Safety-Constrained Policy Optimization for End-to-End Autonomous Driving Using Offline Reinforcement Learning

Julita Wyrostek
Department of Computer Science, George Mason University, Fairfax, VA, USA.
wyrostek1995@gmu.edu

Mattias Kuhl
Department of Computer Science, University of Houston, Houston, TX, USA.
kuhl177@uh.edu

Quinten Schiffer
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
schifferquinten@oregonstate.edu

Rachita J. Chadha
Department of Computer Science, University of North Texas, Denton, TX, USA.
rjchadha@unt.edu

Abstract

End-to-end autonomous driving systems increasingly combine perception, prediction, planning, and control within learned computational pipelines, raising the possibility of more adaptive behavior but also intensifying problems of safety assurance, interpretability, and regulatory oversight. Offline reinforcement learning offers a compelling route for such systems because it allows policies to be trained from previously collected driving data without requiring unsafe online exploration. Yet offline learning alone does not resolve the deeper socio-technical challenge of constraining policy behavior under rare, ambiguous, and ethically charged conditions. This paper examines safety-constrained policy optimization as an architectural and governance strategy for end-to-end autonomous driving using offline reinforcement learning. It argues that safety constraints should not be treated as post hoc filters or narrow reward penalties but as structural commitments that shape data curation, model architecture, training objectives, evaluation protocols, deployment boundaries, and accountability mechanisms. The analysis connects technical advances in offline reinforcement learning, constrained policy optimization, end-to-end driving, simulation infrastructure, and formal safety models with concerns about robustness, fairness, sustainability, and public legitimacy. It also considers how fleet-level learning, over-the-air updates, edge computation, and regulatory auditing create new dependencies between engineering choices and institutional capacity. The paper concludes that safe end-to-end autonomy requires a layered system in which constrained offline policy optimization is embedded within transparent operational design domains, continuous validation, human oversight, and enforceable governance. Without such integration, offline reinforcement learning may improve sample efficiency while leaving unresolved the structural risks that determine whether autonomous driving can be trusted at scale.

Keywords

safety-constrained policy optimization; offline reinforcement learning; end-to-end autonomous driving; socio-technical systems; governance; robustness; fairness; deployment.

1. Introduction

Autonomous driving has become a central test case for the integration of artificial intelligence into high-stakes physical infrastructure. The promise of end-to-end learning is that perception, prediction, planning, and control can be jointly optimized rather than partitioned into manually engineered modules with brittle interfaces. Deep learning has demonstrated that hierarchical representation learning can extract useful features from complex sensory data [1], while reinforcement learning has shown that agents can acquire sophisticated control and decision policies through interaction and reward signals [2], [3]. These developments suggest that driving policies could be learned from data and experience rather than specified through exhaustive hand-coded rules. However, the same flexibility that makes learned policies attractive also makes them difficult to certify, audit, and constrain, particularly when deployed in open-world environments where human safety is at stake.

Offline reinforcement learning has emerged as a practical response to the sample inefficiency and safety risks of online trial-and-error learning. Instead of allowing an agent to explore directly in the physical world, offline methods learn from fixed datasets of previously collected trajectories [4]. This is especially relevant for autonomous driving because fleets, simulators, and human-driven vehicles continuously generate large amounts of observational and behavioral data. Yet offline learning introduces its own pathologies, including distributional shift, extrapolation error, and overestimation of actions that are rare or absent in the dataset [5], [6]. A policy trained offline may appear strong on historical data while behaving unpredictably when deployed in conditions that differ from the training distribution. Consequently, offline reinforcement learning for driving cannot be reduced to a data-fitting problem. It must be framed as a safety-constrained policy optimization problem embedded in a larger system of data infrastructure, validation, and governance.

Safety-constrained policy optimization provides a vocabulary for this challenge. Rather than maximizing expected reward alone, constrained approaches seek to optimize performance subject to explicit safety requirements or cost limits [7]. The broader literature on AI safety has long emphasized that specification, robustness, oversight, and distributional shift are not peripheral issues but core design problems [8]. Safe reinforcement learning surveys similarly show that safety can be approached through risk-sensitive objectives, shielding, control-theoretic constraints, or constrained optimization, each with different assumptions and failure modes [9]. This paper argues that for end-to-end autonomous driving, safety constraints must become architectural commitments. They should determine what data is collected, how policies are represented, how uncertainty is estimated, how updates are gated, how human oversight is organized, and how responsibilities are allocated across manufacturers, regulators, cities, and users.

2. Offline Reinforcement Learning and the End-to-End Driving Problem

End-to-end driving has intellectual roots in early neural network approaches to autonomous navigation, which demonstrated that a learned mapping from sensors to control could operate a vehicle in constrained settings [11]. More recent work has revived this idea with deep networks trained on large-scale driving data, showing that a single model can learn lane following and basic steering behavior from camera inputs [12]. Conditional imitation learning

extended this paradigm by incorporating navigational commands and higher-level intentions, allowing policies to respond to route instructions rather than merely imitate low-level steering [13]. Simulation environments such as CARLA have further enabled reproducible experimentation with urban driving scenarios, sensor configurations, and traffic interactions [14]. These developments collectively support the vision of end-to-end autonomy as a learnable system rather than a modular pipeline.

Offline reinforcement learning adds a distinct dimension to this vision. In conventional imitation learning, the policy is trained to mimic expert actions, which can limit performance when the expert data does not cover rare or safety-critical situations. Reinforcement learning, by contrast, can optimize long-horizon objectives such as comfort, progress, energy efficiency, or rule compliance. Offline reinforcement learning seeks to combine the data efficiency of imitation with the optimization capacity of reinforcement learning by learning from logged transitions without online exploration [4]. Conservative methods penalize uncertain actions or lower their estimated value, reducing the risk that the policy exploits regions of the action space unsupported by data [5]. Batch-constrained approaches similarly restrict the policy to actions observed in the dataset, mitigating extrapolation error [6]. These techniques are promising for autonomous driving because they align with the operational reality that physical exploration by immature policies is unacceptable.

However, the end-to-end setting complicates the application of offline reinforcement learning. A driving policy does not merely select discrete actions in a stationary environment; it operates within a dynamic multi-agent system where other drivers, pedestrians, cyclists, and infrastructure signals co-produce outcomes. The observation space is high-dimensional, partially observable, and affected by weather, lighting, sensor degradation, and occlusion. The reward is difficult to specify because safety, legality, comfort, progress, and social acceptability are not easily collapsed into a single scalar. Moreover, the dataset itself is never neutral. It reflects the routes, environments, and behaviors of the vehicles and drivers that generated it, which may underrepresent nighttime conditions, vulnerable road users, rare intersections, or particular neighborhoods. A recent line of work on post-training in end-to-end autonomous driving highlights that adapting pretrained driving models after initial training is itself a complex safety and performance problem, not merely a final fine-tuning step [10]. This observation reinforces the need to treat offline learning as one phase within a broader lifecycle of constrained policy improvement.

3. Safety-Constrained Policy Optimization as an Architectural Commitment

Constrained policy optimization is often introduced as a method for balancing reward maximization with cost minimization, but its significance for autonomous driving is architectural rather than merely algorithmic. Safety constraints define what the system is not allowed to do, what risks are tolerable, and what evidence is required before an update is accepted. In this sense, constraints are not external limits imposed on an otherwise complete policy; they are constitutive of the policy's operational identity. A vehicle that is optimized only for progress may learn to edge aggressively into intersections, whereas a vehicle optimized under explicit safety costs may adopt more conservative gap acceptance, larger following distances, and clearer yielding behavior. The difference is not just performance tuning but a different relationship between the vehicle and its environment.

Formal safety models provide conceptual resources for translating abstract constraints into operational requirements. Responsibility-sensitive safety frameworks, for example, attempt to define safe following, merging, and intersection behavior through rules that can be checked

and enforced [16]. Such models do not eliminate the need for learning, but they can structure the constraint space within which learned policies operate. In an end-to-end system, these constraints might be represented as cost signals, shielding mechanisms, runtime monitors, or fallback controllers. Yet the integration of constraints with learned perception and planning is difficult because the system may not have reliable access to the semantic variables that formal models require. If the perception module misclassifies a pedestrian as a shadow or a stopped vehicle as a plastic bag, a formally safe planner can still produce unsafe behavior. Therefore, safety-constrained policy optimization must account for uncertainty in perception, prediction, and action, not merely uncertainty in the policy's value estimates.

The constrained optimization literature offers important lessons about trade-offs. Constrained policy optimization methods can improve safety during learning, but they often assume access to reliable cost signals and may struggle when constraints are violated rarely or in complex temporal patterns [7]. Safe reinforcement learning surveys show that different safety definitions lead to different technical commitments: a method may guarantee probabilistic constraint satisfaction, enforce hard constraints under bounded uncertainty, or rely on human oversight to intervene [9]. For autonomous driving, no single approach is likely to suffice. A layered architecture may combine conservative offline policy learning, runtime verification of safety envelopes, uncertainty-aware perception, and human or remote supervision for edge cases. The key point is that safety constraints should be visible across layers rather than confined to a final reward term.

4. Data, Simulation, and Infrastructure for Offline Policy Improvement

Offline reinforcement learning depends on data infrastructure as much as on algorithm design. For autonomous driving, the relevant data includes sensor streams, vehicle trajectories, map information, environmental conditions, human interventions, near misses, and rare-event annotations. The quality, coverage, and provenance of this data determine what policies can learn and what risks remain invisible. A dataset dominated by daytime suburban driving cannot support strong claims about nighttime urban performance, just as a dataset collected in one regulatory jurisdiction cannot automatically transfer to another with different traffic norms. Data curation therefore becomes a safety activity. It requires deliberate sampling of difficult scenarios, audit trails for data sources, privacy-preserving retention policies, and mechanisms for identifying gaps in representation.

Simulation plays a central role in this infrastructure because it allows offline policies to encounter scenarios that are dangerous or rare in real logs. CARLA and similar platforms provide controllable environments for testing perception, planning, and control under varied conditions [14]. Simulation can generate synthetic interventions, adversarial pedestrians, occluded crossings, and sensor failures that would be unethical to stage on public roads. However, simulation also introduces reality gaps. A policy that succeeds in a simulator may exploit rendering artifacts, simplified physics, or unrealistic agent behavior. For offline reinforcement learning, the danger is compounded because the policy may be trained on simulated transitions whose dynamics differ subtly from the real world. Simulation is therefore most valuable when used as part of a mixed-reality validation regime that combines synthetic scenario generation, real-world replay, closed-course testing, and controlled road trials.

Fleet learning creates additional infrastructural tensions. If vehicles upload data and receive updated policies, the system becomes a distributed learning infrastructure with privacy, cybersecurity, and update-management implications. Over-the-air updates can improve safety

by correcting discovered failures, but they can also propagate new errors across an entire fleet. Edge computation can reduce latency and preserve privacy by processing data locally, but it may limit the sophistication of models that can be deployed. Cloud-based training can aggregate diverse experiences, but it raises questions about data ownership, consent, and cross-border transfer. These are not secondary concerns. They determine whether offline policy optimization can be sustained as a trustworthy practice rather than a one-time engineering achievement.

5. Post-Training, Deployment, and Operational Governance

The deployment of an end-to-end autonomous driving policy is not a single event but a sequence of bounded releases, operational design domains, and monitored expansions. Even after offline training and simulation validation, a policy must be post-trained or adapted to specific vehicle platforms, sensor suites, cities, and traffic cultures. The recent emphasis on post-training in end-to-end autonomous driving [10] suggests that the gap between generic pretraining and safe deployment is substantial. Post-training may involve fine-tuning on target-domain data, calibrating uncertainty estimates, adjusting conservative penalties, or integrating rule-based safeguards. Each of these steps can improve performance, but each also introduces new opportunities for overfitting, distributional narrowing, or hidden regressions.

Operational governance is necessary because technical validation cannot anticipate every emergent condition. A vehicle fleet operates within a socio-technical system that includes road authorities, emergency services, insurers, legislators, passengers, pedestrians, and the public. Governance mechanisms might include geofencing, speed and route restrictions, weather-based operational limits, remote assistance, incident reporting, and independent auditing. These mechanisms are often treated as external constraints on a technical system, but they should be co-designed with the policy optimization process. If a policy is trained without knowledge of geofencing or remote assistance, it may behave poorly when those interventions occur. If governance is designed without understanding the policy's uncertainty profile, it may impose rules that are either unnecessarily restrictive or dangerously permissive.

Human oversight remains essential, but it must be realistic. Remote operators cannot continuously monitor every vehicle, and passengers may not have the expertise or authority to intervene effectively. In-vehicle drivers, where present, may suffer from automation complacency or mode confusion. The safety case for end-to-end autonomy therefore cannot rely on an assumption of perfect human backup. Instead, oversight should be designed as a layered system with clear escalation paths, measurable response times, and fallback behaviors that are themselves validated. Constrained policy optimization can support this by producing policies that know when to defer, when to slow down, and when to request assistance. Yet such metacognitive behavior must be trained and tested explicitly rather than expected to emerge from reward maximization alone.

6. Robustness, Fairness, and Sustainability in Socio-Technical Terms

Robustness in autonomous driving is not only a matter of handling sensor noise or adversarial attacks. It also involves maintaining safe behavior across social and infrastructural variation. A policy trained in one city may encounter different lane markings, informal crossing practices, police directions, construction detours, or weather patterns in another. Robustness therefore depends on diversity in training data, conservative behavior under uncertainty, and the ability to detect when the operational design domain has been exceeded. Offline reinforcement learning can contribute by learning from broad historical data, but it can also

amplify biases if the data is skewed. If vulnerable road users are underrepresented, the policy may fail to predict their behavior accurately. If certain neighborhoods are over-policed or under-mapped, the system may perform unevenly across communities.

Fairness concerns are thus intrinsic to the data and deployment choices of end-to-end autonomy. A vehicle that is safer in wealthy, well-mapped areas than in informal settlements is not merely technically incomplete; it distributes risk inequitably. Safety-constrained policy optimization should therefore include distributional considerations in constraint specification and evaluation. This means asking not only whether the average collision rate is low, but whether safety margins are consistent across pedestrian types, age groups, vehicle classes, and geographic areas. It also means involving affected communities in decisions about where autonomous vehicles operate, how data is collected, and how incidents are investigated. Such participation is not a substitute for technical rigor, but it can reveal blind spots that purely quantitative evaluation misses.

Sustainability adds another layer of trade-offs. End-to-end models can be computationally expensive to train and deploy, and fleet-scale learning may require substantial energy and network infrastructure. Edge inference may reduce communication but increase vehicle cost and power consumption. Cloud training may be more efficient in aggregate but depends on data centers with their own environmental footprint. Safety constraints can also affect sustainability because conservative driving may reduce collisions but increase travel time or energy use, while aggressive optimization may improve traffic flow but raise risk. These trade-offs should be made explicit rather than hidden in a single reward function. A socio-technical approach treats safety, fairness, and sustainability as interdependent objectives that require transparent deliberation and adaptive governance.

7. Evaluation, Validation, and Policy Implications

Evaluation is the bridge between offline policy optimization and legitimate deployment. Standard machine learning metrics such as loss, accuracy, or average reward are insufficient for autonomous driving because they do not capture tail risk, causal responsibility, or social acceptability. Testing and validation must include rare-event simulation, closed-course scenarios, shadow mode deployment, and gradually expanding road trials [18]. Real-world reinforcement learning research has identified persistent challenges such as partial observability, delayed rewards, safety constraints, and non-stationarity, all of which apply forcefully to driving [19]. The evaluation regime should therefore combine statistical measures, formal analysis where possible, and qualitative incident review. It should also be continuous, because a policy that was safe at deployment may become unsafe after a software update, a map change, or a shift in traffic patterns.

Regulatory policy must contend with the fact that end-to-end learned systems are difficult to inspect in traditional ways. A rule-based module can be traced through code, but a neural policy's decision may emerge from distributed representations that resist simple explanation. This does not make regulation impossible, but it shifts the emphasis from source-code inspection to performance-based standards, process audits, data governance, and runtime monitoring. Regulators may require manufacturers to define operational design domains, maintain incident logs, demonstrate cybersecurity protections, and provide mechanisms for recall or update. They may also need authority to suspend deployment when systematic failures are detected. Such powers are only legitimate if they are paired with technical capacity and transparent procedures.

The policy implications extend beyond national vehicle safety regulation. Cities control road design, curb management, transit integration, and data-sharing agreements. Insurers shape liability and risk pooling. Labor organizations and disability advocates have stakes in accessibility and employment. International bodies may need to harmonize standards so that vehicles do not face contradictory requirements across borders. In this environment, safety-constrained offline reinforcement learning is not merely a method for producing better driving policies. It is a site where technical choices become public policy. The decision to prioritize certain constraints, to collect certain data, or to permit certain updates is a decision about who bears risk and who benefits from automation.

8. Conclusion

This paper has argued that safety-constrained policy optimization for end-to-end autonomous driving using offline reinforcement learning must be understood as a system-level and socio-technical undertaking. Offline reinforcement learning offers a credible path to learn driving policies without unsafe online exploration, and constrained optimization provides a principled way to encode safety requirements. Yet neither technique alone can guarantee trustworthy autonomy. Safety constraints must shape data collection, model architecture, post-training, runtime monitoring, deployment boundaries, and governance. They must be evaluated not only through average performance but through tail risk, distributional fairness, robustness to distribution shift, and responsiveness to public accountability.

The integration challenge is substantial. End-to-end models blur the boundaries between perception, prediction, planning, and control, making it harder to localize failures and assign responsibility. Offline datasets are never complete, simulators are never perfect, and real-world operations are always partially unpredictable. In response, the field should move toward layered architectures in which conservative offline learning is combined with uncertainty-aware perception, formal safety envelopes, human oversight, and adaptive regulation. Such architectures will not eliminate risk, but they can make risk more visible, contestable, and manageable.

Ultimately, the legitimacy of autonomous driving will depend less on any single algorithm than on the institutional capacity to constrain, monitor, and revise learned behavior over time. Safety-constrained policy optimization is valuable because it forces explicit trade-offs among mobility, efficiency, comfort, equity, and risk. When embedded in robust infrastructure and accountable governance, it can help end-to-end autonomous driving move from demonstration to durable public trust. Without that embedding, offline reinforcement learning may optimize narrowly while leaving the most important safety questions unanswered.

References

1. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

2. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533. https://doi.org/10.1038/nature14236

3. Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe,

D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T., & Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. Nature, 529(7587), 484–489. https://doi.org/10.1038/nature16961

4. Levine, S., Kumar, A., Tucker, G., & Fu, J. (2020). Offline reinforcement learning: Tutorial, review, and perspectives on open problems. arXiv preprint arXiv:2005.01643. https://doi.org/10.48550/arXiv.2005.01643

5. Kumar, A., Zhou, A., Tucker, G., & Levine, S. (2020). Conservative Q-learning for offline reinforcement learning. Advances in Neural Information Processing Systems, 33, 1179–1191.

6. Fujimoto, S., Meger, D., & Precup, D. (2019). Off-policy deep reinforcement learning without exploration. Proceedings of the 36th International Conference on Machine Learning, 97, 2052–2062.

7. Achiam, J., Held, D., Tamar, A., & Abbeel, P. (2017). Constrained policy optimization. Proceedings of the 34th International Conference on Machine Learning, 70, 22–31.

8. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565. https://doi.org/10.48550/arXiv.1606.06565

9. García, J., & Fernández, F. (2015). A comprehensive survey on safe reinforcement learning. Journal of Machine Learning Research, 16(1), 1437–1480.

10. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

11. Pomerleau, D. A. (1991). Efficient training of artificial neural networks for autonomous navigation. Neural Computation, 3(1), 88–97. https://doi.org/10.1162/neco.1991.3.1.88

12. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., & Zieba, K. (2016). End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316. https://doi.org/10.48550/arXiv.1604.07316

13. Codevilla, F., Müller, M., López, A., Koltun, V., & Dosovitskiy, A. (2018). End-to-end driving via conditional imitation learning. 2018 IEEE International Conference on Robotics and Automation (ICRA), 4693–4700. https://doi.org/10.1109/ICRA.2018.8460487

14. Dosovitskiy, A., Ros, G., Codevilla, F., López, A., & Koltun, V. (2017). CARLA: An open urban driving simulator. Proceedings of the 1st Annual Conference on Robot Learning, 78, 1–16.

15. Kendall, A., Hawke, J., Janz, D., Mazur, P., Reda, M., Allen, J.-M., Lam, V.-D., Qiao, K., & Shah, A. (2019). Learning to drive in a day. 2019 International Conference on Robotics and Automation (ICRA), 8248–8254. https://doi.org/10.1109/ICRA.2019.8793742

16. Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2017). On a formal model of safe and scalable self-driving cars. arXiv preprint arXiv:1708.06374. https://doi.org/10.48550/arXiv.1708.06374

17. Yurtsever, E., Lambert, J., Carballo, A., & Takeda, K. (2020). A survey of autonomous driving: Common practices and emerging technologies. IEEE Access, 8, 58443–58469. https://doi.org/10.1109/ACCESS.2020.2983149

18. Koopman, P., & Wagner, M. (2016). Challenges in autonomous vehicle testing and validation. SAE International Journal of Transportation Safety, 4(1), 15–24. https://doi.org/10.4271/2016-01-0128

19. Dulac-Arnold, G., Levine, N., Mankowitz, D. J., Li, J., Paduraru, C., Gowal, S., & Hester, T. (2021). Challenges of real-world reinforcement learning: Definitions, benchmarks and analysis. Machine Learning, 110(9), 2419–2468. https://doi.org/10.1007/s10994-021-05961-4