Data-Centric Post-Training for Improving Long-Tail Scenario Generalization in End-to-End Autonomous Driving

Svetlana Geissler
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
hellosvetlana@unh.edu

Abstract

End-to-end autonomous driving has advanced rapidly by integrating perception, prediction, and control into unified learned systems. Yet deployment remains constrained by long-tail scenarios: rare, ambiguous, or geographically specific situations that are underrepresented in training corpora but disproportionately important for safety. This paper argues that improving long-tail generalization requires a data-centric post-training paradigm rather than isolated architectural refinement. Post-training, in this context, refers to the structured adaptation of a pre-trained driving policy using curated, targeted, and often synthetic data, together with human feedback, simulation, and validation loops. The paper examines the structural trade-offs that arise when data-centric post-training becomes a system-level strategy. It considers how data infrastructure, labeling governance, scenario mining, simulation fidelity, fleet telemetry, and safety assurance interact across the autonomous driving stack. It also explores deployment consequences, including robustness under distribution shift, fairness across regions and road users, sustainability of continual learning pipelines, and policy implications for certification and liability. The central claim is that long-tail performance is not merely a model property but an institutional and infrastructural achievement. Effective post-training therefore depends on governance mechanisms that make rare events discoverable, comparable, auditable, and reusable without creating brittle overfitting to narrow test distributions. The paper concludes that data-centric post-training offers a credible path toward safer end-to-end autonomy, but only if technical pipelines are coupled with transparent data provenance, operational design domain discipline, and cross-organizational evaluation practices.

Keywords

data-centric artificial intelligence, post-training, end-to-end autonomous driving, long-tail generalization, safety assurance, governance, deployment.

1. Introduction

End-to-end autonomous driving represents a significant departure from modular autonomy pipelines. Instead of separating perception, localization, prediction, planning, and control into independently engineered components, end-to-end systems learn mappings from sensor observations to driving actions or trajectories [1]. Early demonstrations showed that convolutional networks could steer a vehicle using raw camera input [1], while later work introduced conditional imitation learning to incorporate navigational commands and route intent [2]. Other approaches learned intermediate affordances such as lane position and distance to obstacles, preserving some interpretability while still relying on learned perception [3]. These developments established a compelling research direction: complex driving

behavior might emerge from data and optimization rather than from exhaustive handcrafted rules.

However, the operational promise of end-to-end driving depends on generalization. Public datasets such as KITTI, nuScenes, and the Waymo Open Dataset have enabled substantial progress in perception and prediction [5], [6], [7]. Simulators such as CARLA have provided scalable environments for training and testing urban driving policies [4]. Yet the distribution of everyday driving is not the distribution that determines safety. Long-tail scenarios, including unusual pedestrian behavior, occluded intersections, temporary construction layouts, severe weather, nonstandard vehicles, and rare interactions among multiple agents, remain difficult to collect, label, and validate. The core problem is not simply that models lack capacity. It is that the data ecosystem surrounding autonomy systematically underrepresents events that are rare individually but inevitable collectively.

Data-centric artificial intelligence responds to this problem by treating data quality, coverage, provenance, and feedback as first-class engineering concerns. In autonomous driving, data-centricity intersects with post-training. A model may be pre-trained on large-scale driving logs, simulation data, or multimodal corpora, and then adapted through targeted fine-tuning, preference optimization, reinforcement learning from feedback, or scenario-specific replay. Recent work on post-training in end-to-end autonomous driving [10] highlights that adaptation after initial training can materially affect robustness and scenario handling. This paper extends that insight into a system-level argument: long-tail generalization is achieved not by a single post-training method but by an institutional and infrastructural architecture that continuously identifies, curates, simulates, validates, and governs rare events.

The argument proceeds in several stages. First, it frames the long-tail problem as a socio-technical challenge rather than a purely statistical one. Second, it explains why data-centric post-training is an architectural strategy with consequences for latency, memory, modularity, and validation. Third, it examines data infrastructure and curation, including scenario mining, synthetic generation, and provenance. Fourth, it discusses governance, labeling, and organizational interfaces. Fifth, it analyzes deployment, robustness, and safety assurance. Sixth, it considers fairness, sustainability, and policy implications. Finally, it proposes evaluation and continual learning principles that align technical performance with public accountability.

2. The Long-Tail Problem in End-to-End Driving

The long tail in autonomous driving is often described as a statistical phenomenon: many rare scenarios account for a large fraction of safety-critical risk. This description is accurate but incomplete. In practice, the tail is produced by the interaction of physical geography, social norms, infrastructure quality, weather, vehicle mix, regulatory regimes, and human behavior. A scenario that is rare in one city may be common in another. A maneuver that is legal but unusual may appear frequently in informal traffic systems. A sensor configuration that performs well in clear daylight may fail under reflective rain or low sun. The long tail is therefore not a fixed list of events. It is a moving, context-dependent frontier shaped by deployment location and operational design domain.

End-to-end models are especially sensitive to this frontier because they learn correlations across the entire input-output mapping. Unlike modular systems, where perception errors can sometimes be bounded by downstream planners, end-to-end policies may propagate subtle distribution shifts directly into control. A model trained predominantly on dense urban traffic

may struggle with rural roads, unpaved surfaces, or livestock crossings. A model trained in one country may misinterpret lane markings, traffic signals, or informal right-of-way conventions. Dataset bias research has long shown that visual datasets encode geographic and cultural regularities that models exploit as shortcuts [8]. In autonomous driving, such shortcuts can become safety failures when deployment moves beyond the original data distribution.

Long-tail generalization is further complicated by the economics of data collection. Rare events are expensive to observe, label, and validate. Fleets may drive millions of miles before encountering a specific combination of weather, road geometry, and agent behavior. Even when such events are recorded, they may be difficult to annotate because their correct handling is contested. Human drivers themselves may disagree about the safest response. This uncertainty means that long-tail data is not simply scarce; it is epistemically complex. The goal of post-training is therefore not only to increase the number of examples but to structure learning around scenarios whose correct treatment can be justified, audited, and reproduced.

Simulation offers one response by allowing rare events to be generated on demand [4]. Synthetic scenario generation can vary pedestrian trajectories, occlusions, sensor noise, and adversarial agent behavior. Yet simulation introduces its own distributional gap. A policy optimized heavily in simulation may exploit simulator artifacts or fail under real sensor physics. The long tail thus becomes a problem of hybrid evidence: real-world logs provide authenticity, simulation provides controllability, and human feedback provides normative judgment. Data-centric post-training must integrate these evidence sources without allowing any one source to dominate in ways that undermine real-world safety.

3. Data-Centric Post-Training as an Architectural Strategy

Post-training is often discussed as a model adaptation step, but in end-to-end autonomous driving it is better understood as an architectural strategy. The pre-trained policy may be large and multimodal, benefiting from broad exposure to driving scenes, language instructions, or world knowledge [11]. Post-training then reshapes behavior using targeted data, reward signals, or constraint-based fine-tuning. In language models, instruction tuning and preference optimization have demonstrated that relatively small amounts of curated data can substantially alter model behavior [12], [13], [14]. Autonomous driving can borrow from this logic, but it must also respect real-time constraints, safety certification, and physical consequences.

The architectural trade-offs are substantial. If post-training is performed centrally and infrequently, updates can be rigorously validated but may lag behind emerging edge cases. If it is performed continuously across a fleet, adaptation can be rapid but risks instability, catastrophic forgetting, and opaque model drift. If post-training focuses on a narrow scenario family, performance may improve sharply on that family while degrading elsewhere. If it uses broad replay and regularization, improvements may be safer but slower and less pronounced. These trade-offs resemble the hidden technical debt described in machine learning systems, where data dependencies, configuration, and feedback loops create maintenance burdens that are not visible in model accuracy alone [9].

A system-level view also reveals that post-training changes the boundary between training and deployment. In traditional software, deployment is a discrete release. In data-centric autonomy, deployment generates data that feeds post-training, which produces new models that are redeployed. This loop can be virtuous if governed well, but it can also amplify biases.

For example, if fleet data is collected predominantly in favorable weather, post-training may overrepresent easy conditions. If human feedback is gathered only from safety drivers in one region, it may encode local driving culture as universal. If simulation scenarios are generated from historical accidents, they may miss novel failure modes. The architecture must therefore include mechanisms for detecting coverage gaps, not merely optimizing on available data.

There is also a modularity question. End-to-end systems aim to avoid brittle interfaces, but data-centric post-training may reintroduce modularity at the data layer. Scenario catalogs, reward models, safety critics, and validation suites can act as semi-independent modules that shape the policy. This can improve auditability, because each data source and reward signal can be inspected. However, it can also fragment responsibility. If a failure occurs, is it due to the pre-trained policy, the post-training data, the reward model, the simulator, or the deployment context? A mature architecture must preserve provenance across these layers so that failures can be traced to specific data and governance decisions.

4. Data Infrastructure and Curation for Rare Events

The effectiveness of data-centric post-training depends on infrastructure that can discover, store, query, and curate long-tail events. Fleet telemetry is a primary source. Vehicles continuously collect sensor streams, vehicle dynamics, driver interventions, and system confidence estimates. The challenge is to identify which segments are valuable. Simple triggers such as harsh braking or system disengagement may capture some failures, but they can miss near misses, ambiguous scenes, or gradual distribution shifts. More sophisticated mining can use uncertainty estimation, novelty detection, disagreement among ensemble models, or semantic search over scenario descriptions. The goal is not to collect more data indiscriminately but to build a searchable memory of rare and instructive situations.

Curation is equally important. A raw event is not automatically a training example. It must be contextualized: what was the operational design domain, what were the sensor conditions, what agents were present, what was the intended route, and what constitutes a safe response? Labeling rare events is expensive and may require multiple annotators with driving, legal, or domain expertise. In some cases, the correct action is not a single trajectory but a set of acceptable behaviors constrained by safety and comfort. Data curation therefore involves negotiation among engineering, legal, ethical, and operational stakeholders. This is a governance activity as much as a technical one.

Synthetic data can supplement real-world curation. Simulation allows controlled variation of parameters that are difficult to observe together in reality [4]. However, synthetic data must be validated against real sensor statistics, agent behavior distributions, and physical dynamics. Otherwise, post-training may overfit to unrealistic patterns. Domain randomization can improve robustness, but it can also blur distinctions that matter for safety. A balanced approach treats simulation as a hypothesis-generating tool: it proposes scenarios that are then validated, refined, or rejected using real-world evidence and expert review. This iterative relationship between real and synthetic data is central to long-tail generalization.

Infrastructure must also support reproducibility. If a post-training run improves long-tail performance, engineers need to know which data subsets, labels, augmentations, and validation cases contributed. Provenance metadata should link model versions to data versions, scenario definitions, annotation guidelines, and approval records. Without such linkage, organizations cannot learn systematically from failures. They may repeatedly fix symptoms

while leaving root causes in the data pipeline. The infrastructure burden is significant, but it is the price of treating data as a safety-critical asset rather than a consumable resource.

5. Governance, Labeling, and Organizational Interfaces

Data-centric post-training is not only a computational pipeline; it is an organizational process. Long-tail scenarios cross departmental boundaries. Perception teams may focus on sensor limitations, planning teams on trajectory feasibility, legal teams on regulatory compliance, and operations teams on customer experience. A rare event that is trivial for one group may be critical for another. Governance mechanisms must create shared definitions of severity, acceptability, and evidentiary sufficiency. Otherwise, data curation becomes politically contested and technically inconsistent.

Labeling governance is especially important. Annotators are often treated as interchangeable labor, but rare-event labeling requires contextual judgment. A cut-in vehicle may require different responses depending on speed, road geometry, visibility, and local law. A pedestrian at the edge of the road may be waiting, walking, or about to cross. These distinctions cannot be reduced to simple bounding boxes. They require scenario-level annotation, often with video, map, and temporal context. Organizations must therefore invest in training, calibration, and audit of annotators, as well as mechanisms for resolving disagreement.

Human feedback can also be used more directly. Preference-based methods ask humans to compare trajectories or behaviors, and then train a reward model or policy to align with those preferences [13], [14]. In driving, this approach is attractive because it can capture subtle notions of comfort, caution, and social compliance. Yet it raises risks. Human preferences may vary by culture, experience, and risk tolerance. A preference model trained on one population may not represent the public interest. Moreover, optimizing too aggressively against a learned reward model can lead to reward hacking, where the policy exploits imperfections in the feedback signal rather than learning safe driving. Governance must therefore constrain post-training with hard safety rules, diverse feedback panels, and independent validation.

Organizational interfaces also determine how quickly rare events move from discovery to remediation. A strong data-centric organization maintains a closed loop: field incidents and near misses are triaged, scenario owners are assigned, data is curated, post-training is performed, validation is conducted, and deployment decisions are documented. This loop resembles safety management systems in aviation, where reporting, investigation, and corrective action are institutionalized. Autonomous driving can learn from that tradition [20]. The difference is that software updates can propagate rapidly across fleets, making both the benefits and the risks of post-training scalable. Governance must be fast enough to address emerging tails but slow enough to prevent unvalidated changes from reaching public roads.

6. Deployment, Robustness, and Safety Assurance

Deployment changes the character of long-tail generalization. A model that performs well on a held-out dataset may still fail when exposed to new geographies, weather, or social practices. Robustness therefore must be evaluated across operational design domains rather than as a single aggregate metric. Safety assurance requires evidence that the system handles rare events not only in simulation but under realistic uncertainty. Formal models of safe driving can provide useful constraints, but they often rely on assumptions that are difficult to satisfy in open-world traffic [18]. Post-training must be integrated with runtime monitoring, fallback behaviors, and conservative decision boundaries.

One deployment strategy is shadow mode, where a candidate policy runs alongside the active system without controlling the vehicle. Shadow mode can reveal disagreements and novel scenarios without immediate risk. Another strategy is geofenced deployment, where the operational design domain is restricted to regions and conditions with sufficient data coverage. Geofencing is often criticized as a limitation, but it is also a governance tool: it aligns deployment with validated competence. As post-training improves long-tail coverage, the geofence can expand. This creates a measurable relationship between data investment and operational scope.

Robustness also depends on how post-training handles uncertainty. A policy should recognize when it is outside its competence and request human intervention or adopt a minimal-risk condition. This is difficult in end-to-end systems because uncertainty is distributed across the network. Post-training can include abstention examples, conservative fallback trajectories, and simulated sensor degradation. However, over-conservatism can make the system unusable or socially disruptive. The trade-off between safety and mobility is not purely technical; it reflects public values and regulatory expectations. Deployment decisions should therefore be transparent about acceptable risk and performance envelopes.

Safety assurance must also address the problem of rare but correlated failures. A model may perform well on individual long-tail scenarios but fail when multiple rare conditions coincide, such as heavy rain, a nonstandard intersection, and a distracted pedestrian. Post-training datasets often isolate scenarios, but real deployment composes them. Validation must therefore include combinatorial stress testing. Scenario catalogs should support composition, parameter variation, and adversarial sequencing. The goal is not to enumerate every possibility but to build confidence that the policy degrades gracefully when conditions exceed training experience.

7. Fairness, Sustainability, and Policy Implications

Long-tail generalization has fairness dimensions. If autonomous vehicles perform better in wealthy, well-mapped, and orderly environments, they may deliver safety benefits unevenly. Regions with informal road use, poor signage, or mixed vehicle types may remain underserved. A data-centric post-training strategy that prioritizes commercially attractive markets can reinforce these disparities. Fairness requires deliberate data collection in diverse environments, engagement with local communities, and evaluation metrics that disaggregate performance by geography, road user type, and scenario category. The long tail is not only a technical residual; it is a map of whose driving conditions are considered normal.

Sustainability is another system-level concern. Continual post-training consumes computation, storage, bandwidth, and human labor. Fleet-scale data collection can generate enormous volumes of video and sensor data. If every rare event triggers retraining, the environmental and financial costs may become prohibitive. Organizations need curation policies that prioritize high-value scenarios, reuse validated datasets, and retire obsolete data. Model updates should be efficient, perhaps through parameter-efficient adaptation or selective replay. Sustainability also includes social sustainability: the labor conditions of annotators, safety drivers, and maintenance workers are part of the autonomy infrastructure. A data-centric approach that ignores these conditions is incomplete.

Policy implications follow from these concerns. Regulators may require documentation of operational design domains, data provenance, validation coverage, and incident reporting. Certification cannot rely solely on aggregate miles driven, because rare events are

disproportionately important [19]. Instead, regulators and industry may need scenario-based evidence, simulation audits, and post-market surveillance. Liability regimes must clarify responsibility when a post-trained model fails on a scenario that was known but underrepresented. If manufacturers are expected to learn continuously, they also need legal safe harbors for good-faith updates that are properly validated. Policy can encourage transparency by standardizing scenario descriptions, safety cases, and data-sharing protocols for rare events.

Cross-domain comparisons are instructive. Aviation safety relies on mandatory incident reporting, independent investigation, and shared databases of failures. Medical devices use post-market surveillance to detect rare adverse events. Pharmaceutical regulation uses phased trials and pharmacovigilance. Autonomous driving needs analogous institutions, but with faster software cycles and less centralized oversight. Data-centric post-training can support these institutions if it produces auditable records of what was learned, from whom, and under what assumptions. Without such records, post-training may improve benchmark performance while weakening public accountability.

8. Evaluation, Metrics, and Organizational Learning

Evaluation for long-tail generalization must move beyond average displacement error or collision rate. These metrics are useful but can hide failure modes. A system may have low average error while performing poorly on a specific class of intersections or vulnerable road users. Evaluation should therefore be stratified by scenario type, operational design domain, and uncertainty level. It should include near misses, comfort violations, rule compliance, and human intervention rates. It should also measure calibration: whether the system knows when it is likely to fail. Calibration is essential for fallback and human-machine interaction.

Organizational learning requires that evaluation results feed back into data curation. If a scenario family shows persistent failures, the response should not be merely to add more of the same data. It may require new sensors, better maps, revised annotation guidelines, or changes to the operational design domain. Post-training is thus one part of a larger learning architecture. The architecture should support experiments that isolate the effect of data interventions. For example, an organization might compare post-training with curated real data, synthetic data, and human feedback to understand which source improves which scenario category. Such experiments are expensive but necessary for responsible scaling.

Finally, evaluation should be adversarial and participatory. Internal teams may unconsciously design tests that favor their own systems. External red teams, independent auditors, and community representatives can surface blind spots. Participatory methods can also reveal local knowledge about road behavior that is absent from corporate datasets. Long-tail generalization is ultimately a public safety goal, not only a product feature. Its evaluation should therefore include affected communities, regulators, and independent researchers. This does not eliminate trade-offs, but it makes them explicit and contestable.

9. Conclusion

Improving long-tail scenario generalization in end-to-end autonomous driving requires more than a better model. It requires a data-centric post-training architecture that treats rare events as a strategic asset and a governance challenge. The paper has argued that post-training should be understood systemically: as a loop connecting fleet telemetry, scenario mining, simulation, human feedback, labeling, validation, deployment, and policy. Each stage introduces structural trade-offs among adaptability, stability, cost, fairness, and safety. Data-

centric methods can improve performance on rare scenarios, but they can also create new brittleness if they overfit to narrow datasets, encode biased feedback, or outpace validation.

The most credible path forward combines technical rigor with institutional discipline. Organizations need provenance-aware data infrastructure, diverse and well-governed annotation, simulation validated against reality, and evaluation that is stratified, adversarial, and transparent. Regulators need scenario-based safety cases, post-market surveillance, and clear rules for continuous learning. Researchers need benchmarks that reward long-tail competence without encouraging benchmark gaming. The long tail will never be fully enumerated, but it can be progressively governed. End-to-end autonomy will succeed not when it has seen everything, but when it has built the capacity to notice, learn from, and responsibly act on what it has not seen.

References

1. 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.

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

3. Chen, C., Seff, A., Kornhauser, A., & Xiao, J. (2015). DeepDriving: Learning affordance for direct perception in autonomous driving. In Proceedings of the IEEE International Conference on Computer Vision (ICCV) (pp. 2722–2730). IEEE.

4. 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 (CoRL) (pp. 1–16). PMLR.

5. 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 11621–11631). IEEE.

6. 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 2446–2454). IEEE.

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

8. Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 1521–1528). IEEE.

9. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In Advances in Neural Information Processing Systems 28 (pp. 2503–2511). Curran Associates.

10. Winner, H., Hakuli, S., Lotz, F., & Singer, C. (Eds.). (2016). Handbook of driver assistance systems: Basic information, components and systems for active safety and comfort. Springer.

11. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.

12. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., 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 35 (pp. 27730–27744). Curran Associates.

13. Ziegler, D. M., Stiennon, N., Wu, J., Brown, T. B., Radford, A., Amodei, D., Christiano, P., & Irving, G. (2019). Fine-tuning language models from human preferences. arXiv preprint arXiv:1909.08593.

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

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

16. Bansal, M., Krizhevsky, A., & Ogale, A. (2018). ChauffeurNet: Learning to drive by imitating the best and synthesizing the worst. arXiv preprint arXiv:1812.03079.

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

18. Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2017). On a formal model of safe and scalable self-driving cars. arXiv preprint arXiv:1708.06374.

19. Kalra, N., & Paddock, S. M. (2016). Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? Transportation Research Part A: Policy and Practice, 94, 182–193.