

Multimodal Learning for Robust Trajectory Prediction under Diverse Autonomous Driving Scenarios

Aili Nikula

Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.

ailin@unr.edu

Charo Cano

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.

ccano@ku.edu

Dawid Kucharz

Department of Computer Science, University of New Hampshire, Durham, NH, USA.

dawidmail@unh.edu

Abstract

Trajectory prediction has become a central functional layer in autonomous driving systems, where vehicles must anticipate the movement of other road users under uncertainty, occlusion, interaction, and distribution shift. Multimodal learning offers a powerful route to robustness because it can combine cameras, lidar, radar, high-definition maps, vehicle telemetry, and infrastructure-side signals into shared or coordinated representations. Yet the promise of multimodal prediction cannot be reduced to model architecture. It is also a system-level problem involving data curation, sensor calibration, edge-cloud orchestration, lifecycle maintenance, fairness, governance, and sustainability. This paper examines multimodal learning for robust trajectory prediction under diverse autonomous driving scenarios through a socio-technical and infrastructural lens. It analyzes how curatorial discipline, fusion strategy, uncertainty management, and policy-aware assurance shape real-world performance. It considers structural trade-offs among latency, redundancy, coverage, cost, and equity, and it discusses how deployment choices influence robustness across urban, highway, adverse-weather, and mixed-traffic environments. The paper argues that robust prediction is achieved when multimodal architectures are embedded within accountable institutions, maintained data pipelines, resilient compute infrastructures, and transparent evaluation regimes. It concludes with a forward-looking research and governance agenda for resilient, fair, and sustainable autonomous mobility.

Keywords

multimodal learning; trajectory prediction; autonomous driving; robust systems; data curation; fairness; governance; edge computing; sustainability; safety assurance.

1. Introduction

Autonomous driving is often described as a perception-and-control problem, but its safety and social legitimacy depend heavily on prediction. A vehicle must not only detect objects but

also anticipate how pedestrians, cyclists, other vehicles, and emerging micromobility actors will move in the next few seconds. This predictive layer is deeply uncertain because road behavior is shaped by intention, social norms, vehicle dynamics, infrastructure design, and local culture. Multimodal learning has emerged as a major response to this uncertainty, since different sensors and data sources expose complementary aspects of the driving scene. Attention-based sequence models have advanced the representation of temporal dependencies [1], while graph-based map representations have improved the encoding of lane topology and road geometry [2]. Heterogeneous forecasting systems have also demonstrated the value of combining agent dynamics with scene context [3], and social interaction models have shown that nearby agents cannot be treated independently [4]. Together, these developments suggest that robustness arises not from a single sensor or model but from a coordinated system of representations, constraints, and assurances.

However, the transition from benchmark performance to operational reliability exposes structural tensions. Multimodal fusion can improve accuracy when sensors agree, yet it can also amplify correlated failures when calibration drifts, weather degrades visibility, or map data become stale. A system that relies on high-definition maps may perform well in mapped urban corridors but degrade in construction zones or rural roads. A system that depends on vehicle-to-infrastructure communication may achieve cooperative prediction in instrumented areas but lose robustness where infrastructure investment is uneven. These are not merely technical edge cases; they are questions of infrastructure, governance, and equity. Safety assurance frameworks for autonomous vehicles emphasize scalable and verifiable models of safe behavior [18], but trajectory prediction introduces probabilistic and socially contingent elements that resist simple certification. The field therefore requires an interdisciplinary account that connects model design to deployment realities.

This paper argues that multimodal learning for robust trajectory prediction should be understood as a socio-technical infrastructure problem. The central claim is that robustness is produced through the alignment of data curation, architectural fusion, uncertainty calibration, operational monitoring, and public accountability. The paper proceeds by examining multimodal data foundations and curatorial governance, system architecture and fusion trade-offs, robustness across diverse scenarios, fairness and policy implications, deployment and sustainability, evaluation and assurance, cross-domain comparisons, and future research pathways. It draws on multimodal machine learning theory [11] and taxonomy [12] to frame fusion as more than concatenation, and it treats robustness as an emergent property of institutions as well as algorithms. The discussion is system-level rather than narrowly model-centric, because the future of autonomous mobility depends on whether prediction systems can remain dependable, fair, and maintainable across heterogeneous environments.

2. Multimodal Data Foundations and Curatorial Governance

The foundation of multimodal trajectory prediction is not simply the presence of multiple sensors but the governance of multimodal data. Large-scale autonomous driving datasets have made significant contributions by combining cameras, lidar, radar, maps, and vehicle state information [6]. Forecasting-focused datasets have further enriched the field with trajectory annotations and map layers [7], while interactive motion datasets have emphasized the social and cooperative dimensions of road behavior [8]. These resources have enabled models to learn from heterogeneous inputs, but they also reveal structural biases. Data are often collected in specific cities, under particular weather conditions, and with sensor suites that reflect the commercial priorities of a small number of firms. Consequently, models trained on

such data may fail to generalize to dense informal traffic, snow-covered roads, night-time rural environments, or regions with different signage and driving norms. Curatorial governance is therefore not a back-office function; it determines the boundary conditions of robust prediction.

Data curation must address class imbalance, rare-event scarcity, labeling ambiguity, and temporal drift. A recent density-balanced data curation and evaluation framework illustrates how trajectory datasets can be rebalanced and assessed to avoid overrepresenting common behaviors while neglecting safety-critical edge cases [5]. Such an approach is important because prediction models tend to optimize for frequent patterns, such as lane-following, while underperforming on unusual but consequential events, such as sudden jaywalking, erratic merging, or emergency vehicle interactions. Curation also involves deciding what counts as a valid trajectory, how to treat missing data, and how to preserve uncertainty in annotations. If labels are treated as deterministic ground truth, models may become overconfident in situations where human annotators themselves disagree. A robust multimodal system should therefore preserve ambiguity and provenance, allowing downstream modules to distinguish between confident predictions and speculative ones.

Governance of multimodal data also raises questions of privacy, consent, data ownership, and cross-border transfer. Cameras in public space record faces, license plates, and behavioral traces, while infrastructure-side sensors may capture entire intersections. The ethical and societal implications of algorithms, data, and artificial intelligence require attention to power asymmetries, accountability, and public trust [17]. A city may provide data to a private operator, but residents may not have consented to continuous behavioral surveillance. A fleet operator may share anonymized data with a mapping provider, but the anonymization may be reversible when combined with other datasets. These issues affect robustness indirectly because they shape what data can be collected, retained, and used for retraining. If privacy rules prevent the collection of certain rare events, models may remain brittle. If data sharing is fragmented across jurisdictions, evaluation may become inconsistent. Multimodal prediction thus depends on curatorial institutions capable of balancing safety, privacy, and innovation.

3. System Architecture for Multimodal Trajectory Prediction

Architecturally, multimodal trajectory prediction can be organized around early fusion, late fusion, or intermediate cross-modal attention. Early fusion combines raw or lightly processed signals before temporal reasoning, which can capture fine-grained correlations but demands precise calibration and high bandwidth. Late fusion keeps modality-specific encoders separate and combines their outputs at the decision level, which improves modularity and fault isolation but may miss subtle cross-modal interactions. Intermediate fusion, often implemented through attention mechanisms, allows modalities to query one another and form shared representations. The transformer architecture has made such cross-modal attention more practical [1], and multimodal machine learning research has long shown that joint representations can outperform unimodal ones when modalities are complementary rather than redundant [11]. Yet the choice among these strategies is not purely technical. It determines latency, explainability, failure modes, and the ease of certification.

Graph-based representations illustrate the value of structural priors in trajectory prediction. Lane graphs can encode connectivity, curvature, and traffic rules, allowing models to reason about feasible future paths rather than treating the scene as an unstructured point cloud [2]. Heterogeneous forecasting systems can integrate agent type, past motion, and map context into a unified probabilistic model [3]. Social pooling and interaction models further show that

nearby agents influence one another, especially in crowded pedestrian areas or dense intersections [4]. Multimodal architectures that combine these ideas may use camera features for semantic cues, lidar for precise geometry, radar for velocity under adverse conditions, and maps for legal and topological constraints. The system-level challenge is to determine which modality should dominate in which context, how to detect when a modality is unreliable, and how to degrade gracefully when one input is missing.

Trade-offs emerge across the entire architecture. More modalities can improve accuracy but increase compute, power consumption, and sensor cost. Redundancy can improve safety but complicate calibration and maintenance. Centralized cloud inference can support large models and fleet-wide learning, but it depends on network availability and raises latency and privacy concerns. Edge inference reduces communication delays but constrains model size and may produce inconsistent behavior across vehicles with different hardware. Multi-task learning can share representations across perception, prediction, and planning, but task interference may harm rare-class performance [14]. A robust system must therefore be designed as an adaptive architecture, one that can shift between fusion modes, adjust computational effort, and communicate uncertainty to downstream planners. In this sense, architecture is not a static blueprint but a governance mechanism that encodes priorities about safety, efficiency, and accountability.

4. Robustness across Diverse Autonomous Driving Scenarios

Robustness in trajectory prediction refers to the ability to maintain acceptable performance when conditions differ from training data or when sensors, maps, and communication links are degraded. Diverse autonomous driving scenarios include dense urban intersections, high-speed highways, parking lots, rural roads, construction zones, adverse weather, night-time driving, and mixed traffic with pedestrians, cyclists, and micromobility devices. Datasets such as nuScenes [6], Argoverse [7], and the Waymo Open Motion Dataset [8] have broadened coverage, but no dataset can fully represent the long tail of real-world events. Robustness therefore depends on how models handle distribution shift, how they represent uncertainty, and how they fail. A model that produces a single trajectory may be dangerously overconfident, whereas a multimodal predictor can represent multiple plausible futures and allow the planner to reason about risk.

Uncertainty management is central to robustness. Aleatoric uncertainty arises from inherent randomness in human behavior and sensor noise, while epistemic uncertainty arises from limited knowledge or unfamiliar situations. Multi-task learning with uncertainty weighting has shown that models can learn to balance competing objectives and express confidence [14]. In trajectory prediction, uncertainty can be propagated to planning so that the vehicle slows down, increases following distance, or requests human intervention when predictions are unreliable. However, uncertainty estimates can themselves be miscalibrated, especially under distribution shift. A robust system should therefore combine probabilistic prediction with runtime monitoring, anomaly detection, and fallback behaviors. The safety literature has long emphasized that autonomous systems must handle rare but consequential failures [15], and trajectory prediction is a prime site where such failures can cascade into unsafe actions.

Scenario diversity also raises questions about adversarial and cooperative behavior. Some road users may behave unpredictably, aggressively, or illegally, while others may cooperate by yielding or signaling. Interactive datasets have attempted to capture such behaviors [9], and multimodal convolutional models have explored diverse future predictions for autonomous driving [10]. Yet robustness cannot be guaranteed by data alone. It requires

scenario-based testing, stress testing with synthetic perturbations, and validation in closed courses and shadow mode before deployment. It also requires cross-modal consistency checks. If camera-based prediction and lidar-based prediction disagree sharply, the system should treat the situation as ambiguous rather than average the outputs. Robustness is therefore not a single metric but a property of the entire prediction pipeline, including sensor health, calibration, map freshness, communication reliability, and the policy constraints imposed by the planner.

5. Fairness, Accountability, and Policy Implications

Fairness in trajectory prediction is often overlooked because the task appears purely technical. Yet prediction systems allocate risk unevenly. If a model is trained predominantly on data from car-centric suburban environments, it may underperform on pedestrians, cyclists, children, older adults, and people with mobility aids. If it learns cultural biases about who yields or who jaywalks, it may systematically mispredict certain groups. Bias and fairness research in machine learning has shown that disparities can arise from data, labels, features, and deployment contexts [16]. In autonomous driving, these disparities can translate into unequal safety outcomes. A vehicle that predicts pedestrians poorly in certain neighborhoods may be more likely to brake abruptly or fail to yield appropriately. A system that predicts cyclists as erratic by default may create hostile interactions even if it avoids collisions.

Accountability requires institutional mechanisms for auditing, explaining, and contesting prediction behavior. Ethical and societal implications of artificial intelligence demand that affected communities have a voice in how systems are deployed and governed [17]. This is particularly important in mobility because streets are public spaces. A private operator may own the vehicle and model, but the public bears the risk of failure. Regulators may require transparency about training data, performance across demographic and geographic groups, and the limits of prediction. However, full model transparency may be impractical for commercial systems. A more feasible approach is independent evaluation using standardized scenarios, disaggregated performance reporting, and third-party access to safety-critical test results. Accountability also involves liability. If a prediction error contributes to a crash, who is responsible: the sensor supplier, the data curator, the model developer, the fleet operator, or the city that installed infrastructure? Clear governance frameworks are needed to avoid a responsibility gap.

Policy implications extend to infrastructure investment and digital inclusion. If robust multimodal prediction depends on smart infrastructure, then regions with fewer resources may receive less safe autonomous mobility. This creates a two-tier geography of automation. Policy can mitigate this by supporting open datasets, shared map standards, and public-interest testing facilities. It can also require that autonomous systems perform safely without relying exclusively on private, proprietary infrastructure. Safety models for self-driving cars emphasize scalable rules and verifiable assumptions [18], but those assumptions must be socially negotiated. What counts as safe following distance? How should vehicles behave around informal crossings? How much risk is acceptable for pedestrians versus passengers? Multimodal prediction cannot answer these normative questions alone. It must be embedded in a policy process that treats fairness, accountability, and public trust as first-order design constraints.

6. Deployment, Infrastructure, and Sustainability

Deployment transforms a promising model into an operational system with maintenance, energy, and cost constraints. Multimodal trajectory prediction may run on vehicle-edge computers, roadside units, or cloud servers. Edge deployment reduces latency and keeps sensitive data local, but it limits model size and requires robust thermal and power management. Cloud deployment supports fleet learning and large-scale simulation, but it depends on network coverage and raises privacy concerns. Hybrid architectures can distribute computation, but they introduce synchronization, versioning, and cybersecurity challenges. Infrastructure is therefore not a neutral substrate. It shapes which multimodal features are available, how quickly they can be fused, and how gracefully the system degrades when connectivity is lost. End-to-end driving models trained from large-scale video datasets have shown the potential of data-driven approaches [13], but operational deployment requires more than end-to-end training; it requires observability, rollback, and maintenance.

Sustainability is both an environmental and an institutional concern. Multimodal models can be computationally expensive to train and serve, especially when they use high-resolution sensors, large map databases, and continuous fleet learning. Energy consumption affects operating costs and carbon emissions. Hardware lifecycle issues include sensor replacement, calibration drift, and electronic waste. A sustainable system should therefore consider model efficiency, data reuse, and the longevity of infrastructure. Redundant sensors may improve safety but increase material and energy footprints. Smaller models may be more sustainable but less robust. There is no universal optimum; the balance depends on context, regulation, and risk tolerance. A system designed for dense urban deployment may prioritize low latency and high redundancy, while a rural shuttle may prioritize energy efficiency and graceful degradation. Sustainability also includes institutional sustainability: data pipelines must be maintained, maps updated, and software patched over many years. A model that cannot be maintained is not robust, regardless of its benchmark performance.

7. Evaluation, Benchmarking, and Operational Assurance

Evaluation of multimodal trajectory prediction must move beyond average displacement error. Common metrics reward accurate predictions for frequent behaviors but may not capture safety-critical failures, multimodal uncertainty, or fairness across groups. A density-balanced evaluation framework highlights the need to assess performance across data distributions rather than aggregate averages [5]. Robustness benchmarks should include adverse weather, sensor dropout, map corruption, and adversarial agent behavior. They should also evaluate calibration, mode diversity, and the consequences of prediction errors for planning. A prediction that is slightly inaccurate but consistently cautious may be safer than a prediction that is accurate on average but occasionally catastrophic. Evaluation must therefore be tied to downstream planning and vehicle dynamics, not treated as an isolated perception task.

Operational assurance requires continuous monitoring after deployment. Fleet data can reveal distribution shift, sensor degradation, and new interaction patterns. However, monitoring must respect privacy and avoid creating surveillance infrastructures. Techniques such as federated learning, differential privacy, and on-device anomaly detection can help balance learning with civil liberties. Assurance also requires incident reporting, root-cause analysis, and corrective action. The AI safety literature has identified concrete problems such as reward hacking, negative side effects, and safe exploration [15], and autonomous driving adds physical risk and public accountability. A robust prediction system should be able to explain why it failed, not merely that it failed. This requires logging multimodal inputs, fusion decisions, uncertainty estimates, and planner responses in a way that supports investigation

without exposing personal data. Operational assurance is thus a socio-technical capability, not a software feature.

8. Cross-Domain Comparisons and Case Illustrations

Cross-domain comparisons can illuminate the challenges of multimodal trajectory prediction. In robotics, multimodal fusion is often used for manipulation and navigation, where sensors can be calibrated in controlled environments. In autonomous driving, the environment is open, adversarial, and socially governed. In aviation, safety assurance relies on redundancy, procedures, and certification, but traffic is highly structured. In healthcare, multimodal prediction combines images, text, and records, but decisions are mediated by clinicians and ethical review. Autonomous driving combines elements of all three: open-world uncertainty, safety-critical certification, and social value conflicts. Multimodal machine learning theory provides general principles for representation, alignment, and fusion [12], but the operational context determines which principles matter most. A driving system cannot simply adopt a healthcare governance model or an aviation certification model; it must develop its own institutions for data stewardship, testing, and accountability.

Case illustrations clarify these tensions. Consider an urban intersection with cameras, lidar, radar, and connected infrastructure. Multimodal prediction can fuse real-time sensor data with signal phase and timing, improving accuracy for vehicles and pedestrians. However, if the infrastructure is maintained by a city and the vehicle by a private company, data sharing agreements, cybersecurity, and liability must be negotiated. Now consider a highway with sparse infrastructure and high speeds. Prediction horizons are longer, and small errors can have severe consequences. Radar may be more reliable than cameras in rain, but maps may be outdated. A robust system might rely more on vehicle dynamics and less on social interaction. Finally, consider a rural road with poor connectivity and irregular traffic. The system may need to operate with minimal infrastructure and conservative fallback behaviors. These cases show that robustness is context-dependent and cannot be reduced to a single multimodal architecture.

9. Future Research Agenda and Governance Pathways

Future research should pursue several interconnected directions. First, multimodal architectures need better mechanisms for dynamic modality weighting, calibration, and graceful degradation. Attention-based models [1] and multimodal taxonomies [12] provide foundations, but operational systems require online assessment of sensor trust. Second, data curation must become a first-class research area, with methods for detecting imbalance, preserving provenance, and evaluating fairness [5]. Third, uncertainty quantification must be integrated with planning so that prediction confidence directly influences speed, distance, and fallback behavior [14]. Fourth, evaluation must include scenario-based, disaggregated, and longitudinal metrics that reflect real-world diversity [8]. Fifth, governance must support public participation, independent audit, and clear liability [17]. These directions are not separate; they form a system of mutual accountability.

Governance pathways should also address international coordination. Autonomous vehicles cross borders, but data protection, liability, and safety standards differ. A patchwork of regulations may fragment markets and create incentives for regulatory arbitrage. Shared technical standards for map formats, sensor calibration, and safety reporting could improve interoperability. Open benchmarks and public-interest datasets could reduce dependence on proprietary data. However, openness must be balanced with privacy and cybersecurity. A

governance pathway might include tiered data access, in which researchers and regulators can evaluate safety-critical scenarios without accessing personal identifiers. It might also include requirements for sustainability reporting, fairness audits, and incident disclosure. The goal is not to slow innovation but to ensure that innovation is robust, equitable, and trustworthy.

10. Conclusion

Multimodal learning offers a compelling foundation for robust trajectory prediction in autonomous driving, but its benefits are realized only when technical design is aligned with data governance, infrastructure, and policy. This paper has argued that robustness is an emergent socio-technical property. It depends on curatorial discipline that addresses imbalance and rare events [5], architectural choices that manage fusion and degradation [1], uncertainty methods that inform planning [14], and evaluation regimes that capture diverse scenarios [8]. It also depends on fairness, accountability, and public trust [16], as well as sustainable deployment and maintenance. The field should move beyond narrow benchmarking toward integrated systems that can explain their confidence, degrade safely, and operate fairly across different communities. Such systems will require collaboration among engineers, policymakers, infrastructure operators, and affected publics. The future of autonomous mobility will not be determined solely by the accuracy of prediction models, but by whether those models are embedded in institutions capable of sustaining safety, equity, and legitimacy over time.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
2. Liang, M., Yang, B., Hu, R., Chen, Y., Liao, R., Feng, S., & Urtasun, R. (2020). Learning lane graph representations for motion forecasting. In *Computer Vision – ECCV 2020* (pp. 541-556). Springer.
3. Salzmann, T., Ivanovic, B., Chakravarty, P., & Pavone, M. (2020). Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision – ECCV 2020* (pp. 683-700). Springer.
4. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., & Savarese, S. (2016). Social LSTM: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 961-971). IEEE.
5. Yang, R., Xu, Y., Fu, Y., & Su, L. (2026). Den-tp: A density-balanced data curation and evaluation framework for trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 10632-10641).
6. 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). IEEE.
7. Chang, M. F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., & Hays, J. (2019). Argoverse: 3D tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8748-8757). IEEE.

8. Ettinger, S., Cheng, S., Caine, B., Liu, C., Zhao, H., Pradhan, S., Chai, Y., Sapp, B., Qi, C. R., Zhou, Y., Yang, Z., Chouard, A., Sun, P., Ngiam, J., Vasudevan, V., McCauley, A., Shlens, J., & Anguelov, D. (2021). Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 9710-9719). IEEE.
9. Zhan, W., Sun, L., Wang, D., Shi, H., Clausse, A., Naumann, M., Kummerle, J., Konigshof, H., Stiller, C., de La Fortelle, A., & Tomizuka, M. (2019). INTERACTION dataset: An INTERNATIONAL, Adversarial and Cooperative moTION dataset in interactive driving scenarios with semantic maps. *arXiv preprint arXiv:1910.03088*.
10. Cui, H., Radosavljevic, V., Chou, F. C., Lin, T. H., Nguyen, T., Huang, T. K., Schneider, J., & Djuric, N. (2019). Multimodal trajectory predictions for autonomous driving using deep convolutional networks. In *2019 International Conference on Robotics and Automation* (pp. 2090-2096). IEEE.
11. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning* (pp. 689-696). Omnipress.
12. Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423-443.
13. Xu, H., Gao, Y., Yu, F., & Darrell, T. (2017). End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 2174-2182). IEEE.
14. Kendall, A., Gal, Y., & Cipolla, R. (2018). Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7482-7491). IEEE.
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. 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.
17. Whittlestone, J., Nyrupe, R., Alexandrova, A., & Cave, S. (2019). Ethical and societal implications of algorithms, data, and artificial intelligence: A roadmap for research. Nuffield Foundation.
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*.