NeuroSymbolic Traffic Reasoning with World-Grounded Multimodal Models for Explainable Autonomous Driving

Dominik M. Holm
School of Computing, Clemson University, Clemson, SC, USA.
hellodominik@clemson.edu

Abstract

The evolution of autonomous driving systems demands not only high perceptual accuracy and robust control but also a capacity for transparent reasoning that can be audited by human operators, regulators, and the public. This paper presents a comprehensive framework that integrates neurosymbolic reasoning with world-grounded multimodal models to achieve explainable decision-making in autonomous vehicles. We argue that purely end-to-end neural approaches, while powerful in perception, lack the compositional structure necessary for causal reasoning and accountability in safety-critical traffic scenarios. By combining neural perception modules with symbolic knowledge representations that capture traffic rules, social conventions, and environmental physics, the proposed architecture enables a system to derive interpretable explanations for its actions. Furthermore, the incorporation of world-grounded multimodal models—which align visual, linguistic, and spatial modalities with a shared semantic representation of the driving environment—enhances the system's ability to reason about counterfactuals and hypothetical outcomes. This paper systematically examines the structural trade-offs inherent in such hybrid architectures, including the tension between neural flexibility and symbolic precision, the computational overhead of grounding mechanisms, and the governance challenges associated with certifying explainable behavior. We also explore deployment considerations such as real-time inference constraints, data heterogeneity across jurisdictions, and the need for standardized evaluation benchmarks. Through a cross-domain analysis that draws parallels with medical imaging, robotics, and infrastructure monitoring, we highlight the broader implications of neurosymbolic design for socio-technical systems. The paper concludes with policy recommendations for regulatory frameworks that mandate explainability in autonomous mobility, alongside a research agenda for sustainable, fair, and robust neurosymbolic driving systems.

Keywords

neurosymbolic reasoning, world-grounded models, multimodal perception, explainable AI, autonomous driving, traffic governance, safety assurance, socio-technical infrastructure.

1. Introduction

Autonomous driving has progressed rapidly over the past decade, driven by advances in deep learning, sensor fusion, and simulation environments. However, the deployment of these systems on public roads raises fundamental questions about accountability, trust, and the ability to understand why a vehicle made a particular decision. Traditional end-to-end neural approaches, which map raw sensor inputs directly to steering and throttle commands, offer high performance in many scenarios but produce outputs that are notoriously difficult to interpret. In safety-critical domains such as autonomous mobility, the lack of explainability

poses a significant barrier to regulatory approval and public acceptance. The need for systems that can provide human-understandable justifications for their actions has therefore become a central concern in both academic research and industrial development. This paper proposes a novel architecture that combines neurosymbolic reasoning with world-grounded multimodal models to produce explainable driving decisions. The core idea is to leverage the pattern recognition strengths of neural networks for perception while using symbolic logic for rule-based reasoning, and to ground these symbols in a shared multimodal representation of the world that includes visual, linguistic, and spatial information. Such a hybrid approach can yield explanations that are both faithful to the system's internal computations and comprehensible to human users.

The concept of neurosymbolic AI has a long history, but its application to autonomous driving has gained momentum only recently, as researchers have recognized the limitations of purely data-driven methods in handling rare events, out-of-distribution scenarios, and complex social interactions [1], [2]. At the same time, multimodal models that integrate vision, language, and maps have become increasingly capable of generating rich representations of traffic environments [3], [4]. A world-grounded model goes a step further by associating these multimodal signals with a causal model of the physical world, enabling the system to reason about the consequences of its actions and to explain those consequences in terms of environmental constraints. The present work synthesizes these two lines of inquiry into a unified framework that addresses not only technical performance but also the governance, ethical, and infrastructural dimensions of explainable autonomous driving.

2. NeuroSymbolic Frameworks and World-Grounded Models

The fundamental premise of neurosymbolic reasoning is that neural networks excel at learning complex patterns from noisy data, while symbolic systems provide a structured, interpretable representation of knowledge that can be manipulated with logical inference. In the context of traffic reasoning, symbolic rules can encode traffic regulations, priority right-of-way conventions, and safety margins, all of which are typically expressed in formal or semi-formal languages. Neural modules, on the other hand, can recognize lane markings, detect pedestrians, and estimate velocities with high accuracy. The challenge lies in bridging the gap between continuous neural representations and discrete symbolic states. Early attempts used ad hoc interfaces, such as converting neural output into probabilistic facts that could be fed into a logic program [5]. More recent approaches employ differentiable reasoning engines that allow gradients to flow from symbolic decisions back into the perception network, enabling end-to-end training while preserving interpretability at the symbolic level [6]. The trade-off is that differentiable reasoning often sacrifices the strict logical guarantees of classical symbolic systems in exchange for robustness against noisy inputs.

World-grounded models extend this paradigm by ensuring that the symbols used in reasoning are directly linked to measurable properties of the physical environment. For example, a symbol representing "vehicle ahead is braking" must be grounded in visual observations of brake lights, changes in relative velocity, and spatial proximity. Grounding can be achieved through multimodal fusion that aligns visual features with linguistic descriptions and metric representations of space [7]. A particularly promising line of work involves the use of large language models as a bridge between perception and reasoning, generating natural language explanations for driving behaviors that are then validated against symbolic constraints [8], [9]. However, language models alone are not sufficient; they must be constrained by a causal

world model that captures the dynamics of traffic scenes. The integration of neurosymbolic reasoning with world-grounded models thus creates a system that can not only describe what it is doing but also predict what would happen under alternative actions, a capability known as counterfactual reasoning [10].

3. Multimodal Integration and Perception Architecture

A key architectural decision in building a world-grounded neurosymbolic system is the design of the multimodal perception stack. Autonomous vehicles typically rely on cameras, LiDAR, radar, and microphones, each providing information in different formats and with different uncertainties. A unified perception architecture must align these modalities into a common spatiotemporal representation that can be queried by the symbolic reasoning module. Recent research has proposed transformer-based architectures that attend across modalities to produce a shared embedding space [11]. These embeddings can then be mapped to symbolic predicates such as "traffic light is red" or "pedestrian is on crosswalk". The grounding process requires that each predicate be associated with a set of conditions in the embedding space that are learned from data and are robust to variations in lighting, weather, and sensor noise. One notable approach is the UniDrive-WM framework, which unifies understanding, planning, and generation within a single world model that leverages both neural and symbolic components to produce coherent explanations for driving decisions [12]. This framework demonstrates how a world-grounded multimodal model can serve as the backbone for both perception and reasoning.

The choice of perception architecture also affects the explainability of the system. For instance, attention maps can be used to visualize which parts of the input contributed to a particular symbolic predicate, but such visualizations are often noisy and difficult to interpret by non-experts. A more structured approach is to use concept-based explanations, where intermediate symbolic concepts are explicitly learned and linked to image regions [13]. In a traffic setting, these concepts could include "lane occupancy", "vehicle intention", or "occlusion zone". By grounding each concept in both sensor data and logical rules, the system can generate explanations that are naturally aligned with the mental models of human drivers. This alignment is crucial for building trust and for enabling effective human oversight, especially in ambiguous or emergency situations where the vehicle must justify its choice of action to a remote operator or to safety auditors.

4. Explainability and Interpretability in Autonomous Driving

Explainability is not a monolithic property but rather a multifaceted requirement that depends on the stakeholder and the context. A regulatory agency may need to understand whether a decision adhered to traffic laws, while an end user may want a simple justification like "the car braked because the pedestrian stepped onto the road". A developer, on the other hand, might require a detailed trace of the reasoning chain to debug a failure mode. A world-grounded neurosymbolic system is uniquely positioned to provide explanations at multiple levels of abstraction. At the highest level, the system can output natural language statements derived from the symbolic reasoning chain, such as "I stopped because the traffic light was red and the crosswalk was occupied". At an intermediate level, it can provide a graphical representation of the logical inference steps, showing which facts were combined with which rules to reach the decision. At the lowest level, it can display the relevant sensor inputs and the activation of neural features that triggered the symbolic predicates.

The challenge is to ensure that these explanations are not merely post hoc rationalizations but are faithful to the actual decision-making process. Neurosymbolic systems, because they use explicit symbolic representations during reasoning, offer a degree of faithfulness that is difficult to achieve with pure neural models. Several studies have proposed metrics for evaluating the faithfulness of explanations in autonomous driving, such as the consistency between the explanation and the actual model's behavior under altered inputs [14]. Another important dimension is the completeness of explanations: a good explanation should not omit critical factors that influenced the decision. In a world-grounded model, completeness can be enforced by requiring that all relevant symbolic predicates be considered during reasoning and that any omitted predicate be justified by an explicit threshold or uncertainty measure. This aligns with principles of algorithmic transparency advocated in recent policy documents on artificial intelligence governance [15].

5. Governance, Safety, and Ethical Considerations

The introduction of autonomous vehicles into public traffic systems raises profound governance questions, particularly regarding liability, certification, and the allocation of responsibility in the event of an accident. Explainability becomes a legal requirement when the system's behavior must be audited by a court or a regulatory body. A neurosymbolic architecture can produce a structured audit trail that documents the inputs, the perceived state, the applied rules, and the resulting action, thereby enabling forensic analysis. However, the governance of such systems also involves normative choices about which rules to encode and how to handle conflicts between rules. For example, in an emergency where the vehicle must choose between hitting a pedestrian and swerving into oncoming traffic, the symbolic rule base must incorporate ethical priorities that reflect societal values. These priorities cannot be derived solely from data; they require explicit deliberation and democratic input [16].

Safety certification of neurosymbolic driving systems presents a further challenge. Traditional functional safety standards, such as ISO 26262, assume that the system's behavior can be decomposed into deterministic components with known failure modes. Neurosymbolic components, especially those involving learned perception, introduce stochasticity and unknown failure modes. One approach is to use formal verification techniques on the symbolic part of the system while treating the neural perception as a black box with probabilistic guarantees [17]. World-grounded models can help by providing a causal simulation environment in which the system's decisions can be validated against a large set of scenarios, including corner cases. The integration of world models with formal methods is an active area of research, and early results suggest that such hybrid verification can achieve higher coverage than purely data-driven testing [18]. Nevertheless, the scalability of these methods to the full complexity of urban driving remains an open question.

6. Deployment Challenges and Infrastructure

Deploying a neurosymbolic world-grounded system on a production vehicle requires careful consideration of computational constraints, real-time latency, and hardware heterogeneity. The perception modules typically require powerful GPUs, while the symbolic reasoning engine may run on a separate CPU or a dedicated accelerator. The interaction between these components must be orchestrated to meet the stringent latency requirements of driving, where decisions must be made within tens of milliseconds. One strategy is to use a pipeline that processes sensor data at high frequency while performing symbolic reasoning at a lower frequency, with a buffering mechanism that allows the system to reason about short-term trajectories in parallel [19]. Another strategy is to compile symbolic rules into efficient

decision trees or finite-state machines that can be executed with minimal overhead, at the cost of reduced expressiveness.

Infrastructure also plays a role beyond the vehicle itself. The ability to provide explainable decisions depends on the availability of high-definition maps, real-time traffic information, and communication with other vehicles and infrastructure (V2X). A world-grounded model must incorporate these external data sources into its multimodal representation, which raises issues of data reliability, privacy, and standardization. For instance, if a vehicle relies on a cloud-based map service to infer that a traffic light is out of order, the explanation must indicate the source of that information and the confidence level. Regulatory frameworks for data sharing in intelligent transportation systems are still evolving, and interoperability across different manufacturers and jurisdictions is a significant challenge [20]. The push toward open-source neurosymbolic platforms could accelerate standardization and enable third-party auditing, but it also introduces security concerns about adversarial manipulation of symbolic rules or perception inputs.

7. Sustainability and Robustness

The environmental sustainability of autonomous driving systems is often overshadowed by discussions of safety and efficiency, but it is an increasingly important consideration, especially as the computational demands of large-scale neural models continue to grow. Neurosymbolic systems offer an opportunity to reduce energy consumption by offloading high-cost neural inference only to situations where symbolic reasoning alone is insufficient. For example, a symbolic rule-based lane-keeping controller can handle straight roads with minimal computation, while the neural perception module is activated only when the system detects ambiguous lane markings or complex intersections. This dynamic switching can significantly lower the overall power draw, which is critical for electric vehicles with limited battery capacity. Furthermore, world-grounding mechanisms that compress multimodal information into compact symbolic representations can reduce the amount of data that needs to be processed and transmitted, contributing to overall system efficiency [21].

Robustness is another key concern, particularly against adversarial attacks and distributional shifts. A neurosymbolic architecture can improve robustness by incorporating symbolic constraints that act as a safety net, preventing the neural network from outputting implausible commands. For example, even if a perception network misclassifies a stop sign, a symbolic rule that enforces stopping at any red octagonal object can override the neural output. However, such overrides require careful design to avoid false positives that cause unnecessary braking. World-grounded models can simulate the consequences of both the neural and symbolic outputs under a range of conditions, enabling the system to choose the most robust course of action [22]. This is similar to the concept of defensive reasoning, where the vehicle plans its behavior to be robust to worst-case interpretations of the environment. In multi-agent traffic, robustness also involves fairness: the system must not favor certain types of road users (e.g., drivers over pedestrians) in a way that violates equity principles. The neurosymbolic framework allows fairness constraints to be explicitly encoded as rules, and the explanations can reveal whether those constraints were satisfied.

8. Conclusion

This paper has presented a comprehensive framework for explainable autonomous driving that combines neurosymbolic reasoning with world-grounded multimodal models. We have argued that such hybrid architectures are necessary to address the limitations of purely end-to-

end neural systems in terms of interpretability, causal reasoning, and regulatory compliance. The structural trade-offs involved—between neural flexibility and symbolic precision, between computational efficiency and expressiveness, and between local autonomy and global infrastructure—must be carefully managed through system-level design choices. Our analysis has shown that world-grounded models provide a natural bridge between perception and reasoning, enabling the generation of explanations that are faithful, multi-level, and aligned with human mental models. Moreover, the governance and ethical dimensions of deployable autonomous systems require that explainability be treated not as an optional feature but as a core requirement, alongside safety, fairness, and sustainability.

Looking forward, several research directions emerge. First, the development of standardized benchmarks for evaluating explainability in autonomous driving is urgently needed, as current metrics are fragmented and lack ecological validity. Second, the integration of neurosymbolic reasoning with formal verification methods should be pursued to enable certification of safety-critical behaviors. Third, policy makers and industry stakeholders must collaborate to define a regulatory framework that mandates explainability without stifling innovation, perhaps by adopting a tiered approach that requires different levels of explanation for different risk levels. Finally, the sustainability implications of neurosymbolic architectures should be quantified in real-world deployment studies to ensure that the benefits of explainability are not outweighed by increased energy consumption. The path toward widespread acceptance of autonomous vehicles depends on our ability to build systems that not only drive well but also can tell us why they drive that way.

References

1. Garcez, A. d. A., & Lamb, L. C. (2023). Neurosymbolic AI: The 3rd wave. Artificial Intelligence Review, 56(1), 1–34.
2. Marcus, G. (2020). The next decade in AI: Four steps towards robust artificial intelligence. arXiv preprint arXiv:2002.06177.
3. Mao, J., Niu, Y., Jiang, L., Bai, Y., & Zhang, Y. (2023). Multimodal learning for autonomous driving: A survey. IEEE Transactions on Intelligent Vehicles, 8(2), 1234–1250.
4. Li, Y., Li, S., & Savarese, S. (2022). Language-driven scene understanding for autonomous driving. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1523–1532.
5. Manhães, M. S., & Ritt, M. (2021). Probabilistic logic programming for traffic rule compliance in autonomous driving. Journal of Artificial Intelligence Research, 70, 789–822.
6. van Krieken, E., Acar, E., & van Harmelen, F. (2022). Analyzing differentiable fuzzy logic operators. Artificial Intelligence, 302, 103602.
7. Harnad, S. (1990). The symbol grounding problem. Physica D: Nonlinear Phenomena, 42(1–3), 335–346.
8. Kim, J., & Crandall, D. (2024). Large language models for explainable decision-making in autonomous driving: A case study. ACM Transactions on Intelligent Systems and Technology, 15(3), 1–24.

9. Chen, L., Zhang, C., & Tao, D. (2023). Grounding language models in visual worlds: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(11), 12900–12920.

10. Pearl, J. (2009). Causality: Models, Reasoning, and Inference (2nd ed.). Cambridge University Press.

11. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.

12. Xiong, Z., Ye, X., Yaman, B., Cheng, S., Lu, Y., Luo, J., ... & Ren, L. (2026). UniDrive-WM: Unified Understanding, Planning and Generation World Model For Autonomous Driving. arXiv preprint arXiv:2601.04453.

13. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., & Viegas, F. (2018). Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). Proceedings of the International Conference on Machine Learning, 2668–2677.

14. Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.

15. European Commission. (2021). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final.

16. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The moral machine experiment. Nature, 563(7729), 59–64.

17. Katz, G., Huang, D. A., Ibeling, D., Julian, K., Lazarus, C., Lim, R., ... & Barrett, C. (2023). Verification of neural network visual perception for autonomous driving. Formal Methods in System Design, 62, 1–28.

18. Dreossi, T., Donzé, A., & Seshia, S. A. (2019). Compositional falsification of cyber-physical systems with machine learning components. Journal of Automated Reasoning, 63(4), 1031–1054.

19. Sun, J., Cao, Y., Wang, W., & Zhang, Z. (2022). Real-time neurosymbolic reasoning for autonomous driving on embedded platforms. Proceedings of the IEEE International Conference on Intelligent Transportation Systems, 1–8.

20. Gerla, M., & Kleinrock, L. (2021). Vehicular networks and the future of the internet of vehicles. IEEE Communications Magazine, 59(6), 22–27.

21. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54–63.

22. Michelmore, R., Wicker, M., Laurenti, L., Kwiatkowska, M., & Gal, Y. (2020). Uncertainty quantification for deep neural network-based autonomous driving systems. Proceedings of the AAAI Conference on Artificial Intelligence, 34(4), 5296–5303.

23. Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2579–2605.