

Continual Learning and Test-Time Adaptation for Autonomous Driving Models Under Distribution Shifts

Liana Casagrande

Department of Computer Science, University of North Texas, Denton, TX, USA.
lianac@unt.edu

Michal Bielas

Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
bielas1996@unr.edu

Rebeca Verdejo

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
rebecamail@ku.edu

Abstract

Autonomous driving models are trained under assumptions of environmental, behavioral, and infrastructural regularity that rarely hold in open-world deployment. Distribution shifts arise from weather, illumination, sensor degradation, map obsolescence, regional driving norms, construction patterns, and evolving pedestrian behavior, producing failures that are difficult to predict from static validation. This paper examines continual learning and test-time adaptation as complementary system-level strategies for maintaining autonomous driving performance under such shifts. Rather than treating adaptation as a model-centric optimization problem, the analysis emphasizes architectural trade-offs, operational governance, infrastructure requirements, sustainability, fairness, and policy implications. Continual learning offers mechanisms for retaining and updating knowledge across fleet-scale data streams, but introduces risks of catastrophic forgetting, feedback loops, memory inflation, and uneven representation across regions. Test-time adaptation enables rapid adjustment during deployment, yet can amplify errors, destabilize safety envelopes, and complicate certification when model behavior changes after release. The paper argues that robust autonomous driving requires a layered adaptation architecture that separates safety-critical invariants from tunable representations, combines cloud-based continual learning with constrained edge adaptation, and embeds auditing, rollback, uncertainty monitoring, and fairness evaluation into the update pipeline. Through cross-domain comparisons with medical imaging, industrial control, and robotics, the discussion identifies governance models and evaluation paradigms needed to make adaptive driving systems trustworthy, sustainable, and equitable across diverse mobility environments.

Keywords

Continual learning; test-time adaptation; autonomous driving; distribution shift; robustness; socio-technical systems; governance; fairness; edge computing; safety assurance.

1. Introduction

Autonomous driving systems represent a demanding test of machine learning under open-world conditions. End-to-end driving models, modular perception-planning stacks, and hybrid architectures all rely on learned representations that must remain reliable as environments, sensors, users, and infrastructure change [14], [16]. The operating domain of a vehicle is not a fixed dataset but a moving socio-technical field shaped by weather, road maintenance, traffic law, local conventions, and human behavior. When a model encounters conditions that differ from its training distribution, performance can degrade sharply, sometimes in ways that are invisible to aggregate accuracy metrics [1], [2]. This problem is not limited to rare corner cases; seasonal change, sensor aging, software updates, and urban redevelopment can produce persistent distribution shifts that affect large portions of a fleet.

Standard supervised learning assumes that training and deployment data are drawn from related distributions, an assumption that is routinely violated in autonomous driving [1]. Dataset bias and benchmark overfitting further obscure the gap between laboratory performance and field reliability [2], [3]. Models trained on large-scale datasets may exploit texture, lighting, or geographic cues that correlate with labels but do not generalize to new cities or sensor configurations [4], [5]. Consequently, autonomous driving requires learning systems that can update, adapt, and forget selectively after deployment. Continual learning addresses this need by enabling models to acquire new knowledge without catastrophically overwriting prior competencies [6], [7], [8]. Test-time adaptation addresses a different temporal scale by allowing inference-time adjustment to unlabeled target data, often with minimal or no access to source data [9], [10], [11], [12], [13].

However, neither continual learning nor test-time adaptation is a simple technical fix. Each introduces structural trade-offs among plasticity, stability, latency, memory, energy consumption, interpretability, and regulatory accountability. A vehicle that adapts too aggressively may become unpredictable; a vehicle that adapts too slowly may fail in precisely the conditions that demand robustness. Fleet-level learning can improve broad coverage but risks homogenizing behavior around dominant regions and manufacturers. Edge adaptation can preserve privacy and responsiveness but may create a fragmented safety landscape in which different vehicles behave differently under identical conditions. These tensions make adaptation a governance problem as much as a modeling problem.

This paper develops a system-level analysis of continual learning and test-time adaptation for autonomous driving under distribution shifts. It argues that effective adaptation must be architected as a layered capability embedded in operational infrastructure, not added as an isolated algorithm. The discussion proceeds through the nature of distribution shifts in driving, the architecture of continual learning, the role and risks of test-time adaptation, governance and fairness, cross-domain comparisons, integrated evaluation, and policy implications. Recent work on post-training in end-to-end autonomous driving [15] illustrates that adaptation can span pre-deployment refinement and post-deployment update, but it must be situated within a broader safety and accountability framework.

2. Distribution Shifts in Autonomous Driving as a Socio-Technical Problem

Distribution shifts in autonomous driving can be categorized in several overlapping ways. Covariate shift occurs when input distributions change while the underlying task remains similar, such as a model trained in clear daylight encountering fog, rain, or glare [5]. Concept shift occurs when the relationship between inputs and safe actions changes, such as when pedestrians behave differently in a new cultural context or when temporary traffic rules alter right-of-way conventions. Label shift occurs when the prevalence of object classes or events

changes, such as a route with unusually high bicycle density. Sensor shift occurs when camera calibration, lidar reflectivity, radar interference, or hardware aging alters the data-generating process. Temporal shift occurs as the world changes after training, including new vehicle models, clothing styles, road markings, and construction patterns. Geographic shift occurs when models trained in one region are deployed in another with different infrastructure, weather, and driving norms [18], [19].

These shifts are not merely technical noise. They are produced by social, economic, and infrastructural processes. A city that repaints lane markings, introduces congestion pricing, or redesigns intersections changes the operational domain for every vehicle. A region with underinvestment in road maintenance may present sensor and planning challenges that are underrepresented in datasets collected in well-maintained areas. A community with different pedestrian behavior, informal transport, or policing practices may be systematically absent from benchmark data [21]. When adaptation systems learn primarily from fleet data, they may reinforce the dominance of regions where data collection is easiest and most profitable. Fairness therefore becomes entangled with data supply chains, update frequency, and infrastructure equity.

The socio-technical nature of distribution shift also affects evaluation. Benchmark datasets such as nuScenes and Waymo Open Dataset have advanced multimodal perception and forecasting, but they remain snapshots of particular cities, sensor suites, and time periods [18], [19]. Simulators such as CARLA provide controllable variation, yet the sim-to-real gap means that robustness in simulation does not automatically transfer to public roads [17]. A model may appear robust under synthetic fog or rain while failing under real precipitation because the physical interaction of light, water, and sensor hardware is different. Static test sets cannot capture the cumulative effect of minor shifts, nor can they represent the feedback loops that occur when adapted vehicles influence the behavior of other road users.

Continual learning and test-time adaptation respond to these shifts at different levels. Continual learning treats deployment as a sequence of tasks or domains and seeks to preserve performance across them [6], [7], [8]. Test-time adaptation treats each deployment episode as an opportunity to adjust representations using unlabeled data [9], [10], [11]. In practice, autonomous driving requires both: slow, governed updates that improve fleet-level competence, and fast, bounded adjustments that maintain safety within a trip. The central design question is not whether to adapt but where adaptation is allowed, how it is validated, who is accountable, and how its benefits and risks are distributed.

3. Continual Learning Architectures and Structural Trade-Offs

Continual learning methods generally fall into regularization, replay, parameter isolation, and dynamic architecture families. Regularization-based approaches constrain updates to important parameters, reducing catastrophic forgetting while allowing some plasticity [6]. Replay-based approaches retain or generate samples from previous distributions, which can be effective but raises memory, privacy, and data governance concerns. Parameter isolation assigns separate capacity to different tasks or domains, which can preserve safety-critical behavior but may fragment the model and increase inference complexity. Dynamic architectures expand or reconfigure networks as new domains arrive, which supports plasticity but complicates verification and hardware deployment [7], [8]. Each approach embodies a trade-off between stability and plasticity, and each has consequences for latency, energy, and maintainability.

In autonomous driving, the stability-plasticity trade-off is not an abstract learning problem. A vehicle must remain predictable to passengers, regulators, and other road users. If continual learning changes perception thresholds, trajectory preferences, or risk calibration, the vehicle's behavior may shift in ways that are difficult to explain after an incident. Safety-critical invariants, such as minimum following distance under defined conditions or conservative pedestrian yielding, should therefore be separated from adaptable components. Adaptation can be permitted in representation layers, domain-specific normalization, or auxiliary perception modules while planning constraints and safety envelopes remain governed by verifiable rules. This layered separation is a system architecture decision, not merely a training objective.

Fleet-scale continual learning introduces infrastructure requirements. Data must be collected, filtered, labeled or pseudo-labeled, stored, and distributed across edge devices and cloud systems. Vehicles may upload difficult examples, but this process raises privacy, bandwidth, and cybersecurity concerns. Edge training can reduce data transfer but may be constrained by compute, thermal, and power limits. Cloud training can leverage larger models and curated datasets but may introduce latency between problem discovery and fleet update. A practical architecture often combines edge inference, cloud-based continual learning, simulation-based validation, and over-the-air deployment. The resulting pipeline is a socio-technical system with organizational roles for data engineers, safety auditors, legal teams, and regulators.

Continual learning also risks feedback loops. If vehicles adapt to data generated by their own policies, they may reinforce existing biases or drift into locally optimal but globally unsafe behavior. For example, if a fleet learns to avoid a difficult intersection rather than solve it, route planning may concentrate traffic elsewhere. If adaptation prioritizes frequently encountered conditions, rare but high-risk scenarios may receive less attention. Governance must therefore include deliberate sampling of underrepresented conditions, adversarial scenario generation, and independent evaluation. The goal is not continuous improvement in average performance but maintained safety across a distribution of conditions, including those that are rare, costly, or politically salient.

4. Test-Time Adaptation and Operational Robustness

Test-time adaptation enables a deployed model to adjust using unlabeled target data. Entropy minimization, self-supervision, normalization statistics updates, and parameter-free methods have shown that models can improve under corruption, domain shift, and sensor variation without labels [9], [10], [11], [12]. Continual test-time adaptation extends this idea across non-stationary domains, allowing adaptation to evolve as conditions change [13]. In autonomous driving, these methods are attractive because conditions can shift within a single trip. A vehicle may move from highway to urban streets, from daylight to night, or from dry pavement to rain. Waiting for cloud-based retraining is not sufficient for such short-timescale changes.

Yet test-time adaptation is dangerous if treated as an unrestricted online learning process. Adaptation on unlabeled data can reinforce incorrect predictions, collapse representations, or amplify noise. A model that becomes overconfident in a misclassified object may make increasingly poor decisions. Safety-critical systems therefore require constraints on what can adapt, how much it can adapt, and how quickly it can change. Adaptation may be limited to normalization parameters, uncertainty calibration, or low-level perception corrections while planning logic remains fixed. It should be monitored by out-of-distribution detectors, uncertainty estimators, and consistency checks across sensors. If confidence drops or

disagreements rise, the system should fall back to a conservative mode rather than continue adapting.

Recent work on post-training in end-to-end autonomous driving [15] suggests that adaptation can be integrated into a broader pipeline that includes pre-deployment refinement, scenario-specific tuning, and post-deployment updates. This is significant because it moves adaptation away from a single algorithm and toward a lifecycle capability. However, post-training and test-time adaptation must still respect operational boundaries. A vehicle cannot continuously rewrite its own safety case in the field. Updates must be traceable, reversible, and validated against regression suites that include rare events, edge cases, and fairness-sensitive scenarios. The system must be able to answer what changed, why it changed, and whether the change affected safety performance.

Operational robustness also depends on infrastructure. Over-the-air updates require secure boot, signed firmware, rollback capability, and version compatibility across vehicle generations. Edge adaptation requires sufficient compute and thermal headroom without compromising real-time perception and control. Cloud services must support data ingestion, model versioning, experiment tracking, and audit logging. These are not incidental engineering details; they determine whether adaptation is sustainable at fleet scale. A technically effective method that consumes excessive energy, requires constant connectivity, or creates unmanageable maintenance burdens may fail in practice.

5. Governance, Deployment, Sustainability, and Fairness

Governance for adaptive autonomous driving must address liability, certification, transparency, and continuous compliance. Traditional vehicle safety cases assume a fixed configuration approved before deployment. Continual learning and test-time adaptation challenge that assumption because behavior can change after approval. Regulators may require a defined adaptation envelope, mandatory logging, human oversight for certain updates, and periodic recertification. Standards bodies can develop procedures for evaluating adaptation methods under controlled shift scenarios while preserving room for innovation [22], [23]. The key is to make adaptation auditable without requiring disclosure of proprietary details that would undermine competition or security.

Fairness is a central but underappreciated dimension. If adaptation is driven by fleet data, regions with more vehicles, better connectivity, or stronger data collection capacity will contribute more to model updates. Regions with older vehicles, limited infrastructure, or privacy-sensitive populations may be underrepresented. This can produce performance disparities in pedestrian detection, sign recognition, or risk estimation [21]. Fairness in autonomous driving is not only about demographic attributes; it also concerns geography, infrastructure quality, vehicle age, and mobility patterns. An adaptive system that performs well in dense, well-mapped cities but poorly in rural or informal settlements may deepen transport inequity.

Sustainability adds another constraint. Continual learning and test-time adaptation both consume computation. Cloud retraining, simulation, data storage, and over-the-air transmission have energy and carbon costs. Edge adaptation may reduce transmission but increase onboard power draw, which affects electric vehicle range and battery degradation. The environmental impact of adaptation must be weighed against safety benefits. In some cases, selective adaptation under high uncertainty may be more sustainable than continuous adaptation. In others, fleet-level consolidation of updates may be more efficient than per-

vehicle learning. Sustainability requires lifecycle assessment rather than narrow accuracy comparisons.

Deployment practices also shape public trust. Shadow mode deployment, in which adaptive models run alongside production systems without controlling the vehicle, can provide evidence before activation. Canary fleets in specific regions can test updates under controlled conditions. Rollback mechanisms allow rapid reversal if anomalies appear. Incident reporting and third-party audits can strengthen accountability. However, these practices require organizational capacity and regulatory coordination. Smaller operators may struggle to meet the same governance standards as large manufacturers, raising concerns about market concentration. Policy design must balance safety assurance with the need for a competitive and innovative mobility sector.

6. Case Illustrations and Cross-Domain Comparisons

Medical imaging offers a useful comparison because it also faces domain shift across scanners, hospitals, and patient populations. Continual learning and test-time adaptation have been explored to maintain diagnostic performance when protocols change [7], [9]. The stakes are high, but the operational environment is more controlled than public roads. A hospital can validate an update before clinical use, and the number of devices is limited. Autonomous driving faces a more open world with real-time safety constraints and heterogeneous hardware. This suggests that driving requires stronger runtime constraints and more conservative adaptation than many medical applications.

Industrial control provides a second comparison. In manufacturing, distribution shift may arise from equipment wear, new materials, or changed operating conditions. Adaptive control systems often separate safety interlocks from optimization layers, allowing adaptation only within certified bounds [23]. This layered safety model is directly relevant to autonomous driving. A vehicle can adapt perception and prediction while planning constraints remain fixed or verified. The analogy also highlights the importance of maintenance schedules, calibration, and human oversight. Adaptation is not a substitute for physical maintenance or operational discipline.

Robotics offers a third comparison. Robots in warehouses or homes must adapt to new objects, layouts, and human behavior. Test-time adaptation and continual learning have been used to improve manipulation and navigation [10], [13]. However, robots often operate in bounded environments with slower dynamics and fewer irreversible consequences than road vehicles. Autonomous driving must handle high-speed interactions, legal liability, and public scrutiny. The comparison suggests that driving should adopt the conservatism of industrial safety systems, the data governance of medical systems, and the flexibility of robotics, while recognizing that no single domain provides a complete template.

Cross-domain lessons also concern evaluation. In medical imaging, external validation on independent sites is standard. In industrial control, simulation and hardware-in-the-loop testing are common. In robotics, benchmark suites and real-world trials are combined. Autonomous driving needs a similar portfolio: simulation, closed-course testing, shadow mode, controlled fleet trials, and independent audits. Adaptation methods should be evaluated not only by average accuracy under shift but by worst-case safety, latency, energy, and fairness outcomes. A method that improves average performance while increasing variance or harming underrepresented regions is not suitable for deployment.

7. Integrated System Design and Evaluation Paradigms

An integrated adaptation architecture for autonomous driving should distinguish among layers. At the sensor layer, calibration and signal processing can adapt to hardware drift and environmental conditions. At the perception layer, domain-specific normalization or feature alignment can address weather, lighting, and geographic variation. At the prediction layer, adaptation can adjust to local behavioral norms. At the planning layer, adaptation should be highly constrained or prohibited unless verified against safety envelopes. At the fleet layer, continual learning can consolidate updates across vehicles, supported by simulation and human review. This layered model allows different rates of adaptation and different governance requirements.

Evaluation must correspondingly become multi-dimensional. Conventional metrics such as mean average precision or collision rate are necessary but insufficient. An adaptive system should be assessed for robustness across shifts, stability under long-tail conditions, calibration of uncertainty, recovery after failure, and transparency of updates. Safety cases should include adaptation-specific hazards such as unintended behavioral drift, delayed response to sensor degradation, or uneven performance across regions. Fairness metrics should examine whether adaptation improves or harms performance for underrepresented environments and populations [21]. Sustainability metrics should account for energy, bandwidth, and hardware wear.

Infrastructure for integrated evaluation includes data curation, scenario generation, simulation, and continuous monitoring. A data flywheel can identify difficult cases, but it must be governed to prevent bias and privacy violations. Simulation can generate rare events, but it must be validated against real-world observations. Monitoring can detect drift, but it must trigger meaningful action rather than alert fatigue. Audit logs can support accountability, but they must be secure and interpretable. These requirements point toward an MLOps and safety engineering convergence in which model updates, vehicle software, and regulatory reporting are managed as one lifecycle.

Finally, integrated design must consider human factors. Drivers, passengers, remote operators, and pedestrians interact with vehicles as social actors. If adaptation changes driving style, signaling, or yielding behavior, it may affect trust and coordination. Test-time adaptation that makes a vehicle more aggressive in dense traffic could improve throughput in simulation but increase risk and discomfort in reality. Continual learning that shifts lane-changing preferences may create inconsistent behavior across a fleet. Human-centered evaluation is therefore essential, especially when adaptation affects observable behavior rather than only internal representations.

8. Policy Implications and Future Research

Policy for adaptive autonomous driving should avoid two extremes. One extreme is to prohibit post-deployment adaptation because it complicates certification, which would leave vehicles brittle in changing environments. The other is to allow unrestricted learning in the field, which would undermine accountability and safety. A middle path defines adaptation envelopes, requires logging and rollback, mandates independent evaluation for safety-critical changes, and enforces transparency about update effects. Regulators can specify what types of adaptation are permitted without recertification, what types require notification, and what types require new approval. This differentiated approach recognizes that adapting normalization statistics is not equivalent to rewriting a planner [23].

Future research should advance several directions. First, methods are needed for safe test-time adaptation with formal or statistically rigorous bounds on behavioral change. Second, continual learning should better handle non-stationary, open-world distributions with unknown boundaries and overlapping domains. Third, uncertainty estimation and out-of-distribution detection must become reliable enough to gate adaptation in real time. Fourth, fairness-aware adaptation should explicitly measure and mitigate disparities across geography, infrastructure, and population. Fifth, energy-efficient adaptation should be co-designed with edge hardware and vehicle power systems. Sixth, governance research should develop audit mechanisms that protect privacy and intellectual property while providing meaningful public accountability [20], [22].

Another promising direction is the integration of foundation models with constrained adaptation. Large pretrained models may provide broad representations, while small adapter modules or post-training procedures [15] can specialize behavior for regions or conditions. However, foundation models also introduce opacity, high compute costs, and centralized update dynamics. Their use in safety-critical driving requires careful scrutiny. Future systems may combine foundation-model perception with rule-based planning, continual learning at the fleet level, and bounded test-time adaptation at the edge. The challenge is to make such hybrid architectures verifiable, sustainable, and equitable.

9. Conclusion

Continual learning and test-time adaptation are essential capabilities for autonomous driving under distribution shifts, but they are not merely algorithmic additions. They reshape the relationship between training and deployment, safety and flexibility, fleet-level learning and local responsiveness, and public accountability and proprietary innovation. Continual learning enables vehicles to acquire new knowledge over time, yet it risks forgetting, bias reinforcement, and infrastructure burdens. Test-time adaptation enables rapid adjustment in changing conditions, yet it can destabilize behavior and complicate certification. The most credible path forward is a layered adaptation architecture that separates safety-critical invariants from tunable components, combines cloud and edge learning, and embeds auditing, rollback, uncertainty monitoring, fairness evaluation, and sustainability assessment into the update lifecycle. Policy must support controlled adaptation rather than prohibit or ignore it. With such system-level design, autonomous driving can become more robust, equitable, and trustworthy as the world around it continues to change.

References

1. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, 79(1-2), 151-175. <https://doi.org/10.1007/s10994-009-5152-4>
2. Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1521-1528). IEEE. <https://doi.org/10.1109/CVPR.2011.5995347>
3. Recht, B., Roelofs, R., Schmidt, L., & Shankar, V. (2019). Do ImageNet classifiers generalize to ImageNet? In *Proceedings of the 36th International Conference on Machine Learning* (pp. 5389-5400). PMLR.
4. Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). ImageNet-trained CNNs are biased towards texture; increasing shape bias

improves accuracy and robustness. In Proceedings of the 7th International Conference on Learning Representations. OpenReview.

5. Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In Proceedings of the 7th International Conference on Learning Representations. OpenReview.
6. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13), 3521-3526. <https://doi.org/10.1073/pnas.1611835114>
7. Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., & Wermter, S. (2019). Continual lifelong learning with neural networks: A review. *Neural Networks*, 113, 54-71. <https://doi.org/10.1016/j.neunet.2019.02.001>
8. De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., & Tuytelaars, T. (2021). A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7), 3366-3385. <https://doi.org/10.1109/TPAMI.2021.3057446>
9. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., & Darrell, T. (2021). Tent: Fully test-time adaptation by entropy minimization. In Proceedings of the 9th International Conference on Learning Representations. OpenReview.
10. Sun, Y., Wang, X., Liu, Z., Miller, J., Efros, A. A., & Hardt, M. (2020). Test-time training with self-supervision for generalization under distribution shifts. In Proceedings of the 37th International Conference on Machine Learning (pp. 9229-9248). PMLR.
11. Boudiaf, M., Mueller, R., Ben Ayed, I., & Bertinetto, L. (2022). Parameter-free online test-time adaptation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 8344-8353). IEEE. <https://doi.org/10.1109/CVPR52688.2022.00816>
12. Schneider, S., Rusak, E., Eck, L., Bringmann, O., Brendel, W., & Bethge, M. (2020). Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 11539-11551). Curran Associates.
13. Wang, Q., Fink, O., Van Gool, L., & Dai, D. (2022). Continual test-time domain adaptation. In Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7201-7211). IEEE. <https://doi.org/10.1109/CVPR52688.2022.00706>
14. 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*.
15. Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2017). On a formal model of safe and scalable self-driving cars. *arXiv preprint arXiv:1708.06374*.
16. Codevilla, F., Müller, M., López, A., Koltun, V., & Dosovitskiy, A. (2018). End-to-end driving via conditional imitation learning. In Proceedings of the 2018 IEEE International

Conference on Robotics and Automation (pp. 4693-4700). IEEE.
<https://doi.org/10.1109/ICRA.2018.8460487>

17. Dosovitskiy, A., Ros, G., Codevilla, F., López, A., & Koltun, V. (2017). CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning* (pp. 1-16). PMLR.
18. 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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11621-11631). IEEE.
<https://doi.org/10.1109/CVPR42600.2020.01164>
19. 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 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2446-2454). IEEE. <https://doi.org/10.1109/CVPR42600.2020.00252>
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. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1-35.
<https://doi.org/10.1145/3457607>
22. Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, 1(9), 389-399. <https://doi.org/10.1038/s42256-019-0088-2>