Vision-Language Model-Guided Post-Training for Explainable End-to-End Autonomous Driving

Ottilia Steen
Department of Computer Science, University of Houston, Houston, TX, USA.
ottiliasteen@uh.edu

Adriano Brunello
Department of Computer Science, Colorado State University, Fort Collins, CO, USA.
adrianobrunello190@colostate.edu

Abstract

Vision-language models have become a significant force in artificial intelligence because they connect visual perception with natural language reasoning, enabling systems to describe, question, and justify what they observe. This paper examines how vision-language model-guided post-training can be used to improve explainability in end-to-end autonomous driving. End-to-end driving systems learn to map sensor data directly to control outputs, but their opacity creates serious problems for safety assurance, regulatory oversight, and public trust. Post-training with vision-language models offers a way to refine driving policies after initial pretraining by aligning them with language-based explanations, human preferences, and semantic constraints. The paper argues that this approach should not be treated as a narrow model improvement technique. Instead, it must be understood as a socio-technical architecture involving data pipelines, simulation infrastructure, edge and cloud computation, safety monitors, validation regimes, governance structures, and sustainability considerations. The analysis covers the foundations of end-to-end driving, the architecture of vision-language model-guided post-training, explainability as a systems property, robustness and validation, fairness and policy, deployment and infrastructure, and cross-domain comparisons with aviation, medicine, and industrial automation. It also identifies limitations, including hallucination risk, distribution shift, computational cost, unclear liability, and the absence of standardized evaluation for explanation quality. The central conclusion is that vision-language model-guided post-training can make end-to-end autonomous driving more interpretable and governable, but only if explainability, safety, fairness, and sustainability are treated as first-class requirements throughout the system lifecycle rather than as retrospective additions.

Keywords

vision-language models; end-to-end autonomous driving; post-training; explainability; safety assurance; governance; deployment; socio-technical systems.

1. Introduction

End-to-end autonomous driving has emerged as a compelling alternative to modular pipelines that separate perception, prediction, planning, and control. By learning a direct mapping from sensor observations to driving actions, end-to-end systems can reduce engineering complexity and capture subtle correlations that hand-designed interfaces may miss [1], [2]. Early work demonstrated that deep networks could steer vehicles from camera images, and subsequent research extended this idea to conditional imitation learning, multimodal datasets, and increasingly realistic simulators [3], [4], [5]. However, the same architectural simplification

that makes end-to-end learning attractive also creates a fundamental problem: the internal decision process is difficult to inspect, audit, or explain. This opacity is not merely an academic concern. It affects accident investigation, regulatory approval, insurance, public acceptance, and the ability of engineers to diagnose rare but dangerous behaviors.

Vision-language models provide a possible response to this problem. These models are pretrained on large-scale image-text and video-text data, allowing them to associate visual scenes with semantic descriptions, relational reasoning, and natural language instructions [6], [7], [8], [9]. When used during post-training, a vision-language model can act as a critic, teacher, annotator, or explanation generator for a driving policy. Post-training is especially important because a base driving model may be trained primarily for imitation or reinforcement objectives, while a vision-language model can introduce language-grounded signals about scene understanding, causal relations, social norms, and human preferences [10], [11], [12]. The result is not simply a more verbose model. It is a system that can potentially produce explanations that are tied to observable scene elements, compare alternative actions, and justify decisions in terms that human operators and regulators can evaluate.

This paper argues that vision-language model-guided post-training for explainable end-to-end autonomous driving should be approached as a large-scale systems problem. The relevant questions are not limited to model accuracy or benchmark performance. They include how explanation signals are generated and validated, how post-training data is governed, how safety monitors interact with learned policies, how deployment across edge and cloud infrastructure affects latency and energy use, how bias in vision-language pretraining may propagate into driving behavior, and how legal and policy frameworks can assign responsibility when explanations are incomplete or misleading. The discussion therefore combines technical architecture with governance, infrastructure, robustness, fairness, and sustainability. The aim is to provide a structured account of the opportunities and trade-offs that arise when vision-language models are used not merely to describe driving scenes but to shape the post-training of end-to-end driving systems.

2. Foundations and Motivation

The motivation for vision-language model-guided post-training begins with the limitations of purely behavioral training. End-to-end driving models are often trained through imitation learning, where the model predicts actions from expert demonstrations [1], [2]. This approach can achieve impressive performance in familiar conditions, but it struggles with causal confusion, distribution shift, and rare events. A model may learn to associate a particular steering angle with irrelevant visual cues, such as time of day or roadside objects, rather than with the actual traffic situation. Datasets such as KITTI, nuScenes, and Waymo Open Dataset have improved scale and diversity, but they cannot cover the full long tail of real-world driving [16], [17], [18]. Interpretability research has attempted to expose attention patterns, saliency maps, and intermediate representations, yet these methods often explain correlation rather than causation and may not align with human reasoning [3], [4], [19].

Vision-language models introduce a different kind of training signal. Because they are pretrained on broad multimodal corpora, they can describe scenes, answer questions, and follow instructions in natural language [6], [7], [8], [9]. Post-training methods such as instruction tuning and reinforcement learning from human feedback have shown that pretrained models can be aligned with human preferences and task-specific constraints [10], [11]. In autonomous driving, a vision-language model can be asked to evaluate whether a proposed action is safe, socially appropriate, or consistent with traffic rules. It can also

generate explanations that compare a chosen action with alternatives. A recent line of work on post-training in end-to-end autonomous driving [12] illustrates the growing interest in using such signals to refine driving policies after initial training. The promise is that language can serve as a bridge between low-level control and high-level human values.

However, the motivation must be tempered by structural realities. Vision-language models are not ground-truth oracles. They can hallucinate objects, misunderstand spatial relations, or produce fluent but incorrect rationales. They may inherit biases from web-scale pretraining, including cultural assumptions about road behavior, pedestrian roles, and urban order. They are computationally expensive, which raises questions about whether post-training and inference can be performed on vehicles or must rely on cloud infrastructure. They also create new dependencies: if a driving policy is aligned to a vision-language critic, then the limitations of that critic become part of the safety case. The motivation for this paper is therefore not to celebrate vision-language post-training uncritically, but to examine how it can be designed, governed, and deployed in a way that genuinely improves explainability without creating new and less visible failure modes.

3. Architecture of Vision-Language Model-Guided Post-Training

A vision-language model-guided post-training architecture for end-to-end driving typically includes several interacting components. The first is a base driving policy that maps multimodal sensor inputs to trajectories or control commands. The second is a vision-language model or ensemble of such models that can observe driving scenes, receive prompts, and produce judgments, rationales, or preference rankings. The third is a data engine that stores driving episodes, sensor data, human annotations, simulation rollouts, and model-generated explanations. The fourth is a training and evaluation loop that uses vision-language outputs to update the policy through imitation, preference optimization, reinforcement learning, or constrained fine-tuning. The fifth is a runtime safety and explanation layer that monitors the deployed policy and produces human-facing justifications. These components are not independent. Their interfaces determine whether post-training improves explainability or merely adds a language veneer over an opaque controller.

The architecture must confront a central trade-off between richness and reliability. Vision-language models can provide detailed semantic feedback, such as noting that a pedestrian is waiting at a crosswalk or that a lane closure requires cooperative merging. Such feedback can be used to shape the policy toward socially aware behavior [6], [7], [9]. Yet the same models may be slow, expensive, and inconsistent. If every training sample requires a large vision-language model query, the post-training pipeline becomes costly and difficult to reproduce. If the model is used only sparsely, its influence may be limited to a narrow set of scenarios. A practical architecture may therefore use distillation, where a large vision-language model generates explanations or preference labels that are then distilled into a smaller critic or directly into the driving policy [10], [12]. This reduces inference cost but introduces a compression risk: important nuances in the explanation may be lost when the teacher signal is transferred to a smaller student.

Another architectural question concerns the relationship between explanation and action. In some designs, the vision-language model generates an explanation after the driving policy has selected an action. This post hoc explanation may be useful for human review, but it does not necessarily influence behavior. In other designs, the vision-language model participates in action selection or post-training, so its reasoning shapes the policy itself. The latter approach is more ambitious because it can improve decisions, but it also makes the system more

dependent on the vision-language model's world knowledge. If the model misunderstands a construction zone or a police officer's hand signal, the resulting policy update may encode that misunderstanding. The architecture must therefore include mechanisms for uncertainty estimation, cross-checking with rule-based systems, and human review of high-stakes post-training examples. Simulation environments such as CARLA and real-world datasets can provide grounding, but they cannot eliminate the need for careful validation [5], [15], [16], [17], [18].

4. Explainability as a Systems Property

Explainability in autonomous driving is often discussed as a model-level property, but it is better understood as a systems property. An explanation is useful only if it reaches the right stakeholder, at the right time, in the right form, and with the right level of abstraction. A safety engineer may need causal details about sensor fusion and trajectory prediction. A regulator may need evidence that the vehicle followed traffic rules and considered vulnerable road users. A passenger may need a concise statement about why the vehicle braked or changed lanes. A fleet operator may need aggregate explanations for incident triage and maintenance. A vision-language model can help generate these different explanation types, but the system must manage their consistency and provenance. If the post hoc explanation contradicts the actual decision process, the result is not transparency but a new form of opacity [19], [20].

Vision-language model-guided post-training can improve explainability by forcing the driving policy to align with language-based rationales during training. For example, if a policy proposes a lane change, the vision-language critic can ask whether the adjacent lane is clear, whether the maneuver is socially appropriate, and whether it creates unnecessary risk. These questions can be translated into preference signals or constraints that shape the policy [10], [11], [12]. Over time, the policy may learn representations that are more semantically structured and therefore easier to explain. However, this benefit depends on the quality of the language supervision. If the vision-language model produces superficial or biased explanations, the policy may learn to mimic those explanations without acquiring the underlying causal competence. Explainability then becomes performative: the system produces plausible language while remaining unsafe in edge cases.

A systems view also requires attention to explanation governance. Who defines what counts as a good explanation? How are explanation datasets curated? How are disagreements between human annotators and vision-language models resolved? How long are explanations stored, and who can access them? These questions connect explainability to privacy, liability, and organizational accountability [21], [22], [23]. An explanation may contain personal data, such as pedestrian appearance or location, and its retention may create surveillance risks. A regulatory audit may require access to training rationales, but releasing them publicly could reveal proprietary information or create security vulnerabilities. The architecture must therefore include access control, audit trails, redaction mechanisms, and clear policies for explanation use. Without such governance, vision-language model-guided post-training may produce explanations that satisfy a technical metric but fail the social and legal purposes for which they are intended.

5. Robustness, Safety, and Validation

Robustness is a prerequisite for any safety-critical deployment. End-to-end driving models are vulnerable to distribution shift, sensor degradation, adversarial perturbations, and rare events

that are underrepresented in training data. Vision-language model-guided post-training can improve robustness by adding semantic constraints that generalize across environments. For instance, a vision-language critic may recognize that a child near the curb requires heightened caution even if that exact scenario was not in the driving dataset [6], [7], [9]. It may also help the policy reason about occlusions, complex intersections, and ambiguous right-of-way situations. Simulation platforms such as CARLA allow controlled stress testing of these scenarios, while datasets such as nuScenes and Waymo Open Dataset provide large-scale real-world diversity [5], [16], [17]. KITTI remains useful for benchmarking perception and geometry, although its scale and sensor configuration are limited relative to newer datasets [18].

Yet robustness cannot be assumed. Vision-language models may themselves be brittle to unusual visual conditions, such as heavy rain, glare, or construction signage. They may confidently describe objects that are not present or fail to notice small but critical cues. If post-training relies on these descriptions, errors can be amplified. Validation must therefore include not only driving performance but also explanation fidelity, uncertainty calibration, and failure mode analysis. An assurance case should specify what the vision-language model is allowed to influence, under what conditions, and how the system degrades when the model is uncertain. Safety monitors that operate independently of the learned policy can provide a final layer of protection, but they too must be validated and governed [15], [20]. The goal is not to prove absolute safety, which is impossible for complex open-world systems, but to make residual risk visible, bounded, and subject to ongoing review.

Validation also raises the problem of ground truth for explanations. Driving outcomes can be measured through collisions, near misses, comfort, and rule compliance, but explanations are more difficult to score. A fluent explanation may be persuasive yet false. A terse explanation may be correct but insufficient for a regulator. Research on interpretable machine learning has emphasized that explanation quality depends on the task, user, and context [19]. In autonomous driving, this means that evaluation must include human factors studies, audit simulations, and adversarial tests in which explanations are deliberately manipulated. Vision-language model-guided post-training should therefore be accompanied by a validation framework that treats explanations as safety-relevant artifacts rather than marketing material.

6. Fairness, Governance, and Policy

Fairness in autonomous driving is a socio-technical challenge that vision-language model-guided post-training can either mitigate or exacerbate. Vision-language models are trained on large, unevenly distributed corpora that reflect historical and cultural biases about roads, pedestrians, neighborhoods, and authority [21], [22]. If these biases enter post-training signals, the driving policy may behave differently across regions, income levels, or demographic contexts. For example, a model might be more cautious in wealthy areas with well-maintained infrastructure and less cautious in areas with informal road use. Such disparities can become safety inequities. Fairness auditing must therefore examine both driving outcomes and explanations across geographic, demographic, and environmental groups. It is not enough to show that the model performs well on average.

Governance must also address the post-training pipeline itself. Who supplies the data? Who labels it? Who decides which vision-language model is used? Who can update the policy after deployment? These questions are political as much as technical, because they determine whose values are embedded in the system [23]. A fleet operator may prioritize efficiency and smooth traffic flow, while a city may prioritize pedestrian safety and equitable access. A

vision-language model may encode default assumptions that favor one set of priorities over another. Transparent governance requires documented decision rights, stakeholder representation, and mechanisms for appeal. Policy instruments such as safety standards, transparency mandates, and liability rules can shape these arrangements, but they must be flexible enough to accommodate rapid technological change. The global landscape of artificial intelligence ethics guidelines offers principles, yet implementation remains fragmented and often voluntary [22]. Autonomous driving needs more concrete accountability mechanisms.

Liability is especially complex when explanations are generated by a vision-language model. If an accident occurs, investigators may ask whether the driving policy made an error, whether the vision-language critic provided flawed post-training signals, whether the safety monitor failed, or whether the explanation misled human operators. The causal chain can span multiple organizations, including model providers, vehicle manufacturers, fleet operators, data vendors, and regulators. Legal frameworks traditionally focus on human drivers or product defects, but they are not fully adapted to systems that learn continuously and rely on third-party foundation models. Clear contracts, audit logs, version control, and incident reporting are necessary. Vision-language model-guided post-training can support liability analysis by preserving rationales and decision traces, but only if those artifacts are reliable, tamper-evident, and accessible to authorized parties. Otherwise, explainability may become a source of dispute rather than a tool for accountability.

7. Deployment, Infrastructure, and Sustainability

Deployment architecture determines whether vision-language model-guided post-training is practical at scale. Large vision-language models require substantial computation, memory, and energy for both training and inference [6], [7]. Running such models on every vehicle in real time may be infeasible, especially for cost-sensitive or battery-constrained platforms. A common approach is to split workloads between edge and cloud. The vehicle may run a distilled driving policy and a lightweight explanation module, while cloud infrastructure handles large-scale post-training, simulation, and periodic model updates. This division reduces latency and bandwidth pressure, but it introduces connectivity dependencies and cybersecurity risks. If the cloud is unavailable, the vehicle must still drive safely using onboard capabilities. The architecture must therefore define graceful degradation modes and ensure that post-training improvements do not become runtime necessities.

Data infrastructure is equally important. Vision-language model-guided post-training depends on large-scale driving logs, human feedback, simulation rollouts, and model-generated rationales. Managing this data requires storage systems, metadata schemas, privacy controls, and versioning. Fleet learning can accelerate improvement by aggregating diverse experiences, but it also raises surveillance and data ownership concerns [17]. Data minimization, on-device preprocessing, and differential privacy can reduce risk, but they may also limit the richness of post-training signals. Infrastructure choices therefore have direct consequences for fairness and trust. A system that collects extensive video and location data may improve performance but alienate users and invite regulatory scrutiny.

Sustainability is another often overlooked dimension. Training and serving large vision-language models consume significant energy, and continuous post-training can multiply that cost. If autonomous driving is deployed across millions of vehicles, the aggregate computational footprint may be substantial. Efficiency techniques such as model distillation, quantization, sparse activation, and selective vision-language queries can help, but they

involve trade-offs between cost and explanation quality [10], [12]. Deployment strategies should also consider lifecycle impacts, including hardware manufacturing, replacement cycles, and electronic waste. A sustainable approach would prioritize reusable models, standardized interfaces, and shared infrastructure rather than duplicating large-scale compute across competing fleets. Sustainability is not only an environmental issue; it affects long-term economic viability and public legitimacy.

8. Case Illustrations and Cross-Domain Comparisons

Comparisons with other safety-critical domains can clarify the promise and limits of vision-language model-guided post-training. Aviation has developed layered safety through checklists, redundancy, incident reporting, and certification regimes. A vision-language explanation layer in autonomous driving could play a role similar to a cockpit voice recorder or maintenance log, but it would need comparable standards for reliability and preservation. Medical artificial intelligence offers another analogy. Clinical decision support systems must provide explanations that clinicians can understand and contest, yet they are also subject to rigorous validation and post-market surveillance [19], [21]. Autonomous driving could borrow from medical device regulation by requiring prospective clinical-style trials, adverse event reporting, and periodic revalidation after model updates.

Industrial automation provides a third comparison. Robotic systems in factories operate in structured environments where safety envelopes, emergency stops, and formal verification are more tractable. Autonomous driving operates in open, unpredictable public space, where social norms and informal behavior matter. Vision-language models can help interpret that social context, but they cannot replace formal safety mechanisms. The most plausible architecture is hybrid: rule-based and control-theoretic safety layers handle hard constraints, while vision-language-guided post-training improves semantic understanding and explanation quality [15], [22]. This mirrors developments in robotics, where learned components are often wrapped by verifiable controllers.

Cross-domain comparison also highlights the importance of institutional readiness. Aviation and medicine have mature regulators, professional norms, and liability regimes. Autonomous driving spans multiple jurisdictions and mixes public and private actors. A vision-language explanation that satisfies a technical team may not satisfy a city transportation department, a civil rights organization, or an insurance regulator. The system must therefore support multiple explanation registers and documentation formats. Case studies of urban deployments, especially in mixed traffic with pedestrians and cyclists, show that trust is built through consistent behavior, transparent communication, and accessible redress mechanisms [23]. Vision-language model-guided post-training can contribute to these goals, but only as part of a broader institutional design.

9. Limitations and Future Directions

Several limitations remain. First, vision-language models can hallucinate and may produce explanations that are fluent but unfaithful to the actual decision process [20]. Faithfulness requires technical methods for tracing how explanations relate to internal representations and actions, as well as organizational processes for auditing them. Second, post-training signals are only as good as the data and feedback that generate them. Human preferences may be inconsistent, culturally specific, or biased, and vision-language models may amplify those tendencies [10], [11]. Third, computational cost limits the frequency and scale of vision-language-guided updates. Distillation and selective querying can help, but they may reduce

the nuance of the supervision. Fourth, evaluation remains immature. There is no widely accepted benchmark for explanation quality in driving, nor a standard way to compare vision-language critics across scenarios.

Future research should pursue several directions. One is the development of grounded explanation benchmarks that combine simulation, real-world logs, and human factors studies. Another is the design of uncertainty-aware post-training, where the vision-language model can abstain or request human review when its confidence is low [12]. A third is the integration of formal methods with learned components, so that safety envelopes and rule-based constraints bound the influence of language-based updates [15]. A fourth is participatory governance, in which cities, residents, and civil society organizations help define what constitutes acceptable driving behavior and adequate explanation [22], [23]. Finally, sustainability must be treated as a design constraint, not an afterthought. Efficient architectures, shared data infrastructure, and transparent lifecycle accounting can reduce the environmental and economic burden of large-scale post-training.

10. Conclusion

Vision-language model-guided post-training offers a promising path toward more explainable end-to-end autonomous driving. By using language and multimodal reasoning to shape driving policies after initial training, it can align behavior with human values, improve semantic robustness, and generate explanations that serve engineers, regulators, passengers, and the public. However, the approach is not a simple add-on to existing models. It creates new architectural dependencies, validation burdens, governance questions, and infrastructure costs. Explainability must be designed as a systems property that spans data collection, model training, runtime monitoring, incident investigation, and policy oversight. Robustness, fairness, privacy, and sustainability must be addressed alongside accuracy. The central lesson is that vision-language models can make autonomous driving more governable, but only if their limitations are acknowledged and their influence is bounded by safety, accountability, and democratic oversight. The future of explainable end-to-end driving will depend less on any single model and more on the socio-technical architecture that surrounds it.

References

1. Codevilla, F., Müller, M., López, A., Koltun, V., & Dosovitskiy, A. (2018). End-to-end driving via conditional imitation learning. In Proceedings of the IEEE International Conference on Robotics and Automation (pp. 4693–4700). IEEE.

2. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., & Zieba, K. (2016). End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316.

3. Kim, J., & Canny, J. (2017). Interpretable learning for self-driving cars by visualizing causal attention. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2942–2950). IEEE.

4. Xu, H., Gao, Y., Yu, F., & Darrell, T. (2017). End-to-end learning of driving models from large-scale video datasets. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2174–2182). IEEE.

5. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An open urban driving simulator. In Proceedings of the 1st Annual Conference on Robot Learning (pp. 1–16). PMLR.

6. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning (pp. 8748–8763). PMLR.

7. Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., ... & Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. In Advances in Neural Information Processing Systems (Vol. 35, pp. 23716–23736). Curran Associates.

8. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the 39th International Conference on Machine Learning (pp. 12888–12900). PMLR.

9. Chen, X., Wang, X., Changpinyo, S., Piergiovanni, A. J., Padlewski, P., Salz, D., Goodman, S., Grycner, A., Mustafa, B., Beyer, L., Kolesnikov, A., Puigcerver, J., Ding, N., Rong, K., Akbari, H., Mishra, G., Xue, L., Thapliyal, A. V., Bradbury, J., ... & Soricut, R. (2023). PaLI: A jointly-scaled multilingual language-image model. In Proceedings of the International Conference on Learning Representations. OpenReview.

10. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In Advances in Neural Information Processing Systems (Vol. 35, pp. 27730–27744). Curran Associates.

11. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems (Vol. 30, pp. 4299–4307). Curran Associates.

12. Winner, L. (1980). Do artifacts have politics? Daedalus, 109(1), 121–136.

13. Kendall, A., Hawke, J., Janz, D., Mazur, P., Reda, D., Allen, J. M., Lam, V. D., Bewley, A., & Shah, A. (2019). Learning to drive in a day. In Proceedings of the IEEE International Conference on Robotics and Automation (pp. 8248–8254). IEEE.

14. Amini, A., Rosman, G., Karaman, S., & Rus, D. (2019). Variational end-to-end navigation and localization. In Proceedings of the IEEE International Conference on Robotics and Automation (pp. 8958–8964). IEEE.

15. Sadat, A., Casas, S., Ren, M., Wu, X., Dhawan, P., & Urtasun, R. (2020). Perceive, predict, and plan: Safe motion planning through interpretable semantic representations. In Proceedings of the European Conference on Computer Vision (pp. 414–430). Springer.

16. Caesar, H., Bankiti, V., Lang, A. H., Vora, S., Liong, V. E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., & Beijbom, O. (2020). nuScenes: A multimodal dataset for autonomous driving. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 11621–11631). IEEE.

17. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., ... & Anguelov, D. (2020). Scalability in

perception for autonomous driving: Waymo Open Dataset. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2446–2454). IEEE.

18. Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The KITTI dataset. The International Journal of Robotics Research, 32(11), 1231–1237.

19. 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.

20. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

21. Mittelstadt, B., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Big Data & Society, 3(2), 1–21.

22. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), 389–399.