

Reinforcement Learning-Based Post-Training Optimization for End-to-End Autonomous Driving Models

Pacifica Porras

Department of Computer Science, University of Alabama at Birmingham, Birmingham, AL, USA.

pporras@uab.edu

Heinz-Werner Nohlmans

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

heinzwerner1998@ku.edu

Micheala Kilgallon

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.

michealakilgallon14@buffalo.edu

Hendrik Bolnbach

Department of Computer Science, University of Central Florida, Orlando, FL, USA.

hbolnbach@ucf.edu

Abstract

End-to-end autonomous driving models increasingly map raw sensor observations directly to control actions, promising reduced engineering complexity and improved adaptability relative to modular pipelines. However, these models remain vulnerable to distribution shift, rare safety-critical events, and reward misalignment when deployed in open-world traffic. This paper examines reinforcement learning-based post-training as a system-level optimization layer for end-to-end driving models. It argues that post-training is not merely a model refinement technique but a socio-technical intervention that reshapes architecture, data infrastructure, validation regimes, governance, and public accountability. The discussion situates imitation-based pre-training alongside reinforcement learning fine-tuning, preference modeling, and constrained policy optimization. It analyzes structural trade-offs involving simulator fidelity, compute intensity, latency constraints, safety envelopes, and regulatory oversight. The paper further considers fairness across geographic contexts, sustainability implications of large-scale training, and policy mechanisms for staged deployment and incident reporting. A central claim is that effective post-training requires an integrated stack in which reward design, human feedback, formal safety constraints, and deployment monitoring are co-engineered. Without such integration, reinforcement learning post-training may improve aggregate performance while amplifying opaque failure modes. The paper concludes with a research agenda for auditable, resource-aware, and socially legitimate post-training pipelines for autonomous driving.

Keywords

end-to-end autonomous driving, reinforcement learning, post-training optimization, safety governance, socio-technical systems, deployment infrastructure.

1. Introduction

End-to-end autonomous driving has emerged as a major alternative to modular perception-planning-control architectures. Early work demonstrated that convolutional networks could learn steering behavior from camera images and human driving data, establishing the feasibility of direct sensor-to-action learning [1]. Subsequent conditional imitation learning frameworks introduced navigational commands and route intent into the policy, allowing models to produce more controllable behaviors in urban settings [2]. Simulators such as CARLA provided reproducible environments for training and benchmarking, while real-world experiments showed that end-to-end policies could acquire basic driving competence within limited data-collection periods [3,4]. These developments shifted the research frontier from hand-engineered pipelines toward learned representations and integrated policy optimization.

Despite these advances, end-to-end models face persistent limitations. Imitation learning tends to inherit the biases of demonstration data and struggles when the vehicle encounters states outside the training distribution. Safety-critical long-tail events are underrepresented, and small prediction errors can compound into hazardous maneuvers. Formal safety models for autonomous vehicles emphasize that scalable deployment requires verifiable assumptions about perception, planning, and fallback behavior [5]. Reinforcement learning has therefore attracted attention as a post-training mechanism that can optimize policies beyond imitation, using reward signals, simulations, and interactive feedback to improve decision quality in rare or complex scenarios [6,7].

This paper focuses on reinforcement learning-based post-training optimization for end-to-end autonomous driving models. The term post-training refers to the phase after initial pre-training, typically imitation-based, in which a policy is refined through reinforcement learning, preference optimization, constraint enforcement, or related adaptation methods. The analysis is deliberately system-oriented. It examines how post-training affects model architecture, simulation infrastructure, compute provisioning, validation pipelines, fairness, sustainability, and policy governance. The paper also considers how post-training interacts with broader regulatory frameworks for automated vehicles and safety assurance [8,9]. A recent survey of post-training in end-to-end autonomous driving provides a useful reference point for organizing these dimensions [11], but the present discussion extends beyond method taxonomy to institutional and infrastructural consequences.

2. Foundations of End-to-End Autonomous Driving and Post-Training

End-to-end driving models generally combine perception, prediction, and planning into a single differentiable or jointly optimized system. Sensor inputs such as cameras, lidar, radar, and vehicle state are encoded into latent representations that inform trajectory generation or direct control outputs. Transformer-based sensor fusion and multi-modal architectures have improved the capacity of such systems to integrate heterogeneous inputs across time [8]. Yet the end-to-end paradigm also concentrates failure risk: an error in representation learning can propagate directly into control without the interpretable interfaces that modular systems sometimes provide. Consequently, post-training must address not only performance improvement but also diagnosability, robustness, and safe fallback.

Post-training can be understood as a second-order optimization phase in which the pre-trained policy is adapted to objectives that are difficult to specify through demonstration alone.

Imitation learning answers the question of how a human might drive in recorded situations, whereas reinforcement learning asks how the policy should act to maximize long-term returns under a reward model. In autonomous driving, return maximization may include collision avoidance, ride comfort, traffic-rule compliance, progress toward destination, and social compatibility with other road users. The challenge is that these objectives are multi-dimensional, context-dependent, and sometimes in tension. A policy that maximizes progress may take unacceptable risks, while an excessively conservative policy may obstruct traffic or create new hazards.

The post-training layer therefore becomes a site of normative design. Reward functions are not neutral descriptions of safety; they encode priorities among road users, operational design domains, and acceptable risk thresholds. Preference-based reinforcement learning and human feedback methods offer one way to learn reward models from comparative judgments rather than hand-crafted scores [14,15,16]. In driving, human preferences might capture subtle notions of comfort, courtesy, and predictability that are difficult to formalize. However, preferences are also socially situated and may vary across cultures, regions, and regulatory environments. Post-training thus raises questions about whose preferences are represented, how disagreement is managed, and whether minority or vulnerable road users are adequately considered.

3. Reinforcement Learning as a Post-Training Layer

Reinforcement learning provides a family of methods for optimizing sequential decision policies through interaction with an environment. Deep reinforcement learning demonstrated that neural networks could learn control policies from high-dimensional observations in domains such as video games and board games [12,13]. Proximal policy optimization and related algorithms have become common because they balance sample efficiency, stability, and implementation complexity [10]. In end-to-end driving, these methods can be applied in simulation, on closed test tracks, or through limited real-world interaction. The appeal is that the policy can explore recovery behaviors, hard braking, lane changes, and negotiation with other agents beyond the demonstrated dataset.

However, applying reinforcement learning to autonomous driving raises structural difficulties. The cost of exploration in the real world is prohibitive because unsafe actions can cause injury or property damage. Simulation is therefore essential, but simulators inevitably simplify sensor physics, agent behavior, and social norms [3,7]. Policies trained in simulation may overfit to simulator artifacts and fail in the real world. Domain randomization, realistic rendering, and data-driven simulation can mitigate this gap, yet no simulator can fully represent the open-ended complexity of public roads. Post-training must therefore be coupled with robust validation, staged deployment, and fallback mechanisms that limit the consequences of simulation-to-reality mismatch.

A further issue is reward hacking, in which a policy discovers unintended ways to maximize reward without achieving the designer's intended outcome. In driving, a policy might learn to exploit simulator traffic rules, drive aggressively in ways that are rewarded by progress metrics, or suppress cautious behavior to reduce perceived hesitation. Safety research has long warned that poorly specified objectives can produce harmful behavior in complex systems [17]. Post-training pipelines must therefore include adversarial reward testing, red-teaming, and constraint-based optimization that treats safety as a boundary condition rather than merely another reward term.

Governance enters at this point because post-training decisions are not purely technical. Regulators, insurers, and municipal authorities may require evidence that policies satisfy minimum safety standards before deployment [19]. Standards bodies have developed taxonomies for driving automation that clarify roles and responsibilities across levels of automation [20]. Reinforcement learning post-training must be documented in ways that support audit, liability assignment, and public accountability. This means recording reward specifications, simulation scenarios, human feedback sources, model versions, and deployment constraints. Without such records, post-training becomes an opaque optimization step that complicates incident investigation and public trust.

4. Architecture and System-Level Trade-Offs

Integrating reinforcement learning post-training into an end-to-end driving architecture involves trade-offs among modularity, latency, interpretability, and update frequency. One design option is to post-train the entire end-to-end network as a single policy. This preserves architectural simplicity and allows gradients or policy updates to influence perception and control jointly. Yet it also makes it difficult to isolate which subsystem caused a failure and can destabilize previously learned representations. An alternative is to freeze the perception backbone and post-train only a planning or control head. This improves stability and interpretability but may limit the policy's ability to adapt representations to new objectives.

Simulation infrastructure is another system-level concern. High-fidelity simulators require substantial computational resources, expert scenario design, and maintenance of asset libraries. Data-driven simulation can generate realistic agent behavior but may inherit biases from the real-world data used to train it [7]. The choice of simulator affects the kinds of policies that emerge. A simulator with overly compliant pedestrians will not prepare a policy for jaywalking, while one with aggressive traffic may encourage excessively defensive driving. Post-training thus depends on a scenario ecosystem that must be curated, versioned, and governed as part of the deployment pipeline.

Compute and latency constraints also shape post-training design. Large-scale reinforcement learning requires parallel environments, distributed rollouts, and frequent policy evaluation [10,12]. This is feasible in cloud data centers but difficult to replicate on vehicles with limited power and thermal budgets. The trained policy must ultimately run under strict latency requirements, often within milliseconds, while maintaining redundancy and fail-safe behavior. Post-training may improve policy quality, but it can also increase model size or inference complexity if it incorporates additional reward models, ensembles, or safety critics. System architects must therefore balance accuracy gains against deployment feasibility.

Data infrastructure is equally important. Post-training benefits from a feedback loop in which fleet data, simulation failures, and human interventions are curated into new training scenarios. This creates a flywheel effect but also raises privacy, ownership, and security concerns. Vehicle data may include location traces, camera imagery, and driver behavior. Aggregating such data across jurisdictions requires compliance with privacy laws and cybersecurity standards. The post-training pipeline must therefore be designed as a governed data system, not merely a machine learning workflow.

5. Safety, Robustness, and Governance

Safety is the central constraint for autonomous driving post-training. Formal models of safe driving often rely on assumptions about perception accuracy, rule compliance, and responsibility allocation [5]. Reinforcement learning complicates these assumptions because

policies are learned rather than explicitly programmed. A post-trained policy may satisfy safety metrics in simulation yet behave unpredictably in rare real-world conditions. Robustness testing must therefore cover sensor degradation, adverse weather, occluded pedestrians, construction zones, emergency vehicles, and adversarial interactions. The policy should degrade gracefully, either by transferring control to a human driver or by entering a minimal-risk condition.

Governance frameworks must address the fact that post-training can change policy behavior without changing the vehicle's physical hardware. Over-the-air updates could introduce new driving behaviors across a fleet, raising questions about re-certification and regulatory approval. If a post-trained model improves efficiency but alters braking distance or lane-change aggressiveness, regulators may need to review the update. Standards for driving automation provide a vocabulary for levels and roles, but they do not by themselves specify how reinforcement learning updates should be validated [20]. Policy mechanisms such as pre-deployment notification, staged rollouts, and mandatory incident reporting can help bridge this gap.

Liability is another governance challenge. When an autonomous vehicle fails, responsibility may be distributed among the manufacturer, software supplier, data provider, simulation vendor, and fleet operator. Reinforcement learning post-training adds further complexity because the policy's behavior emerges from reward design and environmental interaction. Auditable documentation of reward functions, training scenarios, and evaluation results is essential for assigning responsibility and improving safety culture [19]. Without transparency, post-training may weaken accountability by making it difficult to reconstruct why a policy chose a particular action.

Public legitimacy also depends on how safety trade-offs are communicated. Communities may accept autonomous vehicles if they reduce traffic fatalities and improve mobility, but they may resist if deployment disproportionately affects vulnerable populations or if incidents are concealed. Post-training optimization should therefore be accompanied by participatory governance mechanisms, including public reporting, independent audits, and channels for affected communities to raise concerns. Safety is not only a technical property but also a social relationship between institutions and the public they serve [18].

6. Fairness, Sustainability, and Infrastructure

Fairness in autonomous driving post-training concerns both distributive and procedural dimensions. Distributive fairness asks whether the benefits and risks of automated mobility are shared equitably across regions, income levels, age groups, and abilities. A policy optimized primarily on data from dense, well-mapped urban areas may perform poorly in rural roads, informal settlements, or regions with irregular traffic patterns. Procedural fairness asks who participates in defining reward functions and safety thresholds. If post-training is shaped only by engineers and corporate stakeholders, it may neglect the perspectives of pedestrians, cyclists, and marginalized road users. Human feedback methods can incorporate diverse preferences, but they must be designed to avoid amplifying the voices of the most vocal or advantaged groups [15,16].

Sustainability is closely tied to the computational intensity of reinforcement learning post-training. Large-scale simulation and policy optimization require significant energy and water for data centers, as well as hardware resources that have environmental and social costs [12,13]. While autonomous vehicles may improve fuel efficiency through optimized driving,

the training footprint can be substantial. Sustainable post-training should therefore consider model efficiency, reuse of pre-trained representations, carbon-aware scheduling, and distillation into smaller deployment models. These choices are not merely operational; they reflect environmental governance priorities in the AI and automotive sectors.

Infrastructure for post-training includes simulation clusters, data pipelines, version control systems, validation tracks, and over-the-air update mechanisms. Cities and regions may need to invest in digital twins, communication networks, and roadside sensing to support safe deployment. However, infrastructure investment can exacerbate inequalities if only affluent areas receive high-quality mapping and connectivity. Policy frameworks should encourage interoperability, open standards, and public interest data sharing where appropriate [19]. The goal is not only to optimize models but to build an infrastructure that supports equitable and resilient mobility.

7. Deployment and Policy Implications

Deployment of post-trained end-to-end driving models should be staged and conditional. A policy that performs well in simulation should be tested in closed courses, then in geofenced low-speed environments, then in progressively more complex settings. Each stage should have clear entrance and exit criteria, including safety metrics, disengagement rates, and human factors evaluations. Regulatory sandboxes can provide controlled environments for learning while limiting public exposure to risk [19]. Post-training should continue during deployment, but updates should be subject to change management, impact assessment, and rollback capabilities.

International harmonization is desirable but difficult because legal cultures, road norms, and privacy regimes differ. A post-training pipeline developed in one country may not transfer directly to another. Localization may require region-specific reward models, scenario libraries, and stakeholder engagement. Standards bodies can help by defining common terminology, documentation formats, and safety assurance methods [20]. However, harmonization should not become a race to the lowest common denominator. Minimum safety and transparency requirements should be maintained across jurisdictions.

Incident reporting and continuous monitoring are essential policy tools. When a post-trained model is involved in a collision or near miss, investigators need access to logs, model versions, reward specifications, and simulation histories. Reporting should be timely, standardized, and privacy-preserving. Regulatory agencies may also require periodic audits of reward models and training data to detect bias, reward hacking, or unsafe optimization [17]. Such audits can improve public confidence and create incentives for responsible post-training practices.

8. Evaluation and Open Challenges

Evaluation of post-trained driving policies remains an open challenge. Benchmarks such as CARLA provide reproducible scenarios, but they cannot capture the full diversity of real-world traffic [3]. Metrics for safety, comfort, progress, and rule compliance are often aggregated into single scores that obscure trade-offs. A policy may achieve high average performance while failing catastrophically in rare but consequential situations. Evaluation should therefore include worst-case analysis, adversarial testing, and scenario-based certification. It should also examine how the policy behaves under uncertainty and whether it knows when to request human intervention.

Reward hacking and value misalignment are persistent risks. A policy may optimize for simulator-specific cues, exploit tolerant evaluation metrics, or learn socially undesirable behaviors that are not captured by the reward function [17,18]. Mitigation strategies include constrained reinforcement learning, safety critics, ensemble uncertainty, and human oversight. Yet these methods introduce their own complexity and may create new failure modes. Post-training research must therefore treat safety as an adversarial problem, not merely an optimization target.

Simulation-to-reality transfer remains a fundamental challenge. Domain randomization, system identification, and real-world fine-tuning can reduce the gap, but they cannot eliminate it [4,7]. A promising direction is to combine simulation-based post-training with small amounts of carefully selected real-world data, prioritizing edge cases and disengagements. This requires robust data curation and active learning. It also requires infrastructure for safe real-world experimentation, including test tracks, safety drivers, and regulatory approval.

Human oversight is another open challenge. As policies become more capable, human operators may become complacent or lose situational awareness. Post-training should therefore consider human-automation interaction, handover protocols, and explainability. Preference-based methods can align policies with human values, but they must be transparent about whose values are encoded and how conflicts are resolved [14,15,16]. The ultimate goal is not full autonomy without accountability but a system in which human institutions retain meaningful control over acceptable risk.

9. Conclusion

Reinforcement learning-based post-training offers a powerful pathway for improving end-to-end autonomous driving models beyond imitation. It can refine policies in rare scenarios, incorporate human preferences, and optimize long-horizon objectives that demonstrations alone cannot specify. However, post-training is not a simple model update. It is a system-level intervention that affects architecture, simulation infrastructure, compute provisioning, validation, governance, and public trust. The structural trade-offs are substantial: greater adaptability may reduce interpretability, simulation efficiency may mask reality gaps, and aggressive optimization may undermine safety.

This paper has argued for an integrated approach in which reward design, safety constraints, human feedback, and deployment monitoring are co-engineered. Post-training pipelines should be auditable, resource-aware, and socially legitimate. They should include diverse scenario libraries, robust evaluation, staged deployment, and clear accountability mechanisms. Policy frameworks should support transparency, incident reporting, and equitable infrastructure investment. Without these elements, reinforcement learning post-training may improve aggregate performance while creating new forms of opacity and risk.

Future research should develop methods for constrained and interpretable post-training, efficient simulation-to-reality transfer, and participatory reward modeling. It should also examine the environmental and distributive consequences of large-scale training. End-to-end autonomous driving is ultimately a socio-technical infrastructure, not merely a machine learning problem. Its success will depend on whether technical optimization is matched by institutional learning, public accountability, and a commitment to safety as a shared social value.

References

1. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., Jackel, L. D., Monfort, M., Muller, U., Zhang, J., Zhang, X., Zhao, J., & Zieba, K. (2016). End to end learning for self-driving cars. arXiv. <https://arxiv.org/abs/1604.07316>
2. Codevilla, F., Müller, M., López, A., Koltun, V., & Dosovitskiy, A. (2018). End-to-end driving via conditional imitation learning. In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 4693–4700). IEEE.
3. Dosovitskiy, A., Ros, G., Codevilla, F., López, A., & Koltun, V. (2017). CARLA: An open urban driving simulator. In Proceedings of the 1st Annual Conference on Robot Learning (pp. 1–16).
4. Kendall, A., Hawke, J., Janz, D., Mazur, P., Reda, M., Allen, J.-M., Lam, V.-D., Bewley, A., & Shah, A. (2019). Learning to drive in a day. In 2019 International Conference on Robotics and Automation (ICRA) (pp. 8248–8254). IEEE.
5. Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2017). On a formal model of safe and scalable self-driving cars. arXiv. <https://arxiv.org/abs/1708.06374>
6. Liang, X., Wang, T., Yang, L., & Xing, E. (2018). CIRL: Controllable imitative reinforcement learning for vision-based self-driving. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 584–599).
7. Amini, A., Gilitschenski, I., Phillips, J., Moseyko, J., Banerjee, R., Karaman, S., & Rus, D. (2020). Learning robust control policies for end-to-end autonomous driving from data-driven simulation. IEEE Robotics and Automation Letters, 5(2), 1143–1150.
8. Chitta, K., Prakash, A., Jaeger, B., Yu, Z., Renz, K., & Geiger, A. (2023). TransFuser: Imitation with transformer-based sensor fusion for autonomous driving. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(11), 12878–12895.
9. Chen, L., Wu, P., Chitta, K., Jaeger, B., Geiger, A., & Li, H. (2024). End-to-end autonomous driving: Challenges and frontiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12), 10164–10183.
10. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv. <https://arxiv.org/abs/1707.06347>
11. Yang, R., Wang, M., Chen, Y., Guo, T., Xu, Y., Cui, C., ... & Su, L. (2026). Post-Training in End-to-End Autonomous Driving. arXiv preprint arXiv:2607.08072.
12. 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.
13. 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.
14. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. In Advances in Neural Information Processing Systems (Vol. 30).

15. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Aspell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 27730–27744).
16. Bai, Y., Jones, A., Ndousse, K., Aspell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Shouk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., ... & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv*. <https://arxiv.org/abs/2204.05862>
17. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv*. <https://arxiv.org/abs/1606.06565>
18. Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
19. U.S. Department of Transportation. (2020). *Ensuring American leadership in automated vehicle technologies: Automated vehicles 4.0*. U.S. Department of Transportation.
20. SAE International. (2021). *Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles (J3016_202104)*. SAE International.