




Glass-Box Automated Driving: Insights and Future Trends

Mauro Bellone¹ ^a, Raivo Sell² ^b and Ralf-Martin Soe¹ ^c

¹*FinEst Centre for Smart Cities, Tallinn University of Technology, Estonia*

²*Department of Mechanical and Industrial Engineering, Tallinn University of Technology, Estonia*

Keywords: Glass-Box Models, Automated Driving, Performance vs Interpretability Trade-off.

Abstract: Automated driving has advanced significantly through the use of black-box AI models, particularly in perception tasks. However, as these models have grown, concerns over the loss of explainability and interpretability have emerged, prompting a demand for creating 'glass-box' models. Glass-box models in automated driving aim to design AI systems that are transparent, interpretable, and explainable. While such models are essential for understanding how machines operate, achieving perfect transparency in complex systems like autonomous driving may not be entirely practicable nor feasible. This paper explores arguments on both sides, suggesting a shift of the focus towards balancing interpretability and performance rather than considering them as conflicting concepts.

1 INTRODUCTION

Fully automated driving on roads has been a long-sought goal, with little significant progress for many years (Stanton and Young, 1998). The field has only recently advanced notably, driven by improvements in hardware computational capabilities and data-driven models with the promise of end-to-end automated driving in the near future (Yurtsever et al., 2020). However, this progress comes at the cost of losing a clear connection to the fundamentals of process control (Omeiza et al., 2021). Automated driving on roads exemplifies the control of a complex system where data-driven models offer advantages, as creating detailed analytical models for every component is nearly infeasible.

Classical optimal control theory aims to design control processes that achieve optimal tracking of a desired reference signal, offering elegant analytical solutions to many problems. The theory begins with linear, scalar systems and extends to multivariate, non-linear processes, significantly increasing computational complexity and the system's level of abstraction, particularly due to the challenges of solving higher-dimensional and non-linear models.


From this perspective, a model that analytically captures reality at its fundamental level would be the ideal foundation for building deterministic, error-


proof control systems. By definition, such a system would qualify as a glass-box model, offering complete interpretability and explainability. On the other hand, many fundamental physical processes remain poorly understood, and while the human instinct is to seek clear explanations, technological advancement often relies on approximate descriptions to manage complex control systems in an uncertain real world. A deeper understanding leads to improved functionality, highlighting the importance of interpretability. While black-box models may enable control systems to function effectively, only full analytical interpretability can unlock their true potential.


Following the example of automated driving, a full analytical and computational description of dynamic behavior—such as aerodynamics, tire friction, and engine response to driver commands—is complex and not entirely practical yet achievable. Steering robots capable of driving vehicles in structured environments, such as test tracks, have existed since the 1980s (Weisser et al., 1999). The real challenge, however, lies in real-world interactions, where modeling the unpredictable behavior of other road users becomes increasingly difficult. In such scenarios, users expect a robotic driver to act deterministically and make safety-critical decisions within fractions of a second.

Following this line of research, this work discusses the following research questions while focusing on the problem of automated driving:

RQ1. What are the practical motivations for building glass-box analytical models?

^a  <https://orcid.org/0000-0003-3692-0688>

^b  <https://orcid.org/0000-0003-1409-0206>

^c  <https://orcid.org/0000-0002-6782-1677>

RQ2. What are the challenges in achieving full explainability and interpretability in the field of automated driving?

RQ3. What viable strategies exist for achieving transparency in automated driving systems?

This paper is structured as follows: Section 2 provides the motivation for applying glass-box models and presents our perspective on RQ1. Section 3 explores the challenges of developing such models in the field of automated driving, addressing RQ2. Finally, our proposed strategy for RQ3, which focuses on modularizing models and balancing interpretability with performance, is discussed in Section 4.

2 MOTIVATION

Let's assume we want to design a system capable of taking the best action in every situation, following a data-driven approach. For simplicity, we can envision it as a perfect driver: one that always selects the optimal route from A to B while ensuring energy efficiency, minimal travel time, safety, and passenger comfort.

The process of building such a system would involve recording pairs of information points and optimal actions. Each information point can be considered as an abstract representation of everything the robotic driver perceives, including the internal vehicle status and any external sources of information (e.g., driving scenario, traffic conditions, other road users, etc.). For simplicity Fig. 1 provides an abstract representation of these information points on an action-information Cartesian map, where i represents our information content and d represents our best recorded decision in such a scenario. By considering all recorded information, represented as gray dots, one can construct a function—using any analytical fitting method—thus obtaining an action-information mapping function.

The goal of building such a function is to use the action-information model as a decision-making tool to estimate the best action as new information becomes available. If we assume that the blue line in Fig. 1 represents our best model, when new information i_{new} arrives, the model's guess should be no other than \hat{d}_{new} as the intersection point between the information and the model. However, the optimal decision in that situation may be something slightly different, such as d_{new}^* , resulting in a gap between the model's guess and the true optimal decision.

Such an gap in the optimality might result from:

1. *Measurement errors* in both the information used to build the model and the new data point.
2. *Approximation errors* due to complex interactions between variables not fully captured by the model.
3. *Truly unknown circumstances*, highlighting a difference between the current information point and previous knowledge of the system.
4. *A change in the operational domain*, where the model was not trained or calibrated for new conditions, leading to discrepancies in the decision-making process.

Sticking to the example of automated driving, the gap in the optimal decision can lead to a wide range of consequences, from minor delays in travel time or passenger discomfort to more severe outcomes like crashes. How can one identify the cause of a miscalculation in the decision-making process if the system is a fully enclosed black box?

From this perspective, all points mentioned earlier—measurement errors, approximation errors, truly unknown circumstances, and changes in the operational domain—are equally valid sources of error with very little possible action to address and correct such errors to ensure the system functions properly across different scenarios. The current black-box approach to solving this problem tends to add information to the system indefinitely, leading to memorization rather than meaningful knowledge abstraction with minimal validation and verification possibilities (Pikner et al., 2024).

Assuming that our action-information model is an analytical multivariate function underpinning our process, a gap in decision-making can be investigated, studied, and potentially debugged if, and only if, the system can be fully explained and interpreted.

Often, interpretability and explainability are seen as a written form of text providing an explanatory

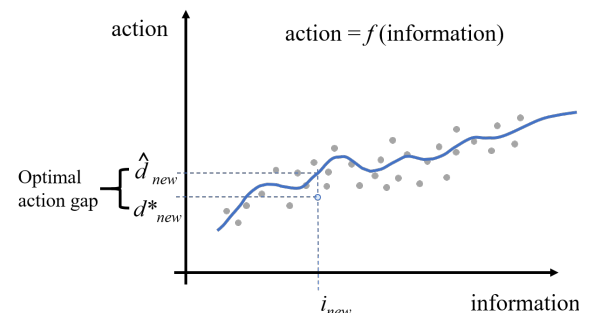


Figure 1: Action - information model depiction. Gray dots constitute information points used to generate the fitting function in blue. The optimality gap arises from the difference between the optimal decision d_{new}^* and the model guess \hat{d}_{new} .

statement about driving decisions and specific actions (Omeiza et al., 2022). However, this approach should not be conflated with true interpretability or accountability. The generated text, often produced by black-box LLMs (large language models), merely describes the situation without offering useful debugging information. As LLMs have shown, it is always possible to generate ‘human-understandable’ explanations that can be terribly misleading and fail to reflect the system’s actual behavior at its fundamental level. Thus, this type of verbal explanation often falls more into the realm of psychological perspective rather than a genuine interpretation of system behavior.

3 CHALLENGES

Without any doubt, full black-box models have demonstrated compelling capabilities in controlling complex systems such as autonomous vehicles (Chen et al., 2024). This achievement cannot be neglected, as it represents a significant opportunity to address long-term practical problems. However, glass-box models offer the opportunity to debug and refine these systems by providing full control and transparency. This allows for a deeper understanding and improvement of each system’s behavior (Kuznietsov et al., 2024).

The first significant challenge in building a glass-box control system for autonomous driving lies in managing the complexity of the environment. As previously mentioned, the challenge does not stem from driving along a predefined path at a precise velocity but rather from interacting effectively with the surrounding environment. Autonomous driving requires processing vast amounts of real-time data from sensors such as cameras, LiDAR, radar, and GPS. The decision-making process entails intricate interactions between perception, prediction, and control systems, which are often modeled using complex deep learning or neural networks—methods that are inherently difficult to interpret. Figure 2 illustrates the control flow from low-level to high-level driving models. The concept is that any state-of-the-art controller, from PID (Emirler et al., 2014) to Lyapunov-based controllers (Alcala et al., 2018) (Karafyllis et al., 2022), can effectively and precisely manage vehicle speed and steering angle, provided a simplified physical model of the vehicle is known. The assumptions required for these controllers to function are often unrealistic and ineffective at predicting dynamic environmental changes. The higher the complexity of the vehicle model, the better and more effective the controller. This capability is sufficient for driving in

structured environments. The system’s internal components, highlighted with a dashed line in Fig. 2, can be fully explainable and interpretable PIDs. Literature offers a robust foundation of analytical analyses and formal solutions for the problem of following reference signals, enabling deterministic control in these scenarios (Fleming and Rishel, 2012).

As the research community approached this problem, it quickly became evident that addressing it required multiple levels of abstraction. At the very low level, classical automatic controllers, such as PID or model predictive controllers, perform their tasks effectively. However, tackling vehicle perception and routing with the same level of detail as low-level control, represented in the outer cycle of Fig. 2, is computationally and practically infeasible.

The routing problem itself can be subdivided into two distinct components: the generation of kinematically or dynamically feasible trajectories for short-range vehicle control, and waypoint generation for high-level vehicle routing. The former requires solving differential equations, which are computationally demanding and often limited to localized regions. The latter typically employs simpler, non-physically compliant algorithms, such as Dijkstra’s algorithm or rapidly-exploring random trees (RRT) (LaValle, 2006). Attempting to solve differential equations for every point along a long path is not only computationally prohibitive but also ineffective, as it assumes a static environment, thereby reducing the system’s ability to adapt to dynamic changes. Conversely, ignoring differential equations altogether can result in unfeasible paths that violate the vehicle’s physical constraints.

The routing problem exemplifies the necessity for different levels of abstraction to address complex challenges in autonomous driving. Each level retains its own interpretability: long-range planning may sacrifice physical precision but provides computational efficiency, while short-range trajectory generation maintains a detailed physical interpretation. This balance allows the system to adapt to a dynamic environment while ensuring feasibility at a local level.

On a different level, the perception problem, which includes tasks such as object detection and segmentation, is predominantly addressed using black-box models (Huang et al., 2022). On one hand, these models excel in providing detection capabilities that can even surpass human performance in certain scenarios, such as low-light conditions or when integrating data from multiple sensor sources. An illustrative example is shown in the two images in Fig. 2 which are extracted from the IseAuto dataset (Gu et al., 2023) and include the addition of a black-box seg-

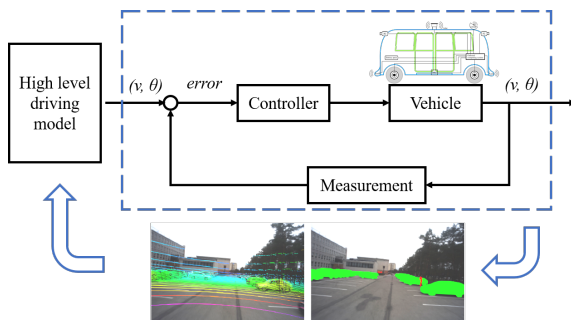


Figure 2: Levels of control abstractions featuring low-level process control and high-level intelligent functions. Visual images from IseAuto dataset (Gu et al., 2023).

mentation result (Gu et al., 2024).

On the other hand, they lack an understanding of the semantic meaning of objects within a scene, treating a tree and a person merely as objects without recognizing their distinct roles or contextual importance. Contextual importance is inherently challenging to capture, as even human observers may assign varying interpretations and significance to similar objects in a scene. An interpretable and explainable model should not aim to produce verbose descriptions of a scene but rather to classify objects accurately while accounting for their high-level roles and importance. For instance, in such a framework, errors or flaws in detecting vehicles or pedestrians, compared to non-critical objects like rocks or foliage, could be identified, debugged, and addressed in a targeted manner.

3.1 Safety-Critical Nature

While it may seem that autonomous vehicles must prioritize split-second decision-making over immediate explanations, it is precisely in these high-stakes scenarios that the effort to build simple, classical control, and rule-based systems becomes essential. Transparency is critical for diagnostics and regulatory compliance, and it is often mischaracterized as compromising the vehicle's ability to respond effectively in emergencies. On the contrary, such transparency ensures that misunderstandings or misinterpretations are minimized, particularly in situations where they cannot be tolerated. In (Abrecht et al., 2024), the authors clearly emphasize the safety concerns that deep learning poses to automated driving, covering aspects such as operational domain definition and limitations, as well as the methods used for data preparation and algorithm development. Moreover, regulations are placing increasing emphasis on the importance of AI explainability in safety-critical industries like transportation. Glass-box models are required to comply with stringent industry standards, such as ISO 26262

(ISO, 2011) (functional safety) and ISO/PAS 21448 (Safety of the Intended Functionality - SOTIF), to ensure reliability, accountability, and safety (Kirovskii and Gorelov, 2019). Additionally, frameworks such as the European Union's Ethics Guidelines for Trustworthy AI and the U.S. NIST's initiatives on AI explainability advocate for more transparent and accountable AI systems. The recently adopted AI Act (*Regulation (EU) 2024/1689 laying down harmonized rules on artificial intelligence*) sets clear requirements and obligations for AI developers, emphasizing explainability and accountability in AI-based systems, including automated driving functionalities.

4 STRATEGIES AND FUTURE TREND

Several strategies exist to achieve transparency, accountability, and explainability in automated driving, many of which can be applied to other complex systems that benefit from black-box models. While a fully glass-box model for automated driving may be impractical in its purest form—particularly in complex real-world scenarios—it remains an ambitious long-term goal. In the interim, hybrid models represent a practical solution, leveraging the strengths of both interpretable and black-box approaches. A promising strategy is to combine interpretable models for high-level decision-making with black-box models for perception tasks, striking a balance between interpretability and efficacy with the goal of reducing black-box models to the minimum level. For example, rule-based logic can be applied to lane-change policies (Malayjerdi et al., 2022), providing clear and explainable decision-making, while deep learning models handle object detection, which often benefits from the data-driven adaptability of black-box approaches.

The application of simpler, domain-specific models also offers notable strengths in achieving transparency and reliability. In constrained environments, such as autonomous shuttles or last-mile delivery vehicles, rule-based systems or simpler machine learning models can effectively balance explainability with functionality. This approach aligns with the concept of maintaining a human-in-the-loop framework, allowing human operators to oversee decisions and intervene when necessary. This ensures that actions taken by the system adhere to ethical and safety considerations. Moreover, adopting a modular system design—breaking the driving stack into smaller, interpretable modules—further aids in achieving transparency. For example:

- *Perception*: Explains how objects are detected and classified.
- *Planning*: Provides insights into why a particular trajectory or decision was selected.
- *Control*: Demonstrates how the vehicle executes the planned commands.

These modular explanations help ensure that the system remains interpretable and debuggable, while still benefiting from advancements in AI and automation. Even in such scenarios, designers must rely on extensive simulation environments and formal verification methods to understand and validate system behaviors under diverse conditions.

Autonomous systems operate in the physical world, a domain governed by the principles of physics (e.g., Maxwell's equations, Newton's laws). These principles provide a robust foundation for defining a governing framework that can enhance both explainability and interpretability. A critical question emerges:

How might one leverage the fundamental properties of physics to build a validation governor around AI-based autonomy systems?

This question remains an open avenue for research, challenging the community to explore innovative ways to integrate physical laws into the validation and explainability frameworks of autonomous systems (Pikner et al., 2024).

The application of Post-Hoc Explainability such as saliency maps, SHAP (Shapley Additive Explanations) (Lundberg and Lee, 2017), (Lundberg et al., 2020), or LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016) to analyze and explain decisions made by black-box models constitute also emergent solutions. One might note that Post-hoc methods do not require the model to be inherently interpretable but rather attempt to interpret and explain the outputs of a AI model after it has been trained, which makes them practical for use with complex models that prioritize performance but sacrifice interpretability.

5 CONCLUSION

A fully glass-box model for automated driving is likely impractical in its purest form, particularly in complex real-world scenarios as highly interpretable models (e.g., decision trees or rule-based systems) may struggle to capture the nuanced decision-making

required in dynamic driving environments. However, hybrid approaches that blend interpretability with black-box models offer a practical way forward while keeping the aspiration for full glass-box designs alive. As is often the case in engineering, the optimal solution lies in the middle ground—making the critical components of the system transparent enough to ensure safety, reliability, and accountability, without compromising performance.

The drive to pursue glass-box models is deeply rooted in the fundamental curiosity of engineers. The question, 'Why does a system work the way it does?' forms the very basis of scientific exploration. Without such curiosity, humanity might still accept a flat Earth as truth, never challenging experiences that seem counterintuitive or seeking alternative perspectives. The essence of progress lies in questioning and reframing our understanding of the world and its complexities.

ACKNOWLEDGMENT

Part of this research has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement No. 856602 (Finest Twins) and from the European Union's Horizon Europe Research and Innovation Programme under grant agreement No. 101135988 (PLIADDES: AI-Enabled Data Lifecycles Optimization and Data Spaces Integration for Increased Efficiency and Interoperability)

REFERENCES

- Abrecht, S., Hirsch, A., Raafatnia, S., and Woehrle, M. (2024). Deep learning safety concerns in automated driving perception. *IEEE Transactions on Intelligent Vehicles*, pages 1–12.
- Alcala, E., Puig, V., Quevedo, J., Escobet, T., and Comasolivas, R. (2018). Autonomous vehicle control using a kinematic lyapunov-based technique with lqr-lmi tuning. *Control engineering practice*, 73:1–12.
- Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., and Li, H. (2024). End-to-end autonomous driving: Challenges and frontiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10164–10183.
- Emirler, M. T., Uygan, İ. M. C., Aksun Güvenç, B., and Güvenç, L. (2014). Robust pid steering control in parameter space for highly automated driving. *International Journal of Vehicular Technology*, 2014(1):259465.
- Fleming, W. H. and Rishel, R. W. (2012). *Deterministic*

- and stochastic optimal control, volume 1. Springer Science & Business Media.
- Gu, J., Bellone, M., Pivoňka, T., and Sell, R. (2024). Clft: Camera-lidar fusion transformer for semantic segmentation in autonomous driving. *arXiv preprint arXiv:2404.17793*.
- Gu, J., Lind, A., Chhetri, T. R., Bellone, M., and Sell, R. (2023). End-to-end multimodal sensor dataset collection framework for autonomous vehicles. *Sensors*, 23(15):6783.
- Huang, K., Shi, B., Li, X., Li, X., Huang, S., and Li, Y. (2022). Multi-modal sensor fusion for auto driving perception: A survey. *arXiv preprint arXiv:2202.02703*.
- ISO (2011). Iso 26262 road vehicles– functional safety. *ISO Standard (2011)*.
- Karafyllis, I., Theodosis, D., and Papageorgiou, M. (2022). Lyapunov-based two-dimensional cruise control of autonomous vehicles on lane-free roads. *Automatica*, 145:110517.
- Kirovskii, O. and Gorelov, V. (2019). Driver assistance systems: analysis, tests and the safety case. iso 26262 and iso pas 21448. In *IOP Conference Series: Materials Science and Engineering*, volume 534, page 012019. IOP Publishing.
- Kuznietsov, A., Gyevar, B., Wang, C., Peters, S., and Albrecht, S. V. (2024). Explainable ai for safe and trustworthy autonomous driving: A systematic review. *arXiv preprint arXiv:2402.10086*.
- LaValle, S. M. (2006). *Planning algorithms*. Cambridge university press.
- Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., and Lee, S.-I. (2020). From local explanations to global understanding with explainable ai for trees. *Nature machine intelligence*, 2(1):56–67.
- Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.
- Malayjerdi, E., Sell, R., Malayjerdi, M., Udal, A., and Bellone, M. (2022). Practical path planning techniques in overtaking for autonomous shuttles. *Journal of Field Robotics*, 39(4):410–425.
- Omeiza, D., Web, H., Jirotko, M., and Kunze, L. (2021). Towards accountability: Providing intelligible explanations in autonomous driving. In *2021 IEEE Intelligent Vehicles Symposium (IV)*, pages 231–237. IEEE.
- Omeiza, D., Webb, H., Jirotko, M., and Kunze, L. (2022). Explanations in autonomous driving: A survey. *IEEE Transactions on Intelligent Transportation Systems*, 23(8):10142–10162.
- Pikner, H., Malayjerdi, M., Bellone, M., Baykara, B. C., and Sell, R. (2024). Autonomous driving validation and verification using digital twins. *VEHITS*, pages 204–211.
- Ribeiro, M. T., Singh, S., and Guestrin, C. (2016). "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, San Francisco, CA, USA, August 13-17, 2016*, pages 1135–1144.
- Stanton, N. A. and Young, M. S. (1998). Vehicle automation and driving performance. *Ergonomics*, 41(7):1014–1028.
- Weisser, H., Schulenberg, P., Gollinger, H., and Michler, T. (1999). Autonomous driving on vehicle test tracks: overview, implementation and vehicle diagnosis. In *Proceedings 199 IEEE/IEEJ/ISAI International Conference on Intelligent Transportation Systems (Cat. No. 99TH8383)*, pages 62–67. IEEE.
- Yurtsever, E., Lambert, J., Carballo, A., and Takeda, K. (2020). A survey of autonomous driving: Common practices and emerging technologies. *IEEE access*, 8:58443–58469.