SafeDrive-DigitalTwin: Digital Twin-Based Risk Prediction and Scenario Generation with Unified World Models for Autonomous Vehicles

Chetan M. Tripathi
School of Computing, Clemson University, Clemson, SC, USA.
cmtripathi@clemson.edu

Xavier Woods
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
xavierwork@ucf.edu

Yuanzhang Tan
Department of Computer Science, Binghamton University, Binghamton, NY, USA.
tanyuan@binghamton.edu

Abstract

Ensuring the safety of autonomous vehicles operating in complex, open-ended environments remains one of the most formidable challenges in contemporary systems engineering. Physical road testing alone cannot economically or temporally cover the combinatorial space of rare and hazardous scenarios. This paper presents SafeDrive-DigitalTwin, a conceptual architecture that fuses high-fidelity digital twin infrastructure with unified world models to enable continuous risk prediction and automated safety-critical scenario generation. The digital twin serves as a live mirror of vehicle state, environmental dynamics, and infrastructure conditions, while a unified generative world model jointly captures perception, planning, and scene evolution across diverse operational design domains. The work provides a system-level analysis of this integration, examining architectural trade-offs between edge-side lightweight twins and cloud-side comprehensive simulation engines, data governance models that span vehicle manufacturers and public agencies, and the sustainability implications of large-scale generative inference. Furthermore, the paper addresses robustness against sensor degradation and adversarial perturbations, fairness in risk allocation across demographic and geographic contexts, and the policy frameworks required to translate virtual safety evidence into regulatory acceptance. By emphasizing structural interdependencies rather than algorithmic minutiae, this discussion illuminates a pathway toward trustworthy, continuously validated autonomous mobility systems. The SafeDrive-DigitalTwin paradigm proposes that persistent synchronization of physical fleets with their digital counterparts, augmented by generative scenario expansion, can fundamentally reshape the safety assurance lifecycle.

Keywords

digital twin, autonomous vehicles, world model, risk prediction, scenario generation, safety assurance, socio-technical infrastructure.

1. Introduction

The deployment of autonomous vehicles at scale confronts developers and regulators with a dilemma that is as much systemic as it is algorithmic. Physical road testing, even when conducted over millions of miles, yields an extremely sparse sampling of the hazardous event space that determines public safety outcomes. Rare combinations of weather, lighting, erratic human road users, degraded sensor modalities, and map inaccuracies can elude both structured track tests and stochastic on-road exposure. The limitations of purely physical validation have motivated a strong pivot toward simulation-based testing, yet conventional simulation environments often lack the dynamic fidelity necessary to generate genuinely surprising and physically plausible edge cases. In parallel, the growing maturity of digital twin concepts and generative world models opens a new design space in which continuous, synchronous representations of vehicles and their surroundings can serve as living laboratories for risk prediction and scenario evolution. The SafeDrive-DigitalTwin framework advanced in this paper seeks to unify these threads, proposing a socio-technical architecture in which every deployed autonomous vehicle remains persistently tethered to a high-fidelity digital representation that is recursively updated and used for forward-looking risk assessment.

The intellectual lineage of digital twins traces its origin to complex product lifecycle management, where a virtual counterpart of a physical artifact captures geometry, physics, and operational history to predict failure modes [1]. Subsequent research has broadened the digital twin paradigm to encompass enabling technologies such as the Industrial Internet of Things, edge-cloud orchestration, and high-bandwidth telemetry, establishing the foundational infrastructure needed to mirror entire vehicle fleets in near real time [2]. When this paradigm is transplanted into the context of autonomous driving, the digital twin becomes not merely a static replica but a probabilistic projection engine that can simulate thousands of possible futures conditioned on current state and learned behavioral priors. However, the fidelity and coverage of such projections are critically dependent on the underlying world model that captures the joint dynamics of perception, planning, and environment interaction. Early formulations of world models demonstrated how an agent could learn compressed spatiotemporal representations to imagine rollouts and perform planning inside its own latent space [3], seeding the idea that generative models could serve as internal simulators. More recently, the autonomous driving community has adapted these principles to construct scene-level world models that jointly generate sensor observations, predict other agents' trajectories, and synthesize realistic driving scenarios for exhaustive verification.

The introduction of unified world models that simultaneously handle understanding, planning, and generation for driving tasks represents a qualitative shift in capability. Instead of relying on a heterogeneous collection of distinct modules for object detection, motion prediction, and scenario sampling, a unified architecture internalizes a holistic representation that can be queried for diverse safety analyses without brittle interface handoffs. When such a world model is embedded within a digital twin loop that continuously ingests fleet telemetry, the system can preemptively identify emerging risks that are invisible to any single-vehicle onboard logic. This paper explores the ramifications of that synthesis from the perspective of large-scale system design, focusing on how architectural decisions distribute computational load, data authority, and ethical accountability across the human and institutional actors that shape autonomous mobility. Rather than presenting algorithmic details, the discussion foregrounds the structural trade-offs, governance challenges, sustainability concerns, and policy adaptations that must be navigated to transform the SafeDrive-DigitalTwin vision into a trustworthy public infrastructure.

2. Background and Related Work

Digital twin technology has undergone a rapid conceptual expansion from its origins in manufacturing to complex cyber-physical systems. In the canonical formulation by Grieves and Vickers, a digital twin comprises a physical entity, a virtual representation, and the connective data flows that ensure the virtual component mirrors the physical object throughout its lifecycle [1]. Transferring this model to autonomous driving implies not only capturing the three-dimensional geometry and kinematic state of a vehicle but continuously assimilating its software stack behavior, perception confidence distributions, control actuator health, and the surrounding dynamic scene graph. Research on the enabling technologies layer has detailed the sensing, communication, and computing substrates that make such real-time mirroring feasible, highlighting the tension between edge-based lightweight twins optimized for low-latency local decisions and cloud-based comprehensive twins that maintain global fleet awareness but incur higher synchronization delays [2]. These trade-offs are not merely technical; they shape business models, data ownership structures, and the degree of centralized oversight that a regulator can exercise over an autonomous fleet.

In parallel with digital twin advances, the concept of a learned world model has emerged as a powerful abstraction for agents operating in partially observable stochastic environments. The seminal work by Ha and Schmidhuber demonstrated that a compact recurrent neural network could learn to represent the dynamics of a visual environment and generate prospective futures, enabling an agent to train entirely inside its own imagined rollouts [3]. Subsequent efforts have sought to adapt this principle to the multifaceted demands of driving, where the model must harmonize visual perception, semantic scene understanding, behavioral prediction of heterogeneous traffic participants, and ego-motion planning within a single cohesive framework. The unification of these functions is particularly relevant because conventional modular pipelines propagate errors across interfaces, often resulting in implausible synthetic scenarios that fail to challenge the autonomy stack in meaningful ways. Instead, a unified world model can learn joint distributions over observations, actions, and outcomes, generating scenarios that respect the causal structure of driving interactions.

Safety validation for autonomous vehicles has historically drawn on scenario-based methodologies that decompose the operational domain into a manageable set of logical scenario descriptions. Criticality analysis techniques quantify the severity of scenarios by measuring risk proxies such as time-to-collision, post-encroachment time, or more sophisticated probabilistic reachability metrics, enabling the identification of scenarios that lie at the boundary of acceptable safety margins [4]. Risk prediction, as surveyed extensively by Lefèvre et al., encompasses both short-horizon collision risk estimation based on motion prediction and longer-horizon situational risk models that incorporate intent recognition and interaction-aware forecasting [5]. These models, however, are generally designed to operate onboard a vehicle with limited computational and observability constraints, whereas a digital twin environment can leverage additional sensor perspectives, infrastructure data, and post-hoc knowledge to generate more comprehensive risk assessments. Scenario generation for autonomous vehicle testing has matured into a distinct field, with ontology-driven approaches systematically varying static environment attributes, dynamic actor behaviors, and sensor degradation effects [6]. Yet the combinatorial explosion inherent in full-factorial variation demands generative models that can intelligently concentrate sampling effort on safety-critical regions of the scene space. It is precisely at this intersection that unified world models offer transformative potential. The UniDrive-WM framework recently demonstrated that a

single architecture can simultaneously perform understanding, planning, and generation for driving applications, yielding coherent scene evolution that respects physical constraints while surfacing rare but plausible hazardous configurations [7]. This capability is foundational for a digital twin system that must anticipate risks not yet recorded in any fleet log.

3. Digital Twin Architecture for Autonomous Vehicle Risk Prediction

Designing a digital twin architecture for fleet-wide risk prediction requires navigating a hierarchy of structural decisions that extend far beyond the selection of a particular neural network architecture or simulation engine. At the first tier of abstraction, the system must define the fidelity spectrum along which vehicle, environment, and actor models are instantiated. A full-order digital twin might incorporate detailed multibody dynamics of the vehicle chassis, physically based sensor models that simulate raw LiDAR point clouds and camera image formation, and a microscopic traffic simulator that reproduces human driving behavior with validated psychometric parameters. Such richness is desirable for exposing subtle failure modes, such as those arising from sensor obstruction when a specific sun angle coincides with a puddle on the road, but it imposes formidable computational costs and challenges the scalability of continuous fleet synchronization. In contrast, a reduced-order twin that abstracts away low-level physics and sensor phenomenology can operate with lower latency and resource consumption, enabling wider coverage at the expense of potentially missing edge cases that depend on precise physical interactions. The SafeDrive-DigitalTwin concept envisions a tiered architecture in which each physical vehicle is shadowed by a lightweight edge-side twin that performs real-time anomaly detection and local risk assessment, while a cloud-side cluster of high-fidelity twins selectively deepens the analysis for scenarios flagged as unusual or high-risk by the fleet-level monitoring system.

A second tier of architectural deliberation concerns the temporality of the digital twin's projection. Risk prediction is not a static classification but a dynamic process that must anticipate how a scene will unfold under both the ego vehicle's planned trajectory and the uncertain behavior of surrounding entities. The integration of a unified world model directly addresses this requirement by providing a learned generative engine that can sample plausible future evolutions conditioned on the current digital twin state. The advantage of embedding a world model inside the twin infrastructure rather than relying on a separate motion prediction module is that the world model encodes covariances between the ego vehicle's own actions and the responses of other agents, thereby generating counterfactual futures that are coherent across all interacting entities. For example, if the twin simulates a scenario in which the ego vehicle performs an assertive lane change, the world model simultaneously synthesizes the braking or swerving reactions of nearby vehicles, the resulting camera and LiDAR depictions, and the associated risk metrics. This holistic generation avoids the unrealistic decoupling that often plagues modular simulation pipelines.

The third architectural dimension involves the interface between the digital twin and the physical vehicle's onboard logic. One possibility is to keep the twin entirely separate, using it only for offline risk auditing and scenario mining. A more ambitious design closes the loop by feeding digital twin risk signals back to the vehicle in near real time, enabling proactive mitigation such as reducing speed when the twin predicts an elevated probability of a previously unseen conflict pattern. Such closed-loop operation forces a careful reevaluation of timing constraints, communication reliability, and safety fallbacks. The system must guarantee that a stale or corrupted risk signal from the twin cannot precipitate a hazardous

action by the vehicle. Consequently, the architecture incorporates a formal safety layer that validates twin-derived recommendations against hard constraints derived from reachability analysis and invariant monitoring, ensuring that the digital twin remains an advisor rather than an unsupervised controller. The use of digital twin enablers such as secure time-sensitive networking, distributed ledger-based state synchronization, and gradual trust calibration over multiple operational phases becomes essential to realizing this architecture at scale [19].

4. Unified World Models for Scenario Generation and Understanding

The role of a unified world model in the SafeDrive-DigitalTwin framework extends beyond risk prediction into the systematic discovery of safety-critical scenarios that are absent from recorded fleet data. Traditional scenario generation pipelines rely on rule-based variations of logical scenarios or on replaying and perturbing collected trajectories, which biases coverage toward regions of the state space that the fleet has already visited. A unified generative world model, trained on diverse urban, highway, and rural driving data, can extrapolate beyond the empirical distribution by exploiting the underlying causal and physical structure that it internalizes. The model's ability to generate coherent sensor observations, lane topologies, and dynamic agent behaviors from a compact latent representation allows it to synthesize scenarios in which the interplay of geometry, weather, and human irrationality creates hazardous situations that have never been recorded but are physically and statistically plausible. For instance, the model can combine a previously encountered roundabout geometry from a European city with the aggressive cut-in behavior typical of a specific North American highway segment and a rare fog condition observed in a coastal region, all while ensuring that the resulting scene remains consistent with sensor physics and traffic norms.

A critical system-level property of this scenario generation capability is its steerability toward regions of the state space that stress specific weaknesses of the autonomous driving policy. By conditioning latent variables on risk heuristics such as the inverse of the policy's confidence or the magnitude of prediction errors in a shadow mode deployment, the digital twin can concentrate generative effort on the boundary of the operational design domain where the autonomy stack is most fragile. This directed exploration, often referred to as adversarial scenario generation when framed through a game-theoretic lens, must be balanced against the risk of producing scenarios that are so extreme that they fall outside any reasonable definition of the operational domain and thus provide little transferable safety evidence. The unified world model's internal regularization, learned from diverse but still real-world-grounded data, provides a natural mechanism to avoid pathological adversarial cases that violate physical laws or commonsense behavioral norms. The challenge lies in calibrating this steerability without inadvertently creating a coverage gap in regions that the model, due to inherent training biases, fails to represent.

Equally important is the model's capacity for long-horizon generation that respects the temporal credit assignment structure of driving interactions. A collision or near-miss event often results from a cascade of decisions over tens of seconds, where early subtle actions by surrounding vehicles set up a context that later becomes hazardous. A world model with a robust memory mechanism can generate these extended precursors, enabling the digital twin to identify not only the final conflict but the chain of events that made it possible. This is essential for moving beyond reactive safety metrics toward a proactive understanding of situational risk that can inform both vehicle planning heuristics and infrastructure design. The integration of scenario generation with the digital twin's continuous state ingestion also means that the generated scenarios can be anchored in the actual physical state of a specific

vehicle at a specific moment, producing personalized risk assessments that incorporate the vehicle's current sensor calibration status, tire wear, and software version.

5. System Integration: SafeDrive-DigitalTwin Platform

Realizing the SafeDrive-DigitalTwin concept requires a platform architecture that orchestrates the interactions among fleet vehicles, edge computing nodes, cloud resources, and the unified world model in a manner that satisfies the often-competing demands of latency, cost, data privacy, and regulatory transparency. The platform adopts an event-driven, microservice-oriented design in which each vehicle publishes a compressed stream of state telemetry, including raw and processed sensor packets, localization estimates, and planner intermediate representations. Edge gateways, deployed at cellular base stations or roadside units, instantiate lightweight digital twins for vehicles in their geographic sector and perform the first stage of anomaly screening using reduced-order world model rollouts. Scenarios that cross a predefined risk threshold are escalated to a cloud-based orchestration layer that assigns them to a pool of high-fidelity twin workers, which can run detailed physics-based simulations augmented by the unified world model's generative capabilities. This hierarchical escalation mirrors the human immune system's interplay between innate and adaptive responses, providing a cost-effective balance between ubiquitous monitoring and deep investigation.

Data governance within the platform must address the multi-stakeholder nature of autonomous mobility. The vehicle manufacturer owns the detailed software stack and sensor specifications, the fleet operator holds a stream of operational data, municipal authorities possess infrastructure state information, and regulators require verifiable safety evidence without necessarily accessing proprietary sensor raw data. The platform's data fabric is therefore organized around a set of privacy-preserving protocols that enable collaborative risk assessment without consolidating all raw data in a single repository. Federated learning and secure multiparty computation allow the unified world model to be trained across organizational boundaries while keeping sensitive data local, while zero-knowledge proofs can attest that a manufacturer has faithfully executed a regulatory scenario suite on its digital twin without revealing the underlying vehicle design parameters. These mechanisms are not merely cryptographic curiosities; they are foundational to building the institutional trust required for a digital twin-based safety case to be accepted by regulators and the public.

The integration of the unified world model with the digital twin's scenario generation loop also demands careful attention to versioning and configuration management. The world model itself will undergo periodic retraining as new fleet data arrives, potentially shifting its generative distribution and altering the risk assessments it produces for the same vehicle state. A robust platform must maintain a provenance trail that links each risk prediction and generated scenario to the exact model version, twin fidelity settings, and environmental data sources used, enabling post-hoc audits and root cause analyses when a physical safety incident occurs that was not predicted by the digital twin. This auditability requirement intersects with emerging standards for safety assurance of artificial intelligence, including the concept of an assurance case that makes explicit the claims, evidence, and assumptions underpinning a safety argument [8]. The platform's logging and traceability subsystems are thus not auxiliary functions but central pillars of the safety argument itself.

6. Governance, Fairness, and Policy Implications

The deployment of a digital twin-based risk prediction and scenario generation system at the scale of a public road network inevitably raises profound questions of governance, fairness, and legal accountability that transcend the narrower engineering focus on false positive and false negative rates. One of the most pressing concerns is the distributional fairness of the risk predictions themselves. If the unified world model is trained predominantly on data collected from affluent urban districts with well-maintained infrastructure and consistent lane markings, the generated scenarios and risk assessments may systematically underestimate the hazards present in rural or under-resourced neighborhoods where road conditions, pedestrian behaviors, and signage differ significantly. Such a bias could lead to a form of algorithmic redlining, wherein the digital twin sanctions more permissive autonomous vehicle behavior in well-represented areas while imposing overly conservative restrictions in underrepresented ones, thereby exacerbating existing mobility inequities. Fairness-aware training regimes and context-adaptive calibration, informed by the expansive literature on machine learning fairness [12], must thus be embedded into the world model's development lifecycle, with explicit performance monitoring stratified by demographic and geographic variables.

Beyond fairness, the governance framework must delineate the roles and responsibilities of the various institutional actors that come into contact with the SafeDrive-DigitalTwin system. The digital twin becomes a shared cognitive artifact that influences decisions made by vehicle manufacturers, fleet management algorithms, municipal traffic control centers, and insurance underwriters. When a crash occurs that the digital twin had flagged as an elevated-risk scenario but the vehicle failed to mitigate, questions of liability become entangled with the interpretation of the risk signal. Was the signal sufficiently timely and actionable? Was its communication channel appropriately reliable? Did the vehicle's safety layer correctly adjudicate it against competing objectives? These questions cannot be resolved solely through technical specifications; they require a regulatory framework that recognizes the digital twin as a form of safety-related information infrastructure, analogous to air traffic control advisories in aviation. The National Highway Traffic Safety Administration's Federal Automated Vehicles Policy [13] provides an initial scaffolding, but it does not yet address the evidentiary weight that a real-time digital twin risk assessment should carry in post-incident investigations or regulatory compliance determinations. Evolving these frameworks necessitates a close dialogue between systems engineers, ethicists, and transportation lawyers.

A related policy dimension involves the public acceptability of a system that continuously mirrors the motion and behavior of every vehicle in a fleet. Even with strong anonymization and differential privacy guarantees, the aggregative power of a fleet-wide digital twin can reveal sensitive patterns about population movement, socioeconomic activity, and individual travel habits. The European Commission's ethics guidelines for trustworthy AI [20] emphasize the principles of transparency, privacy, and human oversight, which must be operationalized within the platform's data handling architecture. One approach is to implement tiered access controls where the granularity of mirroring depends on the social mission of the querying entity: a municipal traffic authority may receive aggregate flow predictions without per-vehicle identifiers, while a safety investigator with judicial authorization may access full-fidelity twin logs for a specific incident. Designing these tiered systems without creating opaque loopholes is a non-trivial governance challenge that will shape the public's willingness to accept the pervasive instrumentation implied by the SafeDrive-DigitalTwin concept.

7. Sustainability and Deployment Scalability

The computational demands of maintaining a fleet-wide digital twin that continuously invokes a unified generative world model for risk prediction and scenario generation raise significant sustainability concerns that cannot be relegated to an afterthought in system design. Training large world models with multimodal perception, planning, and generation capacities already consumes substantial energy and hardware resources, and the inference budget required to run thousands of parallel twin instances across a city-scale fleet can quickly rival the energy footprint of the vehicles themselves. If the carbon cost of the digital twin infrastructure outweighs the safety and efficiency gains it enables, the system's net societal benefit becomes questionable. Mitigating this impact requires a combination of model compression, sparse activation techniques, and energy-proportional computing architectures, as well as a structural commitment to measuring and reporting the full lifecycle environmental impact of the SafeDrive-DigitalTwin platform. Sustainability frameworks that have been developed for smart city digital twins, such as those emphasizing the circular use of cloud resources and heat reuse from data centers [15], can be adapted to the autonomous vehicle context.

Scalability is not solely a matter of computational throughput; it also encompasses the organizational and operational complexity of onboarding heterogeneous vehicle platforms from multiple manufacturers onto a common digital twin substrate. Each vehicle model possesses a distinct sensor suite, chassis dynamics, and software architecture, which means that a universal twin must accommodate a degree of modularity and customization. The platform can address this through the concept of a twin manifest that formally describes the fidelity parameters, sensor models, and behavioral interfaces of each vehicle type, allowing the orchestration layer to instantiate the appropriate twin configuration on demand. Standardization efforts, potentially driven by industry consortia such as the AUTOSAR adaptive platform or emerging digital twin interoperability standards from the Industrial Internet Consortium, will be essential to preventing fragmentation and vendor lock-in. The cost of developing and maintaining the diverse model libraries required for broad fleet coverage also introduces economic sustainability considerations, as the burden of twin development may fall disproportionately on smaller manufacturers and create barriers to market entry.

Deploying the SafeDrive-DigitalTwin system in a geographically distributed manner further introduces challenges related to network heterogeneity and resilience. Urban canyons, tunnels, and rural areas with sparse cellular coverage can interrupt the telemetry feed that keeps the digital twin synchronized, leading to periods of stale or divergent state. The platform must therefore incorporate graceful degradation mechanisms, where the onboard vehicle systems continue to operate with fallback safety strategies when the twin link is lost, and the twin itself can re-synchronize efficiently upon reconnection using checkpointed state snapshots and log replay. The unified world model's generative capabilities can also be leveraged to fill short gaps in telemetry by plausibly interpolating missing observations, but this function must be accompanied by clear uncertainty quantification so that downstream risk assessments do not treat synthetic data as ground truth. These operational resilience properties are critical for maintaining public trust and safety across the full diversity of road infrastructure.

8. Robustness, Security, and Ethical Considerations

Infrastructure-critical systems that fuse physical actuation with machine learning inference must be engineered to withstand a spectrum of adversarial pressures, whether they arise from environmental perturbations, sensor faults, or deliberate cyberattacks. In the SafeDrive-

DigitalTwin architecture, the unified world model represents a particularly attractive attack surface because it can be targeted through data poisoning during training, adversarial inputs crafted to induce catastrophic scenario generation, or model inversion attacks that seek to reconstruct private training data from the model's parameters. Defending against these threats requires a layered security strategy that encompasses robust training techniques such as adversarial augmentation and certified defenses, runtime input monitoring that detects distributional shifts symptomatic of attack, and strict isolation between the world model's inference pipeline and the vehicle's safety-critical control bus. The principle of least privilege must guide every interface through which the digital twin communicates with the physical fleet. Moreover, the scenario generation loop must be designed to prevent a feedback cycle in which an attacker who compromises the world model can force the digital twin to generate misleading safety evidence that masks real vulnerabilities. Formal verification of the safety layer that gates twin-to-vehicle recommendations, possibly drawing on reachability-based online verification methods [18], can provide assurance that even a corrupted digital twin cannot directly cause an unsafe control action.

Robustness to environmental variability is equally critical. The unified world model must generalize across weather conditions, illumination changes, and sensor degradation patterns that were rare or absent in the training distribution. The digital twin's continuous synchronization with physical vehicles provides a unique opportunity to detect and characterize such covariate shifts at fleet scale, enabling the world model to be updated or fine-tuned before a large number of vehicles encounter the novel condition. The challenge is to distinguish between a true environmental shift that requires model adaptation and a transient sensor failure that would pollute the model with spurious correlations. To this end, the platform can employ a dedicated anomaly detection subsystem that cross-references telemetry from multiple co-located vehicles to disambiguate environmental from vehicle-specific artifacts.

Ethical considerations extend beyond fairness in risk distribution to encompass the very nature of scenario generation as a tool for shaping the safety envelope of autonomous vehicles. When the digital twin generates a hazardous scenario and exposes a weakness in the planning policy, the manufacturer faces an ethical obligation to address the vulnerability before deploying an update, even if the scenario is deemed extremely low probability. The temptation to downweight or suppress inconvenient but valid generated scenarios in the interest of accelerating deployment constitutes a new form of ethical hazard that existing regulatory frameworks are poorly equipped to address. The Moral Machine experiment [11] demonstrated that public expectations about autonomous vehicle behavior in ethical dilemmas are culturally conditioned and cannot be resolved through any single optimization metric. Translating such findings into the governance of a digital twin platform requires the establishment of independent oversight bodies with the authority to audit both the scenario generation process and the manufacturer's response to generated risks. Only through such institutional scaffolding can the SafeDrive-DigitalTwin paradigm earn the societal license to operate.

9. Conclusion

This paper has presented SafeDrive-DigitalTwin, a system-level vision that synthesizes digital twin infrastructure with unified world models to enable continuous risk prediction and safety-critical scenario generation for autonomous vehicle fleets. By foregrounding architectural trade-offs, governance structures, sustainability pressures, and ethical requirements, the

discussion has sought to situate the technical components within the broader socio-technical fabric that determines whether an autonomous mobility system can be both safe and publicly trusted. The tiered fidelity architecture balances real-time edge monitoring with cloud-based deep scenario exploration, while the embedment of a unified generative world model endows the digital twin with the capacity to extrapolate beyond historical data and anticipate emergent hazards. Fairness-conscious training, privacy-preserving multi-stakeholder data governance, and formal safety gating mechanisms collectively form the institutional backbone that must accompany the computational substrate. As the transportation ecosystem moves toward increasingly automated and interconnected operation, the integration of persistent digital mirrors with generative foresight will become a cornerstone of safety assurance, not as a replacement for physical testing but as a necessary complement that addresses the fundamental coverage limitations of empirical validation. The research agenda ahead is rich with challenges that span systems engineering, machine learning, human factors, and public policy, and meeting them will demand the kind of deep interdisciplinary collaboration that this paper has aimed to exemplify.

References

1. Grieves, M., & Vickers, J. (2017). Digital twin: mitigating unpredictable, undesirable emergent behavior in complex systems. In Transdisciplinary Perspectives on Complex Systems (pp. 85–113). Springer.
2. Fuller, A., Fan, Z., Day, C., & Barlow, C. (2020). Digital twin: enabling technologies, challenges and open research. IEEE Access, 8, 108952–108971.
3. Ha, D., & Schmidhuber, J. (2018). World models. arXiv preprint arXiv:1803.10122.
4. Neurohr, C., Westhofen, L., & Möhlmann, E. (2020). Criticality analysis for the verification and validation of automated vehicles. IEEE Access, 8, 18049–18066.
5. Lefèvre, S., Vasquez, D., & Laugier, C. (2014). A survey on motion prediction and risk assessment for intelligent vehicles. ROBOMECH Journal, 1(1), 1–14.
6. Menzel, T., Bagschik, G., & Maurer, M. (2018). Scenarios for development, test and validation of automated vehicles. In 2018 IEEE Intelligent Vehicles Symposium (IV) (pp. 1821–1827). IEEE.
7. Zhexiao Xiong, Xin Ye, Burhan Yaman, Sheng Cheng, Yiren Lu, Jingru Luo, Nathan Jacobs, and Liu Ren. UniDrive-WM: Unified understanding, planning and generation world model for autonomous driving. arXiv preprint arXiv:2601.04453, 2026.
8. Koopman, P., & Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. IEEE Intelligent Transportation Systems Magazine, 9(1), 90–96.
9. Bagschik, G., Menzel, T., & Maurer, M. (2018). Ontology based scene creation for the development of automated vehicles. In 2018 IEEE Intelligent Vehicles Symposium (IV) (pp. 1813–1820). IEEE.
10. Li, J., Zhang, J., & Li, K. (2021). A survey on scenario-based testing for automated driving systems. IEEE Access, 9, 87491–87516.
11. Awad, E., Dsouza, S., Kim, R., Schulz, J., Henrich, J., Shariff, A., ... & Rahwan, I. (2018). The Moral Machine experiment. Nature, 563(7729), 59–64.

12. 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.

13. National Highway Traffic Safety Administration. (2016). Federal automated vehicles policy: Accelerating the next revolution in roadway safety. U.S. Department of Transportation.

14. Litman, T. (2020). Autonomous vehicle implementation predictions: Implications for transport planning. Victoria Transport Policy Institute.

15. Bibri, S. E. (2018). The IoT for smart sustainable cities of the future: An analytical framework for sensor-based big data applications for environmental sustainability. Sustainable Cities and Society, 38, 230–253.

16. Minerva, R., Lee, G. M., & Crespi, N. (2020). Digital twin in the IoT context: A survey on technical features, scenarios, and architectural models. Proceedings of the IEEE, 108(10), 1785–1824.

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

18. Althoff, M., & Dolan, J. M. (2014). Online verification of automated road vehicles using reachability analysis. IEEE Transactions on Robotics, 30(4), 903–918.

19. Rasheed, A., San, O., & Kvamsdal, T. (2020). Digital twin: Values, challenges and enablers from a modeling perspective. IEEE Access, 8, 21980–22012.

20. European Commission. (2019). Ethics guidelines for trustworthy AI. High-Level Expert Group on Artificial Intelligence.