

Bridging the Semantic Gap in Autonomous Braking Systems: A Neuro-Symbolic XAI Approach

Amil Roohani Dar

amil.rohani@uokajk.edu.pk

Department of CS&IT, University of Kotli (AJ&K), Kotli, AJ&K, Pakistan

Faisal Riaz

faisal.riaz@must.edu.pk

Control, Automotive and Robotics Lab, Affiliated Lab of National Centre of Robotics and Automation (NCRA), HEC, Islamabad, Pakistan; Department of Computer Science & Information Technology, Mirpur University of Science & Technology (MUST), Mirpur, Pakistan

Corresponding Author: Amil Roohani Dar amil.rohani@uokajk.edu.pk

Received: 05-02-2026

Revised: 28-02-2026

Accepted: 06-03-2026

Published: 22-03-2026

ABSTRACT

Contemporary deep learning systems in autonomous vehicles extend beyond speed and accuracy toward interpretability and trustworthiness. Regulators, safety engineers, and end-users increasingly demand transparency in how these systems operate. Deep learning models are helpful in many fields of life for human beings but their inherent black-box nature creates a significant disconnect between the low-level technical data they process and the high-level concepts that humans use to interpret and trust automated decisions. Traditional tools such as LIME and SHAP help us in understanding the models inner operations but are often too slow and produce visuals that are difficult to interpret. This paper introduces a new framework called Neuro-Symbolic explainable AI (XAI) to address this semantic gap, specifically designed for real-time brake light prediction and distance monitoring. We are combining the advanced perception model with distance monitoring sensors, translating technical detections into a formal knowledge base of traffic rules and safety concepts. This mixed approach allows the system to generate natural language explanations that justify vehicle actions, such as "Warning: Lead vehicle braking at a critical distance."

Keywords: XAI, explainable AI, LIME, SHAP, CAM, Black-box AI, White-box AI.

INTRODUCTION

The continued development of Autonomous Vehicles (AVs) is transforming modern transportation, offering promising improvements in road safety and overall traffic efficiency. These systems execute driving activities without human intervention and depend on a complex integration of algorithms, AI models and sensors to get information of road entities. However, despite high levels of predictive accuracy, the widespread adoption of AVs is currently hindered by the "black-box" nature of deep neural perception and control networks [1]. As road traffic accidents continue to rise—exemplified by a 7.7% increase in major accidents in Australia in 2024—the demand for transparent and interpretable AI is increasingly critical for regulatory compliance and public safety [2], [3]. User acceptance and functional safety depend on easy-to-interpret textual explanations that allow end-users to understand the specific triggers behind a vehicle's behavior. Current research in explainability often distinguishes between introspective explanations, which identify internal model triggers, and rational-Identify applicable funding agency here. If none, delete this izations, which justify outputs after the fact [4]. While vision-based brake detection has been shown to improve the reaction time of Adaptive Cruise Control (ACC) significantly more than traditional radar alone, current state-of-the-art models like YOLOv8 often lack the inherent reasoning to explain why a specific action was triggered [5], [6]. To fix this problem, many researchers have used explainable AI (XAI) techniques like Class Activation Mapping (CAM) [7], LIME, or SHAP [8]. In the field of automotive industry due to real time requirements these methods face significant

challenges. The authors in [9] criticized and highlighted that current black-box explainable AI techniques demand substantial computational resources and lack stability and this make them unsuitable for real-time applications on-board [9]. In addition these methods are struggling to overcome the semantic gap which refers to the disconnect between low-level technical features and high-level human concepts like traffic safety rules. Although the visual attention model can identify image regions influencing a model output but it cannot provide a grounded and logical explanation for the causal behavior of the vehicle.

By presenting an innovative Neuro-Symbolic XAI frame-work this research overcomes these limitations and is illustrated through a real-time case study on brake light prediction and distance tracking. Our approach utilized the perception capabilities with high precision of YOLOv8 with the logical transparency of an ontological reasoning engine. We use the ontology to bridge the semantic gap and translates raw sensor data into a formal knowledge base that represents objects and their relationships on the road [10], [11]. Our framework produces introspective explanations rooted in symbolic logic as compared to traditional AI models that depend solely on mathematical weights. The primary contribution of this study is a to build a framework that effectively balances high accuracy with low computational overhead. Instead of heatmaps or visual features this work offer more specific, linguistically detailed warnings for users of AV technologies which provide a clear and traceable explanation of AI decisions. This ultimately contributes to greater public confidence and safety in autonomous driving systems [12], [13].

Contributions of the Work

1. **Development of a Hybrid "Third-Wave" AI Architecture:** We introduced a single and cohesive approach that combines the deep learning model specifically YOLOv8 with the clear and organized reasoning provided by symbolic logic through the use of semantic ontology.
2. **Bridging the Semantic Gap in Autonomous Perception:** Unlike traditional explanation AI approaches that depend on pixel-level explanations such as heatmaps, this study presents a semantic bridge that connects sensor data with visual detections.
3. **Real-Time Causal Explanation Capability:** Instead of post-hoc explanations of LIME and SHAP this work is the first step toward real-time introspective causal reasoning.

LITERATURE REVIEW

Most recent studies in this area typically offer heatmaps or basic descriptions to illustrate how a vehicle operates. However, we aimed to explore this further. Our work integrates explainable AI with neuro-symbolic reasoning to incorporate formal logic into the analysis. By employing an ontological engine, we are able to generate explanations that are organized and can be validated. In the coming sections we provide a brief summary of the literature relevant to these domains.

Explainability in Autonomous Driving

Although tremendous improvement has been made in the field of AI but new challenges have also emerged every day, including the difficulty of understanding how these AI models make decisions. Users of these technologies demand transparency that is how these models are working like the drivers or passengers of AVs [14], [15]. Many complex machine learning models, particularly deep neural networks, operate as "black boxes" after training, meaning their internal workings are not easily interpretable. In contrast, simpler models such as decision trees and linear classifiers are inherently interpretable [16]. This issue of transparency is not just a model technical issue but also a significant social risk. Significant concerns have been raised regarding the fairness and accountability of automated decisions produced by machine learning and deep learning systems, particularly within communities directly affected by such technologies [17]. Bridging the gap between the complexity of artificial

intelligence and the ability of humans to understand it involves the consistent and structured use of explanations for algorithms. This objective lies at the heart of the field of Interpretable AI. [18], [19].

Within the automotive sector, providing explanations serves as a vital strategy for building public reliance on self-driving systems [20]. However, the characteristics of road setup introduce distinct challenges for autonomous vehicles and often base their actions on environmental data or internal logic that is not immediately apparent to the human occupant. This complicates the process of clarifying vehicle behavior. Whenever an AV makes a decision, it often acts on unseen information provided by intelligent infrastructure or nearby traffic [21].

Since these sources are outside the user's immediate field of view (like remote traffic lights or hidden vehicles), explanations must do more than report what the car sees [22]. They must also translate the influence of these external systems into terms the driver can understand. Although informing drivers about a car's actions and intentions increases safety the way this information is presented is equally important. It also carries the side effect of increased anxiety due to the heavy cognitive demands of multitasking [23]. Users often find them-selves overwhelmed by the volume of information they must interpret in semi-autonomous settings. Hence, moving toward full autonomy changes this dynamic because the passenger is relieved of driving duties, the resulting drop in workload may make them more receptive to detailed system explanations [24].

In the field of automated driving if companies want to gain widespread acceptance then there is need to open the "black-box" model's operations by utilizing XAI. This is particularly crucial for advanced driver assistance systems (ADAS) [25], [26], the kind that lend a hand with steering and staying on course. The ability to explain vehicle decisions is therefore not merely a technical requirement but a critical factor in establishing driver trust and maintaining situational awareness [27]. XAI provides a structured means of making complex model behavior accessible to non-expert users, thereby increasing confidence and supporting broader adoption of autonomous systems [28].

Visual Saliency and Model-Agnostic XAI

Current practitioners in the field of XAI typically rely on tools such as SHAP or LIME or CAM for assistance. Although LIME and SHAP represent the gold standard for model interpretability they often prove difficult to implement within the context of autonomous vehicles. These techniques either require excessive computational power which slows performance or fail to deliver sufficiently granular insights thus rendering them unsuitable for such high stakes environments. Our objective is to adopt a more specialized approach and depart from these established but limited methods. The authors in [29] using SHAP analysis helps the user to see exactly which features drive the decisions of an InceptionV3 model, allowing us to explain its behavior for individual cases as well as its overall logic. The results of SHAP algorithm explain the complex models and give us a way to see how they work. This process enhances overall clarity and empowers different industry specialists to establish significantly more secure protocols and more informed choices [30]. The way AVs interpret the choices and make decisions could be a mystery which is why Cui et al. [31] proposed using both Random Forest and deep RL. Their goal was to move away from 'black-box' systems and create a framework that people can actually understand. Their use of SHAP values allows for a simplified feature space. This type of research bring up the SHAP's versatility as a robust tool for extracting meaningful insights and validating model behavior. The authors in [32] for XAI developments produced a combined LIME-SHAP frame-work tailored for AV systems. By combining the efficiency of LIME with accuracy of SHAP this approach provides clear and more actionable insights without compromising on real-time processing needs. Class Activation Mapping is also a famous explainability technique for black box interpretation [6]. Here are some of the most famous CAM methods: Grad-CAM, Grad-CAM++ [33], Layer-CAM [34], HiResCAM [35],

Eigen-CAM [36]. These model helps in only highlighting the region of input image to deep learning models which affects the model output.

However, these methods are often criticized for their high computational latency and lack of stability in high-speed, real-time environments [6]. Furthermore, these visual "heatmaps" create a semantic gap, as they identify where a model is looking without explaining the causal rules (e.g., traffic safety thresholds) that triggered the decision.

Neuro-Symbolic and Ontology-Based Approaches

To overcome the limitations of purely data-driven models, recent research has increasingly explored Neuro-Symbolic (NeSy) architectures. These approaches aim to combine the perceptual capabilities of deep learning with the structured reasoning provided by symbolic systems [37]. In this context, ontologies have been widely adopted as a formal mechanism for representing domain knowledge, including road environments and traffic rules [38], [39].

A number of studies have investigated the integration of ontologies with explainable AI. For instance, Confalonieri et al. [40] introduced a neural-symbolic framework that extends the TREPAN algorithm by incorporating domain ontologies to guide explanation generation. In the same way the Amara and co authors in [41] mix the ontological knowledge with concept based explanations to enhance the interpretability to diagnose the disease. Few more studies have investigated the use of semantics alongside local explanation methods for example the author Igoche et al. [42] and the author Sharma et al. [43] unified the ontology with LIME to improve explanation fairness and structure across various application areas. Furthermore, Bellucci et al. [44] proposed an ontology-based counterfactual explanation framework to ensure consistency and generate meaningful alternative explanations. Even though the contributions of these studies play a major role in semantics but these approaches are developed for relatively static environments in the domains of healthcare, education, and network security. These are not implemented and tested in real-time environment and still the perception driven systems remains limited. Particularly in driving scenarios where decisions must be made continuously based on dynamic visual inputs.

In comparison the our proposal uses a reasoning approach to combine the perception with decision-making. With the inclusion of important visual cues such as the leading vehicle distance and whether its brake lights are on or off the system creates real-time explanations that are logically sound and based on clear reasoning. This method thinks ahead of descriptive interpretation toward causal reasoning that is both interpretable and verifiable in safety-critical driving conditions.

OVERVIEW OF THE PROPOSED FRAMEWORK

In this section the overview of the proposed framework is given in detail. The suggested system is a real-time risk assessment framework for autonomous driving powered by ontology that combines deep learning-based perception and radar-based distance measurement, and an OWL/DL ontology reasoning engine. We want to transform raw sensor data into semantic knowledge and further into explainable and interpretable driving decisions.

The overall architecture as shown in Figure. 1 follows a multi-layer pipeline consisting of four main components: the Perception Layer, the Mapping and Semantic Layer, the Ontology and Reasoning Engine, and the Output Decision Layer. This hierarchical structure enables a systematic transformation from low-level sensor data to high-level symbolic reasoning and explainable risk assessment.

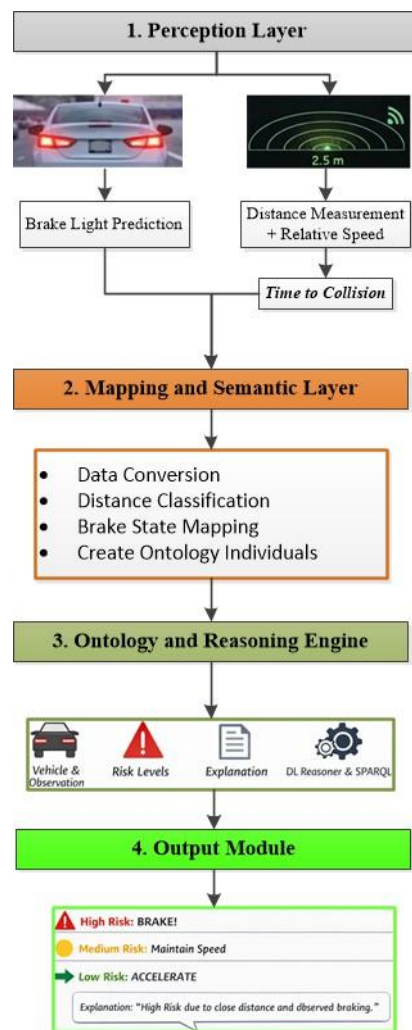


Fig. 1. Workflow of the proposed methodology

Perception Layer (Data Acquisition)

The Perception Layer is responsible for capturing multi-modal environmental information using heterogeneous sensors. In this framework, a YOLOv8-based vision module is employed to detect the brake light status of the leading vehicle [5], [6]. The output of this module is represented as either Brake ON or Brake OFF, providing a semantic interpretation of vehicle behavior from visual data. In parallel, a radar or LiDAR sensor provides continuous distance measurements between the ego vehicle and the leading vehicle. This part helps provide a clear understanding of the space around the environment.

The output results of these these sensor provide critical information but this information is in raw numerical or visual form and is not suitable for direct symbolic reasoning. This limitation motivates the introduction of a semantic transformation layer.

MAPPING AND SEMANTIC LAYER (CRITICAL COMPONENT)

In the proposed framework Figure. 1 the Mapping and Semantic Layer plays the role of a bridge between machine perception and symbolic reasoning. The primary purpose of this is to convert continuous sensor outputs into discrete semantic representations that can be processed by the ontology.

During this phase the numerical distance values are discretized into symbolic categories grounded in physically meaningful highway driving thresholds. Rather than using arbitrary fixed values, the discretization scheme is derived from Time-to-Collision (TTC) analysis, where TTC is computed as the ratio of the observed headway distance to the relative vehicle speed. This ensures that risk classification adapts to vehicle speed rather than relying on context-blind fixed distances. To formally represent driving context, the ontology introduces a *DrivingContext* class with three mutually disjoint sub-classes: *Urban*, *Suburban*, and *Highway*. All scenarios evaluated in this work are linked to the *HighwayContext* individual via the *hasDrivingContext* object property, reflecting real motorway operating conditions. For highway driving scenarios, the distance thresholds are defined as follows. The critical zone, represented by the *Near* distance level, corresponds to headway distances of 30 meters or less. At a typical highway speed of 90 km/h (25 m/s), this corresponds to a TTC of approximately 0.8 seconds, placing the vehicle within the emergency-braking region requiring immediate intervention. The warning zone, represented by the *Medium* distance level, covers headway distances between 30 meters and 80 meters, corresponding to a TTC range of 1.5 to 3.0 seconds at highway speeds, representing an active monitoring condition. The safe zone, represented by the *Far* distance level, corresponds to headway distances exceeding 80 meters, where TTC exceeds 3.0 seconds, indicating a stable and safe following condition. These thresholds are formally encoded as OWL equivalent class axioms on the *NearRiskCondition*, *MediumRiskCondition*, and *FarRiskCondition* classes respectively, with supporting justification annotations attached directly to each class within the ontology. Two new data properties, *hasVehicleSpeed* and *hasTTCValue*, are introduced to record the vehicle speed and computed TTC value for each distance observation. A SWRL rule, formally defined within the ontology, dynamically computes TTC as the ratio of observed distance to vehicle speed, ensuring that risk classification adapts to varying vehicle speeds across different highway scenarios. For example, a distance of 20.0 meters is mapped to *NearDist*, a distance of 50.0 meters is mapped to *MediumDist*, and a distance of 100.0 meters is mapped to *FarDist*.

Similarly, the output of the YOLOv8 brake light detection

Module is transformed into semantic risk factors. A *Brake ON* state is associated with the *BrakeOnFactor*, and a *Brake OFF* state is associated with the *BrakeOffFactor*. This transformation enables raw perception data to be represented in a symbolic format compatible with the ontology. In addition, each sensor observation is created as an OWL individual in real time. For example, a measurement of distance might be represented as *ObsDistance1*, which is an instance of the class *DistanceObservation*, and a detection of a brake light could be represented as *ObsBrake1*, which is an instance of the class *BrakeLightObservation*. These individuals are continuously added to the ontology to enable real-time reasoning. The responsibility of this layer is important because of linking symbols to their semantics as it permit machines to change their perceptions into forms of knowledge that is important for user understanding.

Ontology and Reasoning Engine

The ontology reasoning engine, implemented in Web Ontology Language (OWL/DL), provides the foundational intelligence of the proposed framework. Vehicles are categorized as either *Ego* or *Leading*, and the system evaluates distance and brake light inputs to assign a corresponding risk level of *Low*, *Medium*, or *High*. By treating sensor inputs as empirical observations, the framework derives high-level logical outputs through Description Logic (DL) inference. A vehicle is classified as *HighRisk* when it simultaneously satisfies a *NearDist* observation and a *BrakeOnFactor*, expressed as an OWL equivalent class axiom enabling automatic reasoner classification. The ontology organizes driving scenarios through a *DrivingContext* class with three disjoint subclasses: *Urban*, *Suburban*, and *Highway*. Each ego vehicle is linked to its context via the *hasDrivingContext* property, enabling context-specific threshold schemes. Vehicle speed and Time-to-Collision (TTC) are modeled as semantic properties *hasVehicleSpeed* and *hasTTCValue* on each *DistanceObservation* individual. TTC captures the temporal urgency of a driving situation and is expressed within the ontology as a SWRL rule:

```
DistanceObservation(?o) ^
hasDistanceValue(?o, ?d) ^
hasVehicleSpeed(?o, ?s) ^
swrlb:divide(?ttc, ?d, ?s)
→ hasTTCValue(?o, ?ttc)
```

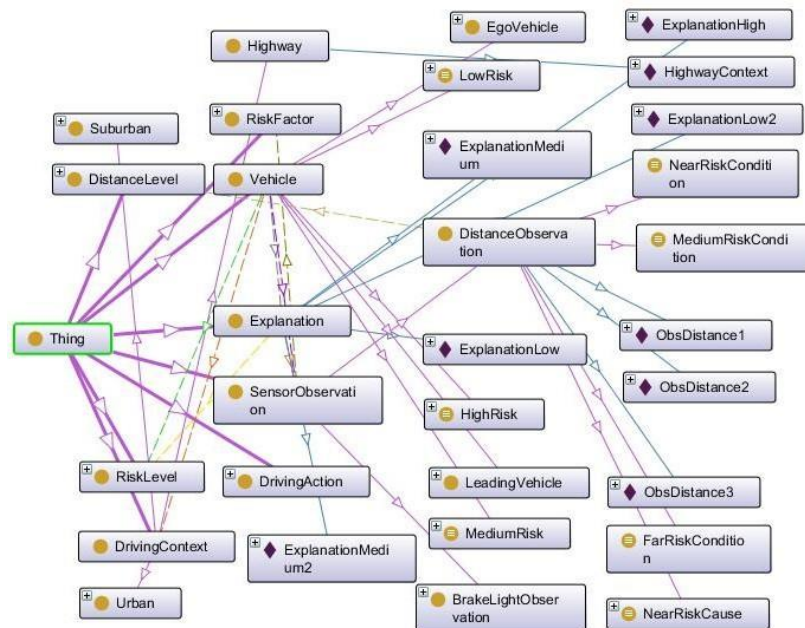


Fig. 2. Ontology structure and reasoning relationships for autonomous driving decision-making.

Each inferred risk level is linked to an explanation entity storing a structured natural language justification via `hasExplanationText`, including driving context, speed, distance, TTC value, and triggered action, ensuring every decision is traceable and human-understandable.

Output Module (Decision and Explainability)

The Output Module is responsible for generating the final risk classifications and the corresponding driving decisions. The system classifies driving scenarios as High Risk, Medium Risk, or Low Risk based on the reasoning provided. Appropriate driving actions are then assigned to each classification, such as immediate braking, maintaining the current speed, or proceeding with safe acceleration. In addition to categorizing risks, the system also creates easy to understand explanations for every decision it makes. For example, a high-risk scenario may be explained as “High risk due to near distance and observed braking behavior.” This guarantees that the process of making decisions is clear and can be understood.

Unlike conventional black-box models, the proposed frame-work generates organized outputs that include a risk level, an explanation for the reasoning, and suggested actions. This enhances trust, safety, and interpretability in autonomous driving systems while also supporting regulatory compliance requirements.

ONTOLOGY REASONING AND EXPLAINABILITY

RESULTS

The proposed ontology was tested by using dummy sensor data that shows the status of brake lights and the distance between vehicles. Protege was used to implement the ontology as shown in Figure. and test the the ontology, and DL reasoner which is freely available in Protege editor to infer risk levels, the driving actions, and the corresponding explanations.

We can visualize and validate the ontology result in the Protege. The visualization of the ontology in Figure 2 demonstrates how sensor observations are semantically mapped to risk levels and driving actions through explicitly defined relationships, enabling transparent reasoning. The Figure. 3 illustrates the hierarchy of classes, and it also provides the instance-level relationships defined within the system.

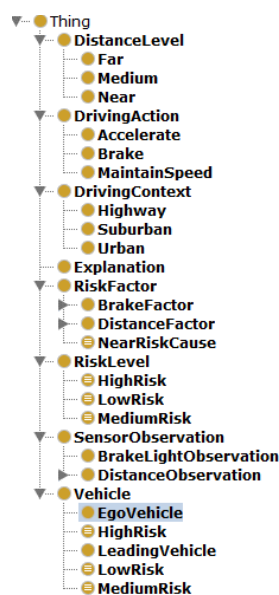


Fig. 3. Class Hierarchy of Explanation Ontology

A representative scenario was considered in which the inter-vehicle headway distance was 20.0 meters at a high-way speed of 90 km/h (25 m/s), and the brake light of the leading vehicle was detected. The computed Time-to-Collision (TTC) for this scenario is 0.8 seconds, placing the vehicle within the emergency-braking region. Based on the physically grounded highway threshold scheme, the ontology classified this distance as a *Near* condition via the *NearRiskCondition* equivalent class axiom, which is defined as a *DistanceObservation* with *hasDistanceValue* ≤ 30.0 meters. The *hasTTCValue* data property recorded a value of 0.8 seconds for this observation, consistent with the critical intervention threshold. Using the defined class axioms and object property relationships, the reasoner inferred the situation as a *HighRisk* condition. Two additional scenarios were evaluated: a medium-risk scenario at 50.0 meters and 108 km/h (TTC = 1.67s), resulting in *MediumRisk* and *MaintainSpeed* action; and a low-risk scenario at 100.0 meters and 120 km/h (TTC = 3.03s), resulting in *LowRisk* and *Accelerate* action.

Consequently, the ontology triggered the corresponding driving action (*BrakeAction*) and generated an associated ex-planation. The results of the ontology reasoning are summarized in Table I. ‘latex ““

Table 1: Ontology-Based Reasoning Results Across Highway Driving Scenarios. Ttc Values Computed as Distance Divided by Vehicle Speed Following Physically Grounded Highway Threshold Scheme.

Scenario	Distance / Brake	Risk / Action	Explanation
S1	20 m / On	High / Brake	Near + braking, TTC=0.8s
S2	50 m / Off	Medium / Maintain	Warning zone, TTC=1.67s
S3	100 m / Off	Low / Accelerate	Safe zone, TTC=3.03s
S4	40 m / Off	Medium / Maintain	Warning zone, no braking

^a Distance levels: Near (≤ 30 m), Medium (30–80m), Far (> 80 m).

^b TTC computed at highway speeds: S1=90 km/h,

S2=108 km/h, S3=120 km/h, S4=108 km/h.

Explanation Generation

The ontology generates human-understandable explanations through explicit semantic relationships between observations, risk factors, and inferred decisions. This explanation is derived from the object properties *hasCause* and *hasExplanationText*, which link the inferred risk level to its contributing factors. Un-like black-box explainability approaches, the proposed method provides structured and interpretable reasoning by explicitly identifying the causal factors behind each decision. To validate the ontology result for explanation generation for driver of AV we use the SPARQL query engine. Figure. 4 shows the query and its associated result below the query.

SPARQL query:		
<pre> PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> PREFIX owl: <http://www.w3.org/2002/07/owl#> PREFIX xsd: <http://www.w3.org/2001/XMLSchema#> PREFIX onto: <http://www.semanticweb.org/amil/ontologies/2026/3/untitled-ontology-8#> SELECT ?vehicle ?risk ?explanation WHERE { ?vehicle rdf:type onto:EgoVehicle . ?vehicle onto:hasRiskLevel ?riskInstance . ?riskInstance rdf:type onto:HighRisk . ?vehicle onto:hasExplanationText ?exp . ?exp onto:hasExplanationText ?explanation . } </pre>		
vehicle	risk	explanation
EgoCar	HighRisk	"HIGH RISK Highway Context Speed: 90 km/h (25 m/s) Distance: 20.0m TTC: 0.8s. TTC is below the critical intervention threshold of 1.5s. Brake light detected on leading vehicle confirming deceleration. Risk classification: HighRisk. Triggered Action: EMERGENCY BRAKE immediately.

Fig. 4. Causal Message for AV user.

Inference Latency Benchmarking

To substantiate the real-time viability of the proposed frame-work, a formal inference latency benchmark was conducted over 100 independent runs, with results contextualized against latency figures reported in the published XAI and autonomous driving literature.

1. **Measurement Methodology:** Ontology inference latency was measured using the ROBOT ontology operations tool (v1.9.5) [45], which invokes the FaCT⁺⁺ Description Logic reasoner as its backend classification engine. Two latency figures were recorded:

- **Wall-clock time:** total elapsed time including JVM initialization, ontology loading, FaCT⁺⁺ reasoning, and output serialization. This represents cold-start latency incurred once at system startup.
 - **Pure reasoning time:** the FaCT⁺⁺ inference duration extracted from the ROBOT execution log, isolating the Description Logic classification step from all initialization and I/O overhead. This represents the per-query latency in a deployed system where the ontology remains loaded in memory between requests. All benchmarks were executed on Google Colab (Intel Xeon CPU, single-threaded, Ubuntu 22.04, Python 3.12, Java 21, ROBOT v1.9.5).
2. **Comparability with Published Latency Figures:** This comparative approach follows the methodology established in the XAI literature. Retzlaff et al. [46] identify runtime performance as one of six essential criteria for XAI method selection and operationalize it as a direct comparative metric between ante-hoc and post-hoc approaches. Rjoub et al. [9] applied the same comparative logic in the autonomous driving domain, concluding that SHAP and LIME demand substantial computational resources that render them unsuitable for real-time on-board deployment. The present benchmark provides the quantitative evidence that those qualitative assessments implied but did not measure.

Description Logic reasoning complexity is determined by the ontology's axiom structure and is invariant to input dimensionality. Post-hoc explanation latency, by contrast, scales with feature count and perturbation sample size [46]. Published post-hoc latency figures are therefore lower-bound estimates relative to real-world high-dimensional autonomous driving in-

Table 2: Inference Latency Comparison Between the Proposed Ontology Framework and Published Post-Hoc Xai Benchmarks ($n = 100$ Runs for This Work; Literature Values Are Reported as Published Ranges).

Method	Mean (ms)	$\pm$ Std	Real-time
Proposed Ontology – FaCT ⁺⁺ Pure Reasoning	0.26	0.44	✓
Proposed Ontology – FaCT ⁺⁺ Cold-Start [†]	2093.70	307.02	✓ [†]
TreeSHAP (tabular) [47]	10–50	—	×
LIME (tabular) [47]	100–500	—	×
Kernel SHAP (tabular) [47]	1000–5000	—	×
Hybrid LIME-SHAP (ResNet-18) [8]	280	—	×
Hybrid LIME-SHAP (SegNet) [8]	3889	—	×

[†] One-time startup overhead incurred during deployment initialization. puts, where the performance gap between ontology reasoning and post-hoc methods would widen further.

Results: Table II presents the complete latency results. Pure FaCT⁺⁺ reasoning time of 0.26 ms (± 0.44 ms) confirms that ontology classification operates well within real-time constraints for autonomous driving applications. The cold-start wall-clock time of 2093.70 ms (± 307.02 ms) reflects one-time JVM initialization and ontology loading overhead incurred at deployment startup. In a deployed system with the ontology pre-loaded in memory, the per-query latency equals the pure reasoning time of 0.26 ms.

Published benchmarks report TreeSHAP latencies of 10–50 ms, LIME latencies of 100–500 ms, and Kernel SHAP latencies of 1,000–5,000 ms per explanation on tabular sensor data [47]. In the autonomous driving domain, Tahir et al. [8] reported hybrid LIME-SHAP inference times of 280 ms for ResNet-18 and 3,889 ms for SegNet on the KITTI dataset. The proposed ontology achieves a pure reasoning time at least 38-fold lower than the minimum published SHAP latency and more than 380-fold lower than published LIME figures.

Post-hoc methods incur their reported latency on every query. This cost cannot be reduced below a theoretical floor determined by perturbation sample count in LIME or Shapley value computation in SHAP. The proposed ontology's per-query cost is fixed at 0.26 ms regardless of query volume, since inference is performed over a pre-loaded knowledge base rather than re-derived per instance. This property is architecturally decisive in autonomous driving, where risk classifications must be generated continuously at sensor frame rates with no tolerance for per-query computational growth.

DISCUSSION

The results confirm that the ontology successfully bridges the gap between raw sensory data and meaningful semantic insights, facilitating transparent autonomous decision-making. The use of physically grounded TTC-based distance thresholds enables continuous sensor values to be categorized into meaningful qualitative levels of Near (≤ 30 m, $TTC < 1.5$ s), Medium (30–80 m, $TTC 1.5\text{--}3.0$ s), and Far (> 80 m, $TTC > 3.0$ s), all evaluated within a Highway driving context. This context-aware discretization scheme ensures that risk classification reflects the physical dynamics of the driving scenario rather than arbitrary fixed values.

The ontology additionally provides a traceable lineage that connects raw observations to their subsequent decision-making outputs:

Sensor Observations $\rightarrow$ Risk Factors $\rightarrow$ Risk Level $\rightarrow$ *Driving Action* $\rightarrow$ *Explanation*

This way of structured process of reasoning improves transparency and interpretability of actions of AVs which are critical requirements for XAI in automotive systems.

Unlike traditional approaches of XAI the proposed frame-work based on ontology provides explicit semantic reasoning and human understandable and readable justifications. Because it explains the logic behind every move, this system works well for safety focused jobs where you can't afford to guess why something happened.

However, the current evaluation is limited to predefined scenarios within the ontology environment. The effectiveness of the approach in real-time autonomous driving systems depends on the accuracy of upstream perception modules and the integration of dynamic sensor data.

CONCLUSION

We present a human-readable, ontology-driven explanation framework for AV users during driving which helps in under-standing the driving actions of the Ego vehicle. The proposed framework is novel because of an ontology based explanation module which is different way of presenting AV actions to enhance user trust. The ontology mapping and semantic layer get the information from deep learning model and vehicle sensors to make the reasoning process fast. The ontology structure and relationships for autonomous driving decision-making make it a suitable candidate for the explainability research community. The ontology additionally provides a traceable lineage that connects raw observations to their sub-sequent decision-making outputs and this reasoning improves transparency and interpretability of actions of AV. This way of explainability of driving actions would be potential future direction.

FUTURE WORK

Our ontology based proposed XAI framework shows effective semantic reasoning. Due to the vast nature of this field several directions can be explored. There is need of real time experiments to assess to proposed framework reaction time and passenger trust. Also, the ontology can be extended to incorporate additional sensor modalities.

REFERENCES

- Y. Hayashi, *Black Box Nature of Deep Learning for Digital Pathology: Beyond Quantitative to Qualitative Algorithmic Performances*. Cham: Springer International Publishing, 2020, pp. 95–101. [Online]. Available: https://doi.org/10.1007/978-3-030-50402-1_image6 (image)
- S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, “Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions,” *IEEE Access*, vol. 12, pp. 101 603–101 625, 2024.
- Bureau of Infrastructure and Transport Research Economics (BITRE). (2025) Road trauma Australia—annual summaries 2024. Accessed: 2026-04-16. [Online]. Available: <https://www.bitre.gov.au/publications/ongoing/roadimagedeathsimageaustraliaimageannualimagesummaries> (image) (image) (image) (image)
- J. Kim, A. Rohrbach, T. Darrell, J. Canny, and Z. Akata, “Textual explanations for self-driving vehicles,” 2018. [Online]. Available: <https://arxiv.org/abs/1807.11546>
- J. Pirhonen, R. Ojala, K. Kivekäs, J. Vepsäläinen, and K. Tammi, “Brake light detection algorithm for predictive braking,” *Applied Sciences*, vol. 12, no. 6, 2022. [Online]. Available: <https://www.mdpi.com/2076-3417/12/6/2804>
- A. Dar and F. Riaz, “Towards real-time explainable ai: Using class activation mapping for brake prediction in yolo-based systems,” *BRAIN: Broad Research in Artificial Intelligence and Neuroscience*, vol. 17, no. 1, pp. 7–25, 2026.
- C. Panati, S. Wagner, and S. Brüggewirth, “Feature relevance evaluation using grad-cam, lime and shap for deep learning sar data classification,” in *2022 23rd International Radar Symposium (IRS)*, 2022, pp. 457–462.
- H. A. Tahir, W. Alayed, W. U. Hassan, and A. Haider, “A novel hybrid xai solution for autonomous vehicles: Real-time interpretability through lime–shap integration,” *Sensors*, vol. 24, no. 21, 2024. [Online]. Available: <https://www.mdpi.com/1424-8220/24/21/6776>
- S. Nazat, O. Arreche, and M. Abdallah, “On evaluating black-box explainable ai methods for enhancing anomaly detection in autonomous driving systems,” *Sensors*, vol. 24, no. 11, 2024. [Online]. Available: <https://www.mdpi.com/1424-8220/24/11/3515>
- R. Ambalavanan, R. S. Snead, J. Marczika, G. Towett, A. Malioukis, and M. Mbogori-Kairichi, “Ontologies as the semantic bridge between artificial intelligence and healthcare,” *Frontiers in Digital Health*, vol. 7, 2025. [Online]. Available: <https://www.frontiersin.org/journals/digital-health/articles/10.3389/fdgth.2025.1668385>
- Y. Li, J. Tao, and F. Wotawa, “Ontology-based test generation for automated and autonomous driving functions,” *Inf. Softw. Technol.*, vol. 117, no. C, Jan. 2020. [Online]. Available: <https://doi.org/10.1016/j.infsof.2019.106200>
- A. S. d’Avila Garcez and L. C. Lamb, “Neurosymbolic AI: the 3rd wave,” *CoRR*, vol. abs/2012.05876, 2020. [Online]. Available: <https://arxiv.org/abs/2012.05876>

- U. Nawaz, M. A. ur Rahaman, and Z. Saeed, "A review of neuro-symbolic ai integrating reasoning and learning for advanced cognitive systems," *Intelligent Systems with Applications*, vol. 26, p. 200541, 2025. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S2667305325000675>
- M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, and H. P. Beck, "The role of trust in automation reliance," *Int. J. Hum. Comput. Stud.*, vol. 58, pp. 697–718, 2003. [Online]. Available: <https://api.semanticscholar.org/CorpusID:29006669>
- J. Lyons, G. Sadler, K. Koltai, H. Battiste, N. Ho, L. Hoffmann,
- A. Smith, W. Johnson, and R. Shively, *Shaping Trust Through Trans-parent Design: Theoretical and Experimental Guidelines*, 01 2017, vol. 499, pp. 127–136.
- H. Cramer, V. Evers, S. Ramlal, M. van Someren, L. Rutledge,
- N. Stash, L. Aroyo, and B. J. Wielinga, "The effects of transparency on trust in and acceptance of a content-based art recommender," *User Modeling and User-Adapted Interaction*, vol. 18, pp. 455–496, 2008. [Online]. Available: <https://api.semanticscholar.org/CorpusID:26115746>
- M. I. V. Eale, "Slave to the algorithm ? why a ' right to an explanation ' is probably not the remedy you are looking for," 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:9481286>
- A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda," *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018. [Online]. Available: <https://api.semanticscholar.org/CorpusID:5063596>
- F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv: Machine Learning*, 2017. [Online]. Available: <https://api.semanticscholar.org/CorpusID:11319376>
- N. Du, J. Haspiel, Q. Zhang, D. M. Tilbury, A. K. Pradhan, X. J. Yang, and L. P. R. Jr., "Look who's talking now: Implications of av's explanations on driver's trust, AV preference, anxiety and mental workload," *CoRR*, vol. abs/1905.08878, 2019. [Online]. Available: <http://arxiv.org/abs/1905.08878>
- C. Creß, Z. Bing, and A. C. Knoll, "Intelligent transportation systems using roadside infrastructure: A literature survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 25, no. 7, pp. 6309–6327, 2024.
- H. Zhao, X. Li, C. Xu, B. Xu, and H. Liu, "A survey of automatic driving environment perception," in *2024 IEEE 24th International Conference on Software Quality, Reliability, and Security Companion (QRS-C)*, 2024, pp. 1038–1047.
- S. Xu *et al.*, "Cognitive load and task switching in drivers: Implications for road safety in semi-autonomous vehicles," *Transportation Research Part F: Traffic Psychology and Behaviour*, vol. 107, pp. 1175–1191, 2024.

G. Wiegand, M. Eiband, M. Haubelt, and H. Hussmann, ““I’d like an explanation for that!” exploring reactions to unexpected autonomous driving,” *22nd International Conference on Human-Computer Interaction with Mobile Devices and Services*, 2020. [Online].

Available: <https://api.semanticscholar.org/CorpusID:222081054>

F. Riaz and M. A. Niazi, “Enhanced emotion enabled cognitive agent based rear end collision avoidance controller for autonomous vehicles,” *CoRR*, vol. abs/1708.01930, 2017. [Online]. Available: <http://arxiv.org/abs/1708.01930>

M. A. Butt, F. Riaz, Y. Mehmood, and S. Akram, “Reec-agent: human driver cognition and emotions-inspired rear-end collision avoidance method for autonomous vehicles,” *SIMULATION*, vol. 97, pp. 601 – 617, 2021. [Online]. Available: <https://api.semanticscholar.org/CorpusID: 234831611>

M. P. S. Lorente, E. M. Lopez, L. A. Florez, A. L. Espino, J. A. I. Mart’inez, and A. S. de Miguel, “Explaining deep learning-based driver models,” *Applied Sciences*, vol. 11, no. 8, 2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/8/3321>

A. M. A. Salih, Z. Raisi-Estabragh, I. B. Galazzo, P. Radeva,

S. E. Petersen, G. Menegaz, and K. Lekadir, “Commentary on explainable artificial intelligence methods: Shap and lime,” *ArXiv*, vol. abs/2305.02012, 2023. [Online]. Available: <https://api.semanticscholar.org/CorpusID:282894857>

B. Aldughayfiq, F. Ashfaq, N. Z. Jhanjhi, and M. Humayun, “Explainable ai for retinoblastoma diagnosis: Interpreting deep learning models with lime and shap,” *Diagnostics*, vol. 13, no. 11, 2023. [Online]. Available: <https://www.mdpi.com/2075-4418/13/11/1932>

X. Zeng and K. Zhu, “Enhancing the interpretability of shap values using large language models,” 2026. [Online]. Available: <https://arxiv.org/abs/2409.00079>

Z. Cui, M. Li, Y. Huang, Y. Wang, and H. Chen, “An interpretation framework for autonomous vehicles decision-making via shap and rf,” *2022 6th CAA International Conference on Vehicular Control and Intelligence (CVCI)*, pp. 1–7, 2022. [Online]. Available: <https://api.semanticscholar.org/CorpusID:254459278>

//api.semanticscholar.org/CorpusID:254459278

H. A. Tahir, W. Alayed, W. U. Hassan, and A. Haider, “A novel hybrid xai solution for autonomous vehicles: Real-time interpretability through lime–shap integration,” *Sensors*, vol. 24, no. 21, 2024. [Online].

Available: <https://www.mdpi.com/1424-8220/24/21/6776>

A. Chattopadhyay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, “Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks,” *CoRR*, vol. abs/1710.11063, 2017. [Online].

Available: <http://arxiv.org/abs/1710.11063>

P.-T. Jiang, C.-B. Zhang, Q. Hou, M.-M. Cheng, and Y. Wei, “Layercam: Exploring hierarchical class activation maps for localization,” *IEEE Transactions on Image Processing*, vol. 30, pp. 5875–5888, 2021.

R. L. Draelos and L. Carin, "Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks," 2021. [Online].

Available: <https://arxiv.org/abs/2011.08891>

M. B. Muhammad and M. Yeasin, "Eigen-cam: Class activation map using principal components," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–7.

D. Gunning, E. Vorm, J. Y. Wang, and M. Turek, "Darpa's explainable ai (xai) program: A retrospective," *Applied AI Letters*, vol. 2, no. 4, p. e61, 2021. [Online]. Available: <https://onlinelibrary.wiley.com/doi/abs/10.1002/ail2.61>

I. Urbieto, M. Nieto, M. Garc'ia, and O. Otaegui, "Design and implementation of an ontology for semantic labeling and testing: Automotive global ontology (ago)," *Applied Sciences*, vol. 11, no. 17,

2021. [Online]. Available: <https://www.mdpi.com/2076-3417/11/17/7782>

D. Yu, B. Yang, D. Liu, and H. Wang, "A survey on neural-symbolic systems," *CoRR*, vol. abs/2111.08164, 2021. [Online]. Available: <https://arxiv.org/abs/2111.08164>

R. Confalonieri, T. Weyde, T. R. Besold, and F. Moscoso del Prado Mart'ın, "Using ontologies to enhance human understandability of global post-hoc explanations of black-box models," *Artificial Intelligence*, vol. 296, p. 103471, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0004370221000229>

<https://www.sciencedirect.com/science/article/pii/S0004370221000229>

J. Amara, S. Samuel, and B. Ko'nig-Ries, "Integrating domain knowledge for enhanced concept model explainability in plant disease classification," in *Extended Semantic Web Conference*, 2024. [Online].

Available: <https://api.semanticscholar.org/CorpusID:269758257>

B. Igoche, O. Matthew, P. Bednar, and A. Gegov, "Integrating structural causal model ontologies with lime for fair machine learning explanations in educational admissions," *Journal of Computing Theories and Applications*, vol. 2, pp. 65–85, 06 2024.

S. Sharma, N. Rakesh, M. Jagga, H. Singh, M. Kirola, R. Subbiah,

N. Varshney, L. Hussein, and N. I. Vatin, "An ontological-based explainable intrusion detection system," *International Journal of Information Technology*, 2025. [Online]. Available: <https://api.semanticscholar.org/CorpusID:280605079>

M. Bellucci, N. Delestre, N. Malandain, and C. Zanni-Merk, "Towards counterfactual explanations for ontologies," *Semantic Web*, vol. 15, pp. 1–26, 03 2024.

R. C. Jackson, J. P. Balhoff, E. Douglass, N. L. Harris, C. J. Mungall, and

J. A. Overton, "ROBOT: A Tool for Automating Ontology Workflows," *BMC Bioinformatics*, vol. 20, no. 1, p. 407, 2019.

C. O. Retzlaff, A. Angerschmid, A. Saranti, D. Schneeberger,

R. Röttger, H. Müller, and A. Holzinger, "Post-hoc vs ante-hoc explanations: xai design guidelines for data scientists," *Cognitive Systems Research*, vol. 86, p. 101243, 2024. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1389041724000378>

Lernia, "AI Interpretability with LIME and SHAP: A Practical Guide," <https://learn-prompting.fr/blog/ai-interpretability-lime-shap>, 2026, accessed: 2026-05-01.