(a|s)} \tag{2.2}$$

The function $f_{\theta,\phi}(s, a, s')$ is decomposed as:

$$f_{\theta,\phi}(s, a, s') = g_\theta(s, a) + \gamma h_\phi(s') - h_\phi(s) \tag{2.3}$$

In this formulation, $g_\theta(s, a)$ is the learned reward function we aim to recover, parameterized by θ, and $h_\phi(s)$ is a learned potential function over states, parameterized by ϕ. The potential function serves as a reward shaping term. This decomposition ensures that the optimal policy under f remains optimal under g_θ, enabling theoretical transferability of the recovered reward across environments with different dynamics.

The discriminator $D_{\theta,\phi}$ is trained using a binary cross-entropy loss to classify expert transitions as 'real' and policy-generated transitions as 'fake':

$$\mathcal{L}_D = -\mathbb{E}_{(s,a,s')\sim D}[\log D_{\theta,\phi}(s, a, s')] - \mathbb{E}_{(s,a,s')\sim \pi_G}[\log(1 - D_{\theta,\phi}(s, a, s'))] \tag{2.4}$$

The generator policy π_G is trained using Proximal Policy Optimization (PPO) [27] to maximize the reward signal derived from the discriminator. The overall training procedure is outlined in Algorithm 1. Upon convergence, g_θ represents the inferred human reward function.

Algorithm 1 PPO-IRL

1: Collect expert trajectories $D = \{\tau_i^E\}$ from VanderTest dataset.
2: Initialize policy π_G (parameterized by ω), value function V_ξ.
3: Initialize reward approximator g_θ and shaping term h_ϕ.
4: **for** iteration $k = 0, 1, 2, \ldots, K$ **do**
5: Collect trajectories $G_k = \{\tau_i^G\}$ by executing policy π_G in the environment.
6: Sample batches of transitions from D and G_k.
7: Update discriminator parameters θ, ϕ by minimizing $\mathcal{L}_D$.
8: Compute rewards for generated trajectories G_k using $r_{\theta,\phi}(s, a, s')$.
9: Update policy π_G and value function V_ξ using PPO with rewards $r_{\theta,\phi}$.
10: **end for**
11: **return** learned reward function g_θ

Policy Mixture Framework

After obtaining both the learned human reward function g_θ and the engineered energy reward function r_{energy}, we introduce a Policy Mixture framework to balance the competing objectives of human-like behavior and energy efficiency. This framework combines the two reward functions using a scalar mixing coefficient $\alpha \in [0, 1]$:

$$r_{combined}(s, a, s') = (1 - \alpha) \cdot g_\theta(s, a) + \alpha \cdot r_{energy}(s, a, s') \tag{2.5}$$

The combined reward function $r_{combined}$ serves as the objective for training the final AV control policy π_{final} using PPO. The parameter α controls the interpolation between

a policy that fully optimizes the learned human preferences at $\alpha = 0$ and a policy that fully optimizes the engineered energy objectives at $\alpha = 1$. By varying α, we can explore the trade-off space between behavioral human-likeness and system efficiency. The training process for this mixture policy is outlined in Algorithm 2.

Algorithm 2 Policy Mixture Training

Load learned human reward function g_θ from Algorithm 1.
Define engineered energy reward function r_{energy}.
Choose a mixture coefficient $\alpha \in [0, 1]$.
Initialize final policy π_{final} (parameterized by ω').
for iteration $k = 0, 1, 2, \ldots, K'$ **do**
    Collect trajectories by executing policy π_{final}.
    Compute combined rewards using $r_{combined} = (1 - \alpha)g_\theta + \alpha r_{energy}$.
    Update policy π_{final} and value function using PPO with rewards $r_{combined}$.
end for
return final policy π_{final} (tuned by α)

Implementation Details

The reward functions (g_θ, h_ϕ) and the policy/value functions are implemented as multi-layer perceptrons (MLPs) with 64 hidden units per layer. Each MLP consists of fully connected layers with ReLU activations.

Our observation space consists of the ego vehicle speed, headway distance to the leader vehicle, and the leader vehicle speed. The action represents a continuous space of the acceleration of the car.

Both AIRL training and the Policy Mixture training use the PPO algorithm with the following hyperparameters: discount factor $\gamma = 0.99$, GAE parameter $\lambda = 0.97$, policy learning rate: 3×10^{-4}, and 250,000 training iterations.

2.4 Experimental Results

Experimental Setup

We evaluate our Policy Mixture framework using real-world driving trajectories from the I-24 highway westbound [40], analyzing its performance in three different platoon configurations (Figure 2.1) that capture various mixed-autonomy driving scenarios:

1. Config 1: AV alone (baseline case)
2. Config 2: AV-Human*4 (AV leading four human vehicles)

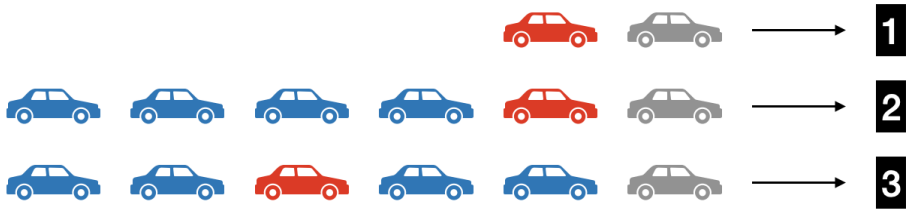


Figure 2.1: Platoon Configurations. The direction of the arrow represents the direction in which the vehicles are driving on a single lane. AV is depicted in red. The human vehicle is depicted in blue. The leader vehicle trajectory is depicted in gray.

3. Config 3: Human*2-AV-Human*2 (AV in middle of human vehicles)

Our primary goal is to demonstrate how the balance between energy optimization and human-like behavior (controlled by α) impacts system performance. We use the Intelligent Driver Model (IDM) [32] to simulate the behavior of human drivers in the platoon. The IDM provides a model of human car-following behavior, including reactions to changes in the lead vehicle's speed and spacing.

To understand the defining behaviors of these two contrasting policies, we utilize three metrics. First, system energy consumption, defined as the total energy expenditure across all five vehicles in the platoon, measured in kJ/m. Second, mean headway, which captures the average distance between successive vehicles in the platoon, measured in meters. Third, fuel efficiency, expressed in miles per gallon(MPG) and computed using the fuel consumption model.

Analysis on Human Reward Function

Following the convergence of the PPO-IRL algorithm, we notice some key characteristics of human driving behavior on the freeways. While the energy-optimal reward function tends to encourage larger headways to provide buffer space for smoother deceleration, the human-derived reward function shows a preference for more moderate following distances.

Balancing Energy Efficiency and Human-Likeness

Table 2.1 presents the impact of α on both energy consumption (kJ/m) and mean headway (m) for Configuration 2 (AV-Human*4). As expected, a purely energy-optimal policy of $\alpha = 1$ achieves the lowest energy consumption but maintains unrealistically large headways of 89.14 meters. Conversely, a purely human-like policy of $\alpha = 0$ results in smaller, more natural headways of 67.8 meters but at the cost of an average 1.324 kJ/m.

Table 2.1: Policy Mixture Performance on Platoon Configuration 2

Alpha Level	**Mean Headway (m)**	**Energy Consumption (kJ/m)**
Human-like ($\alpha = 0$)	67.8	1.324
$\alpha = 0.2$	72.2	1.298
$\alpha = 0.5$	74.6	1.276
$\alpha = 0.8$	77.3	1.247
Energy-Optimal ($\alpha = 1$)	89.14	1.225

The middle ground, particularly at $\alpha = 0.5$, represents an attractive compromise, achieving headways of 74.6 meters with only 4.2% higher energy consumption compared to the purely energy-optimal policy.

Impact of Platoon Configuration on Efficiency

Table 2.2 shows how system-wide fuel efficiency (MPG) achieved by Policy Mixture ($\alpha = 0.5$) varies significantly across different platoon configurations. This highlights a critical factor for real-world AV deployment: strategic placement within a mixed-autonomy traffic stream is essential for effective congestion mitigation.

Table 2.2: Platoon Configurations for $\alpha = 0.5$

Platoon Configuration	**MPG**
Config 1: AV-Human*4	35.67
Config 2: Human*2-AV-Human*2	32.14
Config 3: AV alone	36.12

While a solo AV (Config 3, 36.12 MPG) achieves the highest efficiency, it represents an unrealistic scenario. More importantly, we observe a significant efficiency difference between Config 1 (32.14 MPG) and Config 2 (35.67 MPG). When the AV is surrounded by human drivers as in Config 2, its ability to influence overall traffic flow is limited. However, an AV leading the platoon, in Config 1, can directly improve efficiency by guiding the vehicles behind it. This finding show that early AV deployment might be most effective when focused on lead vehicles in convoys.

2.5 Discussion

Interpretability of Learned Rewards

One of the key advantages of our IRL approach is the interpretability it offers. By recovering a reward function rather than directly learning a policy, we gain insights into the underlying

objectives that shape human driving behavior. Analysis of the learned reward function reveals patterns of how human drivers appear to balance competing objectives of maintaining target speeds and safe following distances. These along with future insights could inform not only AV control design but also traffic modeling and infrastructure planning more broadly. The learned rewards capture implicit social norms and expectations that are difficult to specify manually but critical for developing AVs that integrate naturally into human-dominated traffic flows.

Limitations and Future Work

Despite promising results, limitations in this work should be acknowledged. First, our approach relies on the assumption that human driving behavior can be effectively modeled as a reward-maximizing process, potentially overlooking factors that influence human driving. Second, our experiments focus on a simple car following scenario, whereas real world traffic involves complex multilane interactions. Future work should address these limitations by incorporating more sophisticated models of human cognition and expanding the application to richer traffic environments.

Chapter 3

CLEAR: Verifiable Language Model Explanations for Traffic Smoothing

3.1 Introduction

Reinforcement learning (RL) policies have demonstrated strong performance in traffic flow optimization, including reducing stop-and-go waves and fuel consumption in mixed-autonomy settings [35, 36]. Field deployments on Interstate 24 in Nashville showed that even limited adoption of RL-controlled autonomous vehicles (AVs) can significantly improve flow and energy efficiency [12]. However, the lack of interpretability in deep RL policies poses a barrier to practical deployment, particularly in scenarios where human operators must trust and supervise these systems.

During the MegaVanderTest deployment, one of the largest traffic smoothing field experiments to date, vehicle operators frequently disengaged the RL controllers despite having received training and preparation. Engagement rates were as low as 38% on the first day of testing [12], primarily because operators couldn't understand why the vehicle was behaving in specific situations. This lack of transparency undermines user confidence and raises legitimate safety concerns in critical operational environments.

In this chapter, we introduce CLEAR (Contextual Language Explanations for Actions from RL), a framework designed to generate verifiable natural language explanations for decisions made by RL traffic controllers. CLEAR emphasizes verifiability through multiple validation mechanisms, ensuring that explanations accurately reflect the underlying policy's decision-making process.

The Verifiability Challenge in Explanation Generation

Traditional explanation methods for black-box models often fall short in two critical dimensions: faithfulness to the model's actual decision process and logical coherence in the explanations provided. Large Language Models (LLMs) show promise in generating natural language explanations, but face significant challenges when applied to safety-critical domains

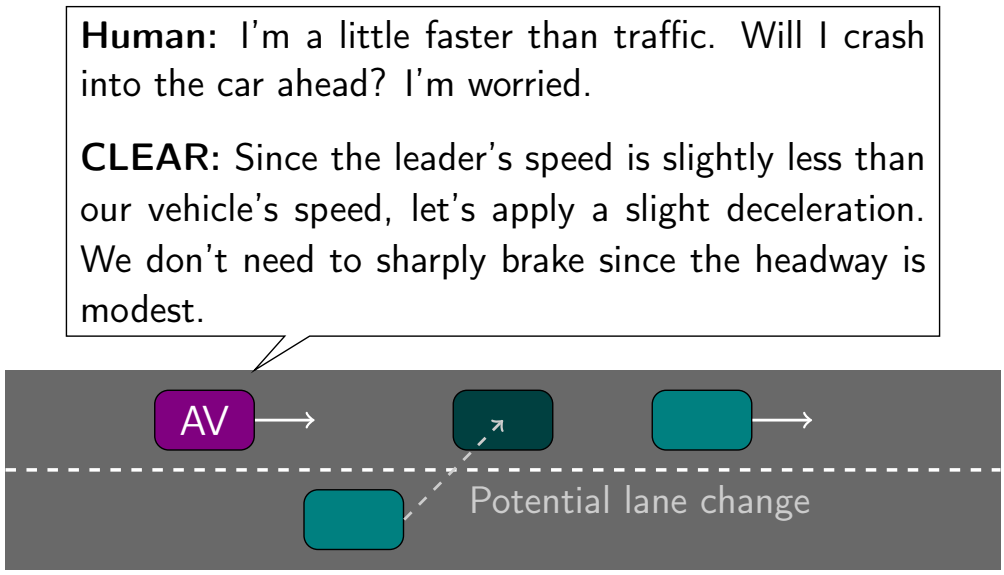


Figure 3.1: During the MegaVanderTest, operators of RL-controlled AVs reported discomfort caused by unusual space gaps compared to typical driving patterns. Several also raised concerns about potential dangers such as cut-ins, unpredictable lead vehicle behavior, and varying driving conditions [12]. CLEAR addresses these issues by providing real-time explanations for AV behavior, enhancing transparency and user trust.

like autonomous driving. Language models can generate explanations that sound plausible but contain invented reasoning or factual inaccuracies not grounded in the model's actual computation [5].

This hallucination risk is particularly problematic in autonomous systems where trust depends on accurate representations of the underlying decision process. Even when factually accurate, explanations may contain flawed reasoning chains or contradictory statements that undermine their utility for operators needing to make split-second trust decisions [41]. Moreover, static explanation methods often fail to account for the environmental context in which decisions are made, particularly in dynamic settings like traffic where conditions rapidly evolve [8].

In autonomous driving contexts, misaligned explanations can undermine operator trust or, worse, prompt inappropriate interventions during safety-critical situations. The verifiability gap in existing explanation methods remains a barrier to the practical deployment of RL–based traffic controllers.

3.2 Related Work

Explainability in Autonomous Driving and Reinforcement Learning

Explainability has been widely studied in machine learning, with model-agnostic methods such as LIME [22] and SHAP [15] providing post-hoc feature attributions for complex black-box models. In reinforcement learning (RL), explanation remains more challenging due to the sequential and high-dimensional nature of decision-making. Surveys on explainable RL [21] categorize methods based on whether they aim to explain individual components, such as observations, actions, or rewards, or whether they pursue inherently interpretable policy representations.

Language Models for Contextual Reasoning and Explanation

Large language models (LLMs) have emerged as powerful tools for generating structured explanations and supporting reasoning across diverse tasks without additional training. One prominent approach is in-context learning (ICL) [3], where the model adapts to new tasks using only a few examples embedded in the prompt.

Chain-of-thought (CoT) prompting [38] further improves reasoning quality by encouraging the model to generate intermediate steps before producing a final answer. This structured output aligns well with how humans decompose complex problems, and it improves transparency.

To support factual grounding, retrieval-augmented generation (RAG) [13] combines the LLM's generative ability with dynamic access to external information. By retrieving task-relevant documents at inference time, RAG enables models to operate on specialized domains without relying on static pretraining, improving adaptability and accuracy.

Despite their strengths, LLMs are prone to hallucinations: outputs that are syntactically fluent but factually incorrect or logically flawed [11]. To reduce hallucination risk, researchers have developed verification strategies that operate at inference time. One approach is self-consistency [37], which samples multiple completions from the model and selects the majority-voted response across samples, improving stability and reducing spurious outputs. Another strategy leverages LLMs as judges, where a secondary model evaluates and ranks candidate responses based on criteria such as logical soundness, factual correctness, and alignment with context [10].

Verifier-Based Feedback in Language Model Training

Verifier-based methods incorporate feedback from humans or learned reward models to guide generation, while verifier-free methods such as supervised fine-tuning rely solely on labeled examples. One line of work examines the limitations of scaling test-time compute without verification, showing that reinforcement learning (RL) offers more reliable alignment than prompting-based strategies without feedback mechanisms [28]. Another study com-

pares supervised fine-tuning (SFT) and RL post-training, demonstrating that SFT tends to memorize training data while RL generalizes better to unseen instructions and contexts [4]. These works highlight the importance of incorporating feedback into post-training to ensure robust and goal-aligned model behavior.

3.3 CLEAR: A Framework for Verifiable Explanations

CLEAR addresses the verifiability challenge through a two-layer architecture designed to generate, validate, and refine explanations for RL traffic controllers. The framework comprises a Generation Layer responsible for producing initial explanations and a Correctional Layer that systematically verifies and improves these explanations across multiple dimensions.

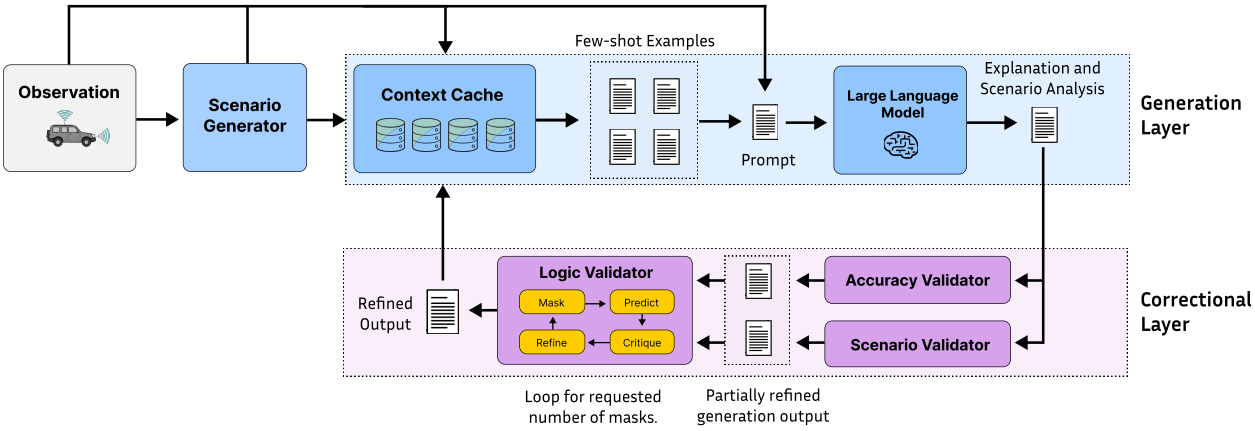


Figure 3.2: **Overview of CLEAR (Contextual Language Explanations for Actions from RL).** CLEAR consists of two components: the Generation Layer produces language explanations using a context cache of recent driving data, while the Correctional Layer refines each output for clarity and accuracy. As the AV operates, new data is continuously added to the context cache, allowing explanations to become more tailored to specific driving scenarios over time. The multi-level validation pipeline ensures explanations are factually accurate, physically plausible, and logically coherent.

Architecture Overview

CLEAR processes new vehicle observations through a structured pipeline that ensures explanation verifiability. As illustrated in Figure 3.2, the Generation Layer creates initial explanations by retrieving relevant examples from a context cache and utilizing a scenario generator to construct hypothetical future states. This is followed by the Correctional Layer, which applies a sequence of validators to ensure explanations are accurate, logically coherent,

and aligned with the RL controller's actual decision process. This modular design allows for systematic verification at multiple stages of the explanation process, creating a pipeline where each component addresses a specific aspect of explanation verifiability.

Generation Layer: Contextual Understanding

The Generation Layer addresses the need for contextually aware explanations by maintaining a diverse memory of past driving experiences and generating analyses of potential future scenarios. This layer consists of two primary components: the Context Cache and the Scenario Generator.

The context cache maintains an expanding repository of experiences as the AV operates, accumulating tuples of (observation, scenario, refined explanation, feedback) with each new driving situation. This memory enhances performance over time by retrieving relevant prior experiences. To ensure diversity and manage memory dimensions, an eviction module eliminates redundant entries based on action similarity, preserving a comprehensive range of wide-spanning examples while adapting to deployment constraints.

To examine the LLM's reasoning about potential future events, the Scenario Generator constructs "what-if" perturbations originating from the current observed state (ego speed, leader speed, space gap). It samples from a collection of scenario archetypes including lead vehicle braking, accelerating, maintaining speed, or a sudden cut-in maneuver. The generator synthesizes these elements into a natural language description representing the hypothetical scenario, which is incorporated into the LLM prompt. To subsequently validate the LLM's analysis, we employ a physics simulation for executing rollouts based on the scenario parameters. This simulation progresses in discrete increments, determining the leader vehicle's state according to the scenario, while simulating the ego vehicle's acceleration at each step by executing the original RL policy on the preceding state. This allows us to generate a trajectory used to verify the LLM's predictions against a physically plausible outcome.

The integration of historical context and forward-looking scenario analysis enables CLEAR to generate explanations that account for both past patterns and potential future states, addressing a key limitation of static explanation methods. By grounding explanations in both observed and hypothetical contexts, the Generation Layer establishes a robust foundation for the subsequent verification process.

Correctional Layer

Since human-in-the-loop evaluations and annotations are expensive and challenging to scale in practical driving environments, we introduce a set of specialized validators that facilitate automated, high-fidelity refinement. These validators critique and enhance specific aspects of each generated response before it is stored in the context cache.

The Accuracy Validator facilitates self-correction by providing the Language Model with access to the base RL controller's forward pass as a callable function. With this access, the

model can compare its predicted action to the controller's actual output. If a discrepancy is detected, the validator instructs the model to revise its explanation to better reflect the authentic behavior of the controller, ensuring the final response remains aligned with the true decision even if the initial generation is inaccurate.

The Scenario Validator verifies the accuracy of the generated analysis regarding hypothetical future scenarios. These scenarios, injected into the prompt, assume the RL controller acts on-policy under modified environmental conditions. To validate predictions about how the state would evolve in such settings, the validator is equipped with forward simulation tools, which is a learned model that approximates the AV's dynamics. The Language Model utilizes this model as an oracle to simulate the effects of actions over time and identify inconsistencies in the predicted outcomes. A final rule-based verifier provides an additional correctness check, ensuring that the revised explanation aligns with the physical plausibility and policy behavior under the simulated conditions.

The Logic Validator focuses on ensuring the reasoning process within explanations is logically sound, addressing a key limitation of previous explanation methods. The validator operates in three steps: first, it categorizes statements as either observational (factual) or inferential (reasoning); second, it tests logical coherence by masking key reasoning components and evaluating the model's ability to reconstruct them; and third, it performs a logical flow analysis to ensure that conclusions follow from premises through valid inferential steps. Logical flaws are identified through reconstruction accuracy, which can be described as the average semantic similarity between original statements and their reconstructions after masking the surrounding context. Low scores in this reconstruction accuracy indicate potential logical inconsistencies that require correction. This validation step is crucial for building operator trust, as it ensures that explanations not only align with the controller's actions but also follow logically consistent reasoning patterns.

3.4 Experimental Validation

To evaluate CLEAR's effectiveness in generating verifiable explanations, we conducted experiments using real-world trajectory data collected during the VanderTest deployment on Interstate 24. This data reflects realistic mixed-autonomy traffic scenarios, providing a comprehensive testbed for assessing explanation quality across diverse driving conditions.

Experimental Setup

Our experiments focused on two primary tasks designed to assess CLEAR's performance across different aspects of explanation generation. The first task was state to action mapping with rationale generation, which required the system to predict the RL controller's actions and provide explanations for observed traffic states. We evaluated both the accuracy of action prediction and the quality of generated explanations. The second task was hypothetical future state prediction, which focused on analyzing controller behavior under synthetically

perturbed scenarios to assess generalization capability. We evaluated the system's ability to predict controller responses to novel situations not present in the observed data.

We implemented CLEAR using Gemini Flash 2.0 as the base language model and compared its performance against four baselines to isolate the impact of verification mechanisms on explanation quality. These baselines were Zero-shot Gemini, which involved standard prompting without examples or verification; Few-shot Gemini, which used prompting with representative examples but no verification; CLEAR without validators, an ablated version of our framework without explicit verification components; and Supervised Fine-tuning (SFT) with LLaMA 3.2 8B trained on 770 expert-annotated examples. This comparative framework allowed us to systematically evaluate the contribution of each component to overall explanation quality. All models were evaluated on a held-out test set comprising 120 real-world traffic scenarios.

Evaluation Metrics

We employed multiple complementary metrics to assess different aspects of explanation quality. These included Action Prediction Accuracy, measured by the Mean Absolute Error (MAE) between predicted and actual controller actions; Explanation Quality, assessed by the cosine similarity between generated explanations and expert-annotated ground-truth rationales; and Scenario Prediction Error, determined by the L1 distance between predicted and simulated states in hypothetical scenarios. For the expert-annotated ground-truth rationales, three domain experts independently provided explanations for each scenario in the test set, and we used the consensus explanation (average embedding) as the reference point for evaluation.

Results and Analysis

Action Prediction Accuracy

CLEAR demonstrated superior performance in predicting the RL controller's actions, achieving a mean absolute error (MAE) that significantly outperformed all baselines, as shown in the top panel of Figure 3.3. The CLEAR variant without validators performed better than Zero-Shot and Few-Shot Gemini, and LLaMA with SFT also showed its respective performance. This substantial performance gap demonstrates the effectiveness of CLEAR's verification approach in ensuring explanations accurately reflect the controller's decision process. Notably, the SFT approach, while achieving competitive action prediction accuracy, exhibited significant deficiencies in explanation quality, often generating generic justifications that failed to capture the nuanced reasoning behind the controller's decisions. This highlights a fundamental limitation of supervised approaches in their ability to articulate the complex decision-making logic, even when they can reasonably approximate the resulting actions.

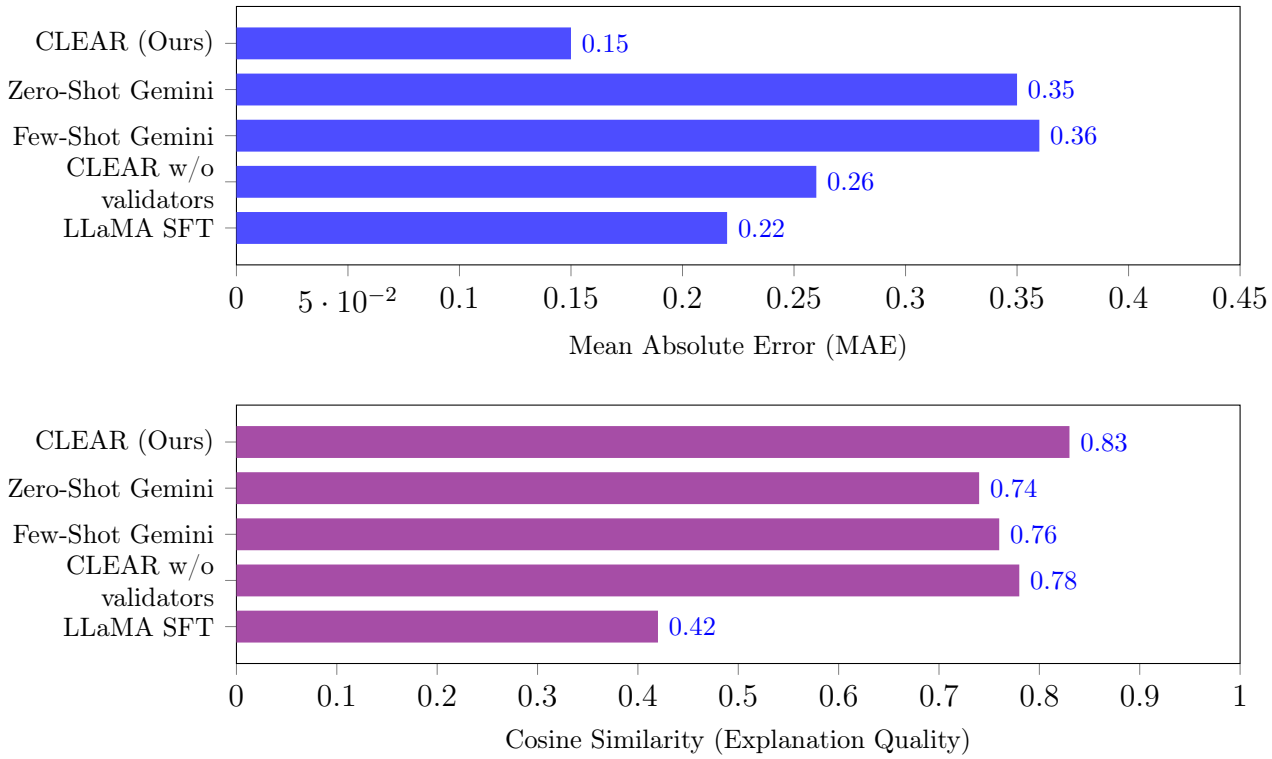


Figure 3.3: **Top:** Mean Absolute Error (MAE) of each method compared to the ground truth controller actions (lower is better). **Bottom:** Cosine similarity between generated explanations and expert-annotated ground truth rationales, reflecting explanation quality (higher is better).

Explanation Quality

To assess the semantic quality of generated explanations, we computed cosine similarity between generated explanations and expert-annotated ground-truth rationales. CLEAR achieved the highest similarity score, as shown in the bottom panel of Figure 3.3. This compared favorably to Few-Shot Gemini, Zero-Shot Gemini, CLEAR without validators, and LLaMA with SFT. While most methods correctly identified key observational features, CLEAR's explanations demonstrated superior inferential reasoning, particularly in cases requiring multi-step logical chains. The SFT approach notably struggled with explanation quality despite reasonable action prediction, highlighting the limitations of supervised learning for generating contextually appropriate explanations.

Qualitative analysis revealed that CLEAR's explanations included more specific references to relevant traffic features and demonstrated stronger causal reasoning about controller behavior. For example, in a case involving a decelerating leader vehicle, CLEAR correctly identified not only the deceleration itself but also its implications for future safety,

while baseline methods often provided generic explanations about maintaining safe distance without capturing the specific dynamics of the situation.

Performance Under Varied Scenarios

CLEAR demonstrated robust performance across diverse traffic scenarios, including challenging cases like emergency braking and cut-in maneuvers. As shown in Figure 3.4, in simulated emergency braking scenarios, CLEAR achieved a lower error compared to Zero-Shot Gemini, representing a substantial improvement in prediction accuracy under extreme conditions. Similarly, for cut-in maneuvers, CLEAR reduced error compared to the Zero-Shot baseline, demonstrating strong generalization to complex multi-vehicle interactions.

This performance across varied scenarios demonstrates the effectiveness of CLEAR's scenario-based verification approach in developing a robust, generalizable understanding of the controller's behavior across diverse traffic conditions. The framework's ability to maintain accuracy in challenging scenarios is particularly important for building operator trust in real-world deployments, where unusual traffic conditions can arise unexpectedly. Furthermore, as depicted in Figure 3.5, CLEAR maintains lower error rates over longer prediction horizons compared to baselines, showcasing its ability to make stable long-term forecasts in hypothetical scenarios.

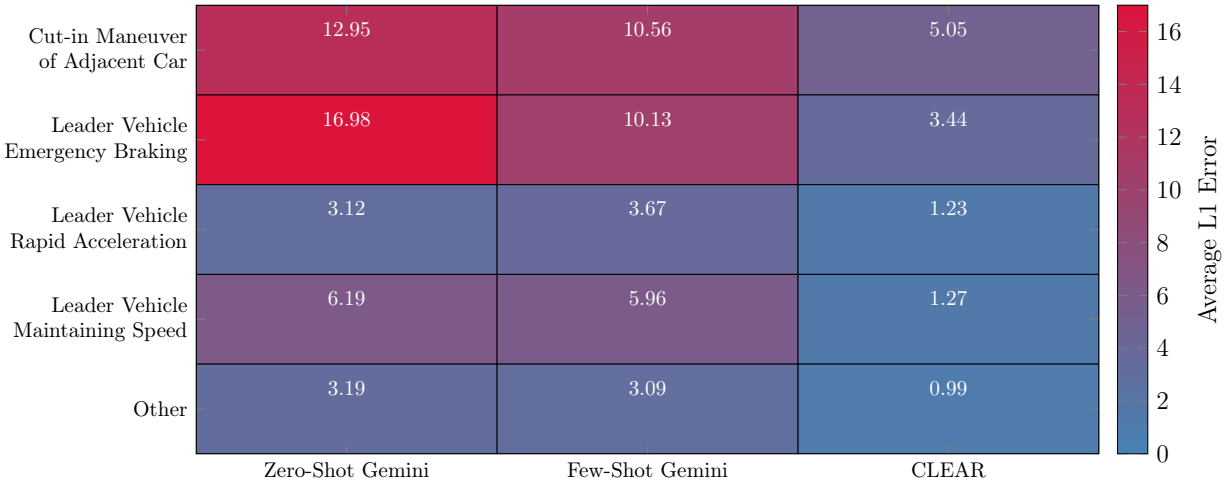


Figure 3.4: **Error by Archetype of Hypothetical Scenarios.** Average L1 error of predicted state by category of imagined scenario, across methods. Lower values (blue) are better. Prediction time horizon: 5 seconds.

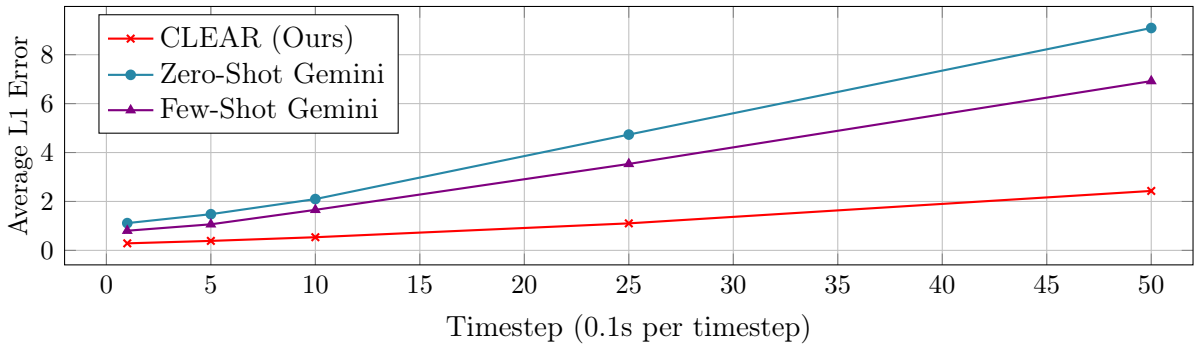


Figure 3.5: **Environment prediction accuracy with varying forecast horizons.** This shows the L1 error evolution across hypothetical scenario prediction horizons, demonstrating error propagation trends for all methods.

3.5 Discussion

Our experiments highlight limitations in standard approaches to generating explanations for RL controllers. Zero-Shot and Few-Shot prompting, while capable of generating plausible-sounding explanations, struggle with accurately capturing the controller's decision process, particularly in complex or unusual traffic scenarios. These approaches tend to rely on general knowledge about driving rather than specific understanding of the controller's policy, leading to explanations that may sound reasonable but fail to align with the controller's actual behavior.

Supervised Fine-Tuning (SFT), despite being trained directly on explanation examples, showed a tendency toward degenerate behavior, often generating generic explanations with poor logical coherence. While the model performs well on predicting the appropriate actions themselves, our analysis revealed significant flaws in its logical explanations. The SFT model struggles to provide coherent justifications that accurately reflect the controller's decision-making process, despite correctly mimicking the actions. Additionally, SFT approaches struggle with generalization to novel scenarios not represented in the training data, limiting their practical utility in real-world deployments where traffic conditions can vary widely.

These limitations underscore the importance of CLEAR's verification-focused approach, which enables more robust explanation generation without requiring extensive labeled training data. By directly integrating with the controller's decision process and incorporating multiple verification mechanisms, CLEAR overcomes the limitations of both prompting-based and supervised approaches.

3.6 Conclusion

This chapter presented CLEAR, a framework for generating verifiable natural language explanations for RL-based traffic controllers. By integrating multi-dimensional verification mechanisms, including accuracy validation, scenario-based testing, and logical coherence checking, CLEAR addresses key challenges in explanation generation for safety-critical autonomous systems.

Our experimental results, based on real-world data from the VanderTest deployment, demonstrate CLEAR's effectiveness in generating high-quality explanations that accurately reflect the controller's decision process, reason coherently about traffic dynamics, and generalize robustly across diverse scenarios. These capabilities represent a significant advancement over standard explanation approaches and address a critical need for transparency in autonomous vehicle deployment.

The verification-focused approach developed in CLEAR has implications beyond traffic control, potentially serving as a template for explanation generation in other safety-critical domains where algorithmic transparency and human trust are paramount. By prioritizing verifiability through multiple validation mechanisms, CLEAR represents an important step toward building autonomous systems that are not only effective but also transparent and trustworthy.

Chapter 4

Conclusion

4.1 Summary of Contributions

This thesis has addressed the bidirectional challenge of human-AI alignment in autonomous driving through an integrated framework that advances both behavioral compatibility and verifiable transparency. By tackling both dimensions, we have demonstrated that autonomous systems can be designed for harmonious integration into real world environments.

In the first part of this work, we leveraged human driving demonstrations to infer latent behavioral objectives and integrate them into an autonomous driving policy. Our IRL-based approach recovers a reward function that captures nuanced human driving preferences, such as comfortable headways and smooth acceleration, that are difficult to hand-engineer. Building on this learned reward, we introduce a *policy mixture* mechanism that blends human-like behavior with classical efficiency optimization. This approach yields driving policies that respect human norms while still reducing stop-and-go waves and improving fuel economy. Empirically, the mixed policy was able to maintain realistic headways (tens of meters) with only a few percent increase in energy consumption compared to a purely fuel-optimal policy. This balance illustrates that our method can align AV behavior with human expectations without sacrificing the benefits of traffic smoothing. Notably, in simulation experiments the behaviorally-aligned AV served as a *traffic stabilizer*, dampening oscillatory congestion and even achieving a fuel consumption reduction relative to human-driven traffic.

In the second part, we presented CLEAR, a novel explainability framework that uses large language models(LLMs) to generate natural-language rationales for an AV's actions, augmented with domain-specific verification modules. Our approach produces explanations grounded in the true state and decision logic of the AV's reinforcement learning controller, addressing key limitations of prompt-based or supervised explanation methods. We designed multiple validators to systematically check an explanation's factual alignment with the controller's behavior, adherence to traffic context, and logical consistency. By filtering and refining LLM outputs through these validators, CLEAR ensures that the final explanations remain accurate and coherent even in complex driving scenarios. In evaluations on realis-

tic traffic scenarios, CLEAR substantially outperformed all baseline methods, achieving the highest agreement with expert-annotated rationales and the lowest action-prediction error. These results demonstrate that integrating verifiability into the explanation-generation process yields explanations that are more truthful and provide more nuanced, context-specific justifications of the AV's behavior.

Together, these contributions are a step towards addressing the human-AI alignment problem. A driving policy that internalizes human-like objectives provides a sound foundation for explanation, as its decisions are inherently more interpretable and acceptable to human stakeholders. Conversely, the ability to explain an AV's actions builds trust and can reveal whether the policy's behavior truly aligns with human values. By jointly modeling human driving behavior and enabling humans to understand the AV's reasoning, we make progress toward autonomous systems that can be safely and harmoniously integrated into human-dominated traffic.

4.2 Future Directions

While the results are promising, there remain several opportunities to extend this work. One immediate avenue is to deploy and evaluate our behaviorally aligned policy in richer traffic environments, such as multi lane highways or urban intersections, where interactions are more complex. This would test the scalability of our IRL-derived policy under more diverse and realistic conditions. On the explainability side, future work could explore scaling the CLEAR framework to handle real-time decision explanation and integrating it with driver feedback loops. Enhancing the depth of logical verification, by using formal methods, may further improve the reliability of the generated explanations. Pursuing these extensions will move us closer to autonomous driving systems that are optimized for performance and safety, while also being deeply aligned with human behavior and transparently accountable to their human partners.

In conclusion, this thesis has demonstrated that addressing both dimensions of human-AI alignment, behavioral compatibility and decision transparency, creates a foundation for autonomous systems that can function effectively within real world environments. By learning from demonstrations rather than relying solely on engineered objectives, and by providing verifiable explanations, we enable autonomous vehicles that are socially intelligent. As autonomous systems increasingly enter our shared spaces, this integrated approach to alignment offers a pathway toward technologies that benefit society while respecting human values, preferences, and understanding.

Bibliography

[1] Dario Amodei et al. "Concrete problems in AI safety". In: *arXiv preprint arXiv:1606.06565* (2016).

[2] Matthew Barth and Kanok Boriboonsomsin. "Energy and emissions impacts of a freeway-based dynamic eco-driving system". In: *Transportation Research Part D: Transport and Environment* 14.6 (2009), pp. 400–410.

[3] Tom B. Brown et al. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[4] Tianzhe Chu et al. "SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training". In: *arXiv preprint arXiv:2501.17161* (2025).

[5] Shehzaad Dhuliawala et al. "Chain of verification reduces hallucination in large language models". In: *arXiv preprint arXiv:2309.11495* (2023).

[6] Justin Fu, Katie Luo, and Sergey Levine. "Learning robust rewards with adversarial inverse reinforcement learning". In: *International Conference on Learning Representations*. 2018.

[7] Ian Goodfellow et al. "Generative adversarial nets". In: *Advances in neural information processing systems*. Vol. 27. 2014.

[8] Amaury Heuillet, Florian Coutarel, and Nicolas Baskiotis. "Explainable reinforcement learning: A survey and comparative study". In: *Artificial Intelligence* 316 (2023), p. 103825.

[9] Jonathan Ho and Stefano Ermon. "Generative adversarial imitation learning". In: *Advances in neural information processing systems*. Vol. 29. 2016.

[10] Susmit Jha et al. "Dehallucinating Large Language Models Using Formal Methods Guided Iterative Prompting". In: *Proceedings of the IEEE International Conference on Assured Autonomy (ICAA)*. 2023.

[11] Ziwei Ji et al. "Survey of Hallucination in Natural Language Generation". In: *ACM Computing Surveys* 55.12 (2023), p. 248.

[12] K. Lee et al. "Traffic smoothing via learned autonomous vehicle control: Results from the MegaVanderTest field experiment". In: *Transportation Research Part C: Emerging Technologies* (2025).

[13] Patrick S. H. Lewis et al. "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2020.

[14] Nathan Lichtlé et al. "Deploying traffic smoothing cruise controllers learned from trajectory data". In: *2022 International Conference on Robotics and Automation (ICRA)*. IEEE. 2022, pp. 2884–2890.

[15] Scott M. Lundberg and Su-In Lee. "A Unified Approach to Interpreting Model Predictions". In: *Advances in Neural Information Processing Systems (NIPS)*. 2017.

[16] Negar Mehr and Dorsa Sadigh. "Communication-Based Cooperative Planning of Vehicle Platoons in Congested Traffic". In: *IEEE Transactions on Intelligent Transportation Systems* 22.10 (2021), pp. 6229–6239.

[17] Andrew Y Ng, Daishi Harada, and Stuart Russell. "Policy invariance under reward transformations: Theory and application to reward shaping". In: *Proceedings of the Sixteenth International Conference on Machine Learning* (1999), pp. 278–287.

[18] Andrew Y Ng and Stuart J Russell. "Algorithms for inverse reinforcement learning". In: *Proceedings of the Seventeenth International Conference on Machine Learning*. 2000, pp. 663–670.

[19] Zhenghao Peng et al. "Improving Agent Behaviors with RL Fine-tuning for Autonomous Driving". In: *Proceedings of the Conference on Robot Learning (CoRL)*. 2021.

[20] Dean A Pomerleau. "Efficient training of artificial neural networks for autonomous navigation". In: *Neural computation* 3.1 (1991), pp. 88–97.

[21] Elisabeth Puiutta and Eric Veith. "Explainable Reinforcement Learning: A Survey". In: *Lecture Notes in Computer Science (XAI: Concepts, Algorithms, and Applications)* 12200 (2020), pp. 77–95.

[22] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. ""Why Should I Trust You?": Explaining the Predictions of Any Classifier". In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*. 2016.

[23] Stéphane Ross and Drew Bagnell. "Efficient reductions for imitation learning". In: *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2010, pp. 661–668.

[24] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. "A reduction of imitation learning and structured prediction to no-regret online learning". In: *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. JMLR Workshop and Conference Proceedings. 2011, pp. 627–635.

[25] Stuart Russell. *Human compatible: Artificial intelligence and the problem of control*. Penguin, 2019.

[26] David Schrank, Tim Lomax, and Bill Eisele. "2023 Urban Mobility Report". In: *Texas A&M Transportation Institute* (2023).

[27] John Schulman et al. "Proximal Policy Optimization Algorithms". In: *arXiv preprint arXiv:1707.06347*. 2017.

[28] Amrith Setlur et al. "Scaling Test-Time Compute Without Verification or RL is Suboptimal". In: *arXiv preprint arXiv:2502.12118* (2025).

[29] Raphael E Stern et al. "Dissipation of stop-and-go waves via control of autonomous vehicles: Field experiments". In: *Transportation Research Part C: Emerging Technologies* 89 (2018), pp. 205–221.

[30] Yuki Sugiyama et al. "Traffic jams without bottlenecks—experimental evidence for the physical mechanism of the formation of a jam". In: *New Journal of Physics* 10.3 (2008), p. 033001.

[31] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. 2nd ed. MIT press, 2018.

[32] Martin Treiber, Andreas Hennecke, and Dirk Helbing. "Congested traffic states in empirical observations and microscopic simulations". In: *Physical Review E* 62.2 (2000), pp. 1805–1824. DOI: 10.1103/PhysRevE.62.1805.

[33] Martin Treiber and Arne Kesting. *Traffic Flow Dynamics: Data, Models and Simulation*. Springer-Verlag Berlin Heidelberg, 2013.

[34] Eugene Vinitsky et al. "Benchmarks for reinforcement learning in mixed-autonomy traffic". In: *Conference on Robot Learning*. PMLR. 2018, pp. 399–409.

[35] Eugene Vinitsky et al. "Benchmarks for reinforcement learning in mixed-autonomy traffic". In: *Conference on Robot Learning*. PMLR. 2018, pp. 399–409.

[36] Eugene Vinitsky et al. "Optimizing mixed autonomy traffic flow with decentralized learning-based control". In: *IEEE Transactions on Intelligent Transportation Systems* (2023).

[37] Xuezhi Wang et al. "Self-Consistency Improves Chain of Thought Reasoning in Language Models". In: *arXiv preprint arXiv:2203.11171* (2022).

[38] Jason Wei et al. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". In: *Advances in Neural Information Processing Systems (NeurIPS)*. 2022.

[39] Cathy Wu et al. "Flow: A Modular Learning Framework for Autonomy in Traffic". In: *IEEE Transactions on Robotics* 38.4 (2022), pp. 2014–2031. DOI: 10.1109/TRO.2021.3133564.

[40] Fangyu Wu et al. "MegaVanderTest: Large-scale field deployment of autonomous vehicles for traffic flow stabilization". In: *Nature Communications* 15.1 (2024), p. 782.

[41] Yongqi Xue, Jin Zhao, and George Karypis. "RCoT: Detecting and Rectifying Factual Inconsistency in Reasoning by Reversing Chain-of-Thought". In: *arXiv preprint arXiv:2305.11499* (2023).

[42] Brian D Ziebart et al. "Maximum entropy inverse reinforcement learning". In: *Proceedings of the 23rd national conference on Artificial intelligence*. Vol. 3. 2008, pp. 1433–1438.