

Digital Twin-Assisted Post-Training for Autonomous Driving Validation in Complex Urban Scenarios

Yfke Lafleur

School of Computing, Clemson University, Clemson, SC, USA.
yfkell@clemson.edu

Lakshmi L. Sehgal

Department of Computer Science, Binghamton University, Binghamton, NY, USA.
sehgal1975@binghamton.edu

Abstract

Autonomous driving validation remains one of the most difficult socio-technical assurance problems because safety must be demonstrated across rare, ambiguous, and rapidly changing urban conditions that cannot be exhaustively reproduced on public roads. This paper examines digital twin-assisted post-training as a system-level strategy for validating autonomous driving systems in complex urban scenarios. Rather than treating digital twins as merely high-fidelity simulators, the analysis frames them as governed validation infrastructures that mediate data collection, scenario generation, model adaptation, safety assessment, and regulatory evidence. The discussion connects post-training methods for end-to-end driving with urban simulation, multimodal datasets, imitation and reinforcement learning, and emerging assurance practices. It then analyzes architectural trade-offs among edge, cloud, and hybrid deployments, focusing on latency, synchronization, privacy, data sovereignty, and resilience. The paper also considers robustness, fairness, sustainability, and policy implications, including the risk that digital twin validation may privilege well-instrumented cities while underrepresenting marginalized neighborhoods, informal transport patterns, and vulnerable road users. It argues that digital twin-assisted post-training can improve coverage and reduce physical testing burdens only when paired with transparent scenario taxonomies, auditable data governance, independent oversight, and cross-domain learning from aerospace, industrial automation, and critical infrastructure assurance. The conclusion emphasizes that technical gains in simulation realism and model adaptation must be embedded in institutional accountability, otherwise validation may become more efficient without becoming more legitimate.

Keywords

autonomous driving, digital twins, post-training, urban validation, scenario generation, safety assurance, data governance, socio-technical systems.

1. Introduction

Autonomous driving systems are often evaluated through the lens of mileage, disengagement rates, and simulation coverage, but these metrics reveal only part of the assurance problem. Complex urban environments contain layered uncertainties: pedestrians may behave unpredictably, cyclists may occupy ambiguous spaces, construction zones may alter lane semantics, and emergency vehicles may impose temporary rules that contradict ordinary traffic conventions. Foundational work on safe self-driving models has emphasized the need for formal or structured reasoning about perception, planning, and fault tolerance, yet such

models remain difficult to translate into comprehensive validation for end-to-end learning systems [1]. At the same time, statistical studies of reliability show that demonstrating acceptable failure rates through road testing alone would require distances and timeframes that are practically infeasible [2]. This mismatch between empirical road testing and the scale of rare-event validation has driven interest in simulation, scenario libraries, and continuous assurance architectures [3].

Digital twins offer a broader concept than conventional simulators. In industrial and complex-systems contexts, a digital twin is a virtual representation that is connected to a physical asset or process through data flows that support monitoring, prediction, and control [4]. Industrial digital twins have evolved from static models into dynamic, data-intensive systems that integrate sensor streams, physics-based models, and machine learning for optimization and lifecycle management [5]. The convergence of digital twins with Internet of Things platforms and artificial intelligence further allows these representations to act as decision-support infrastructures rather than passive visualizations [6]. For autonomous driving, a digital twin of an urban district could combine road geometry, traffic signals, weather, pedestrian flows, vehicle telemetry, and regulatory constraints into a continually updated validation environment. The central question is not whether such a twin can be built, but how it should be governed, synchronized, audited, and integrated into post-training workflows without creating new blind spots or institutional dependencies.

This paper argues that digital twin-assisted post-training should be understood as a socio-technical validation regime. Post-training, in this context, refers to the adaptation, refinement, or alignment of pre-trained driving models using targeted data, simulated scenarios, reward signals, or human feedback, including methods developed for end-to-end autonomous driving [7]. When post-training is coupled with digital twins, the validation environment becomes both a training ground and an evidentiary instrument. This dual role creates structural trade-offs. Greater simulation fidelity can improve coverage of dangerous scenarios, but it can also amplify the assumptions embedded in the model, the map, or the scenario generator. More centralized data collection can improve consistency, yet it may weaken local accountability and create privacy risks. Faster deployment cycles can accelerate innovation, but they can also outpace regulatory capacity and public understanding. The sections that follow examine these tensions across architecture, governance, robustness, fairness, sustainability, and policy.

2. Digital Twins as Validation Infrastructure for Autonomous Systems

A digital twin for autonomous driving validation is not simply a three-dimensional city model. It is an infrastructure that maintains partial, probabilistic, and often contested correspondence between a physical urban environment and a computational representation. That correspondence depends on sensor calibration, map updates, communication latency, data quality, and assumptions about which features of the city matter for driving behavior. In the industrial domain, digital twins are valued because they can reduce downtime, predict maintenance, and optimize operations across a lifecycle [5]. In urban mobility, the same logic appears as continuous validation, where vehicles, roadside units, traffic management centers, and cloud services exchange data to update scenario libraries and evaluate model behavior. The twin becomes a boundary object among engineers, city agencies, insurers, regulators, and the public, each of whom may need different forms of evidence from the same virtual environment [4].

The architectural choice of what to twin is consequential. A corridor-level twin may focus on intersections, pedestrian crossings, and signal timing, while a metropolitan twin may model

regional traffic flows, freight routes, and emergency response networks. A vehicle-level twin may represent the autonomous system itself, including perception, planning, and control modules, while an environment-level twin represents the world in which the vehicle operates. The most ambitious approach combines both, creating a closed-loop system in which the vehicle model is tested against simulated urban dynamics that are themselves updated by real-world data. This closed loop can support post-training by generating targeted edge cases, but it also raises the risk of circular validation. If the same data sources, ontologies, and assumptions are used to train and evaluate the model, the validation may confirm internal consistency rather than real-world safety. Digital twin governance must therefore separate, at least partially, the data used for adaptation from the data used for independent assessment.

Infrastructure requirements further shape what kinds of twins are feasible. High-fidelity urban twins require pervasive sensing, stable connectivity, precise maps, and computational resources for simulation and data fusion. Edge computing can reduce latency for safety-critical updates, but it may struggle with large-scale scenario generation. Cloud computing can aggregate data across regions, but it introduces bandwidth costs, privacy concerns, and dependency on network availability. Hybrid architectures attempt to balance these constraints by processing time-sensitive perception and control locally while using cloud resources for scenario mining, model retraining, and regulatory reporting. The trade-off is not merely technical. It determines which cities can participate in digital twin validation, which vehicles can be certified, and which populations are represented in the resulting safety evidence. Without deliberate equity provisions, digital twin infrastructure may reproduce existing asymmetries in urban investment and data capacity [6].

3. Post-Training, End-to-End Learning, and Complex Urban Scenario Generation

Post-training in end-to-end autonomous driving has become an important focus because pre-trained models often fail to generalize to rare, interactive, or long-tail urban situations [7]. End-to-end driving systems map sensor inputs to driving actions through learned representations, which makes them adaptable but also difficult to interpret and verify. Post-training can refine these models using targeted demonstrations, simulated rollouts, preference data, or reinforcement learning signals. In complex urban scenarios, the most valuable post-training examples are not ordinary lane-following episodes. They are cases involving occluded pedestrians, aggressive merging, double-parked vehicles, temporary signage, jaywalking, and mixed traffic that includes bicycles, scooters, and informal transport. These scenarios are difficult to collect at scale because they are rare, ethically sensitive, and context-dependent. Digital twins can help generate them, but the generation process must be carefully designed to avoid unrealistic or biased distributions [7].

Simulation platforms have provided the technical basis for much of this work. CARLA, for example, offers an open urban driving simulator with configurable sensors, maps, weather, and traffic actors, enabling reproducible experimentation in complex environments [8]. Multimodal datasets such as nuScenes and the Waymo Open Dataset provide real-world sensor data with annotations for perception, tracking, and prediction, supporting the development and evaluation of driving models [9] [10]. These resources are essential, but they also reveal a tension between scale and specificity. Large datasets improve general perception, while local digital twins are better suited to capturing the particular geometry, culture, and regulatory context of a city. A validation regime that relies only on global datasets may miss local practices, such as informal crossing patterns or unusual traffic enforcement, that shape real-world risk.

Imitation and reinforcement learning have been widely used to connect perception with control. Conditional imitation learning allows driving policies to be conditioned on navigational commands, improving behavioral diversity and controllability [11]. Controllable imitative reinforcement learning combines demonstration data with reward-driven exploration, which can improve performance in interactive settings [12]. When applied within a digital twin, these methods can be used for post-training by generating counterfactual scenarios and evaluating policy responses under modified conditions. However, the same flexibility that makes these methods useful also makes them vulnerable to reward hacking, distributional shift, and overfitting to simulator artifacts. A model may learn to exploit unrealistic physics, sensor noise patterns, or scenario generator biases rather than acquiring robust driving competence. Digital twin-assisted post-training therefore requires validation protocols that test for simulator-induced shortcuts and that compare simulated behavior with independent real-world observations.

4. System Architecture and Data Governance

The architecture of a digital twin-assisted validation system determines not only performance but also accountability. A typical architecture includes data acquisition from vehicles and roadside infrastructure, a data plane for storage and streaming, a modeling layer for scenario generation and simulation, a post-training layer for model adaptation, and an assurance layer for testing, auditing, and reporting. Each layer can be centralized, distributed, or federated. Centralized architectures simplify version control and regulatory inspection, but they concentrate power and risk. Distributed architectures preserve local control and reduce bandwidth, but they complicate consistency and certification. Federated approaches can train models across multiple sites without sharing raw data, yet they still require shared ontologies, aggregation rules, and mechanisms for detecting poisoned or low-quality updates. These are governance choices as much as engineering choices [13].

Data governance is particularly important because digital twins combine personal, commercial, and public data. Vehicle telemetry may reveal driving behavior, location histories, and passenger information. Roadside cameras may capture pedestrians and license plates. Municipal maps may contain sensitive infrastructure details. Post-training on such data can improve safety, but it can also enable surveillance, re-identification, or commercial exploitation. Effective governance requires purpose limitation, data minimization, retention limits, access control, and independent audit. It also requires clarity about liability when a model trained in a digital twin fails in the physical world. If a city provides data, a vendor supplies the twin, and a manufacturer performs post-training, responsibility can become fragmented across contracts and jurisdictions. Governance frameworks must therefore define traceability from scenario generation to model update to deployment decision [14].

The integration of digital twins with AI safety practices also raises questions about model documentation and change management. A post-trained driving model may be updated frequently, which challenges traditional type approval and certification models that assume stable configurations. Regulators need mechanisms to inspect not only final model weights but also training data provenance, scenario coverage, evaluation results, and rollback procedures. Standards such as the SAE taxonomy for driving automation provide common vocabulary for levels and system boundaries [16], while legal and social analyses of autonomous driving emphasize that technical classification alone cannot resolve responsibility, privacy, or accessibility concerns [15]. Digital twin governance should therefore connect technical metadata with institutional records, enabling auditors to

reconstruct why a model behaved as it did during validation and how that validation related to public road deployment.

5. Robustness, Fairness, and Socio-Technical Trade-Offs

Robustness in digital twin-assisted post-training has multiple meanings. It can refer to the model's resistance to adversarial inputs, sensor degradation, weather changes, and distributional shift. It can also refer to the resilience of the validation infrastructure itself when networks fail, data streams are corrupted, or simulation services are unavailable. AI safety research has identified concrete problems such as negative side effects, reward hacking, scalable oversight, and safe exploration, many of which appear in autonomous driving validation [13]. A model that performs well in a digital twin may still fail when the twin's assumptions break down. For example, a twin may model pedestrians as rational agents, but real pedestrians may act on social cues, local norms, or momentary distraction. Robustness therefore requires deliberate stress testing against model mismatch, not just parameter variation within the same simulator.

Fairness is equally important and often underdeveloped in autonomous driving validation. Complex urban scenarios are not distributed evenly across neighborhoods. Wealthier districts may have better road markings, clearer signage, more consistent enforcement, and more comprehensive digital infrastructure. Marginalized neighborhoods may have poorer road maintenance, informal crossings, mixed-use streets, and different patterns of pedestrian and commercial activity. If digital twins are built primarily in well-instrumented areas, post-training may optimize for those conditions and underperform elsewhere. This is not only a technical generalization problem but a distributive justice problem. Fairness requires representative scenario collection, community consultation, and evaluation metrics that disaggregate performance by location, road user type, time of day, and socio-economic context. Without such measures, validation may improve average safety while widening disparities in actual protection [19].

There are also trade-offs between transparency and security, and between innovation and precaution. Detailed digital twins can support open auditing, but they may also expose sensitive infrastructure or enable adversarial scenario discovery. Post-training can adapt quickly to new urban conditions, but rapid updates can undermine regulatory oversight and public trust. A socio-technical approach recognizes that these tensions cannot be resolved by a single technical fix. They require institutional design: independent validation bodies, multi-stakeholder scenario committees, public reporting, and channels for affected communities to contest model behavior. The goal is not to eliminate trade-offs but to make them visible, negotiable, and accountable [14] [15].

6. Deployment, Infrastructure, Sustainability, and Policy

Deployment of digital twin-assisted validation depends on physical and digital infrastructure that is unevenly distributed. Cities need reliable power, connectivity, computing, sensor maintenance, and skilled personnel. Cloud-based twins can reduce local hardware requirements, but they depend on network latency and data transfer costs. Edge-based twins can support real-time safety functions, but they require capital investment and ongoing maintenance. In regions with limited infrastructure, lightweight twins, open datasets, and shared validation services may be more feasible than full metropolitan replicas. Policy should therefore avoid a one-size-fits-all certification model that implicitly favors wealthy jurisdictions. Instead, it can support tiered validation pathways, shared public-interest

infrastructure, and interoperability standards that allow smaller cities to participate without surrendering local control.

Sustainability is another dimension. Digital twins can reduce the need for physical road testing, which lowers emissions, congestion, and risk to test drivers and the public. However, large-scale simulation, data storage, and continuous sensor streaming consume energy and materials. The net sustainability benefit depends on how efficiently the twin is used and whether it substitutes for or merely adds to physical testing. If digital twin validation becomes a parallel requirement rather than a replacement for road testing, the environmental and financial costs may increase. Policy instruments such as lifecycle assessment, energy-aware scheduling, and reuse of scenario assets can help align validation with sustainability goals. Urban planners should also consider how digital twin infrastructure can serve multiple public purposes, such as traffic safety analysis, emergency response planning, and accessibility mapping, rather than remaining a proprietary asset for autonomous vehicle vendors [18].

Regulatory frameworks are beginning to address these issues. Incident reporting requirements for automated driving systems create data streams that can inform digital twins and post-training, but they also raise privacy and liability questions [17]. The European Union's artificial intelligence regulation proposal introduces risk-based obligations for high-risk AI systems, which may apply to autonomous driving components and their validation environments [18]. International organizations such as the OECD have emphasized human-centered AI principles, including transparency, accountability, and inclusive growth [19]. These instruments provide direction, but they do not yet specify how digital twin evidence should be weighted, audited, or compared with physical testing. Future policy should define standards for scenario provenance, simulation fidelity, model update traceability, and independent evaluation. It should also clarify when digital twin results are sufficient for deployment and when physical validation remains necessary.

7. Evaluation, Assurance, and Cross-Domain Lessons

Evaluation of digital twin-assisted post-training must go beyond aggregate success rates. It should examine scenario coverage, realism, adversarial robustness, uncertainty calibration, and failure severity. A model that avoids collisions in common scenarios but fails catastrophically in rare interactions may appear safe in average metrics while remaining unacceptable in practice. Assurance should therefore use layered evidence: simulation for breadth, closed-course testing for controlled edge cases, public road testing for ecological validity, and incident reporting for post-deployment monitoring. Planning and decision-making research in autonomous vehicles has long emphasized the need to handle uncertainty, interaction, and dynamic constraints [20]. Digital twins can support this by generating interactive scenarios that stress the planner rather than only the perception system. However, the evaluation must include human factors, because the behavior of pedestrians, cyclists, and other drivers is shaped by expectations, trust, and social signaling.

Cross-domain comparisons offer useful lessons. Aerospace validation relies on rigorous certification, redundancy, and independent oversight, but it also benefits from relatively bounded operating conditions. Industrial digital twins emphasize lifecycle monitoring and predictive maintenance, but they operate in controlled environments with clear asset boundaries [5]. Critical infrastructure assurance combines technical standards with resilience planning and public accountability. Autonomous driving sits at the intersection of all three: it is safety-critical, socially embedded, and highly variable. A digital twin-assisted validation regime should therefore adopt the strongest elements of each domain, including independent

verification, continuous monitoring, transparent incident analysis, and community engagement. It should avoid the temptation to treat simulation realism as a proxy for legitimacy. Realism is necessary but not sufficient for trust.

Assurance also requires organizational capacity. Regulators need expertise in machine learning, simulation, cybersecurity, and urban planning. Municipalities need resources to maintain data infrastructure and participate in standards development. Manufacturers need incentives to document negative results and share safety-relevant scenarios without exposing proprietary secrets. Third-party auditors need access to enough data and model behavior to conduct meaningful review. These requirements imply new institutions, or at least new functions within existing institutions. Public-private partnerships can help, but they must be designed to prevent regulatory capture and to ensure that safety evidence is available to those affected by deployment. The legitimacy of digital twin-assisted post-training depends on whether it is seen as a shared safety infrastructure or as a private optimization tool.

8. Conclusion

Digital twin-assisted post-training offers a promising pathway for validating autonomous driving systems in complex urban scenarios. It can generate rare and dangerous situations, adapt models to local conditions, reduce physical testing burdens, and support continuous assurance. Yet its value depends on how it is architected and governed. A digital twin is not a neutral mirror of the city. It encodes choices about what to sense, what to model, what to ignore, and who gets to decide. Post-training amplifies those choices by turning validation data into model behavior. If digital twins are deployed without independent oversight, transparent provenance, and inclusive scenario design, they may improve simulation metrics while reproducing existing inequalities and creating new forms of unaccountability.

The paper has argued for a socio-technical approach that treats digital twins as governed validation infrastructures rather than merely advanced simulators. Such an approach requires architectural pluralism, with edge, cloud, and federated systems matched to local constraints. It requires data governance that protects privacy, clarifies liability, and enables audit. It requires robustness testing against simulator artifacts, distributional shift, and adversarial manipulation. It requires fairness metrics that disaggregate safety outcomes across neighborhoods and road user groups. It requires sustainability assessments that weigh the benefits of reduced road testing against the energy and material costs of digital infrastructure. Finally, it requires policy frameworks that specify when digital twin evidence is sufficient, how post-trained models are certified, and how affected communities can participate in decisions about deployment.

The future of autonomous driving validation will likely involve a hybrid of digital and physical evidence, human and automated oversight, and local and global knowledge. Digital twin-assisted post-training can be a constructive part of that future if it is designed as a public-interest capability rather than a proprietary shortcut. The central challenge is not only to make autonomous vehicles safer in complex urban environments, but to make the validation process itself more transparent, equitable, and accountable. That challenge is as much institutional as it is technical, and it should shape the next generation of validation architectures, standards, and policies.

References

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

2. 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. <https://doi.org/10.1016/j.tra.2016.09.010>
3. Koopman, P., & Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. *IEEE Intelligent Transportation Systems Magazine*, 9(1), 90-96. <https://doi.org/10.1109/MITS.2016.2583491>
4. Grieves, M., & Vickers, J. (2017). Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. In F.-J. Kahlen, S. Flumerfelt, & A. Alves (Eds.), *Transdisciplinary perspectives on complex systems* (pp. 85-113). Springer. https://doi.org/10.1007/978-3-319-38756-7_4
5. Tao, F., Zhang, H., Liu, A., & Nee, A. Y. C. (2019). Digital twin in industry: State-of-the-art. *IEEE Transactions on Industrial Informatics*, 15(4), 2405-2415. <https://doi.org/10.1109/TII.2018.2873186>
6. Kaur, M. J., Mishra, V. P., & Maheshwari, P. (2020). The convergence of digital twin, IoT, and machine learning: Transforming data into action. In M. Farsi, A. Daneshkhah, A. Hosseinian-Far, & H. Jahankhani (Eds.), *Digital twin technologies and smart cities* (pp. 3-17). Springer. https://doi.org/10.1007/978-3-030-18732-3_1
7. Yang, Ruining, et al. "Post-Training in End-to-End Autonomous Driving." arXiv preprint [arXiv:2607.08072](https://arxiv.org/abs/2607.08072) (2026).
8. 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).
9. 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* (pp. 11621-11631). <https://doi.org/10.1109/CVPR42600.2020.01164>
10. Sun, P., Kretschmar, 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* (pp. 2446-2454). <https://doi.org/10.1109/CVPR42600.2020.00252>
11. 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. <https://doi.org/10.1109/ICRA.2018.8460487>
12. Liang, X., Wang, T., Yang, L., & Xing, E. (2018). CIRL: Controllable imitative reinforcement learning for vision-based self-driving. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 584-599). Springer. https://doi.org/10.1007/978-3-030-01234-2_36
13. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint [arXiv:1606.06565](https://arxiv.org/abs/1606.06565).

14. Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
15. Maurer, M., Gerdes, J. C., Lenz, B., & Winner, H. (Eds.). (2016). *Autonomous driving: Technical, legal and social aspects*. Springer. <https://doi.org/10.1007/978-3-662-48847-8>
16. SAE International. (2021). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles (J3016_202104)*. SAE International.
17. National Highway Traffic Safety Administration. (2021). *Standing general order on crash reporting for incidents involving ADS and Level 2 ADAS*. U.S. Department of Transportation.
18. European Commission. (2022). *Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. European Commission.
19. OECD. (2019). *Artificial intelligence in society*. OECD Publishing. <https://doi.org/10.1787/eedfee77-en>
20. Schwarting, W., Alonso-Mora, J., & Rus, D. (2018). Planning and decision-making for autonomous vehicles. *Annual Review of Control, Robotics, and Autonomous Systems*, 1, 187-210. <https://doi.org/10.1146/annurev-control-060117-105157>