

Reinforcement Learning-Based Post-Training for Robust End-to-End Autonomous Driving Under Dynamic Traffic Conditions

Tage Haraldsson

Department of Computer Science, University of North Texas, Denton, TX, USA.

tage.work@unt.edu

Sorcha Haran

School of Computing, Clemson University, Clemson, SC, USA.

haransorcha@clemson.edu

Abstract

End-to-end autonomous driving has advanced rapidly by coupling perception, prediction, planning, and control within unified neural architectures, yet its deployment under dynamic traffic conditions remains constrained by brittleness, distributional shift, and limited assurance. This paper examines reinforcement learning-based post-training as a system-level intervention for robust end-to-end driving. Rather than treating post-training as merely a final optimization step, the analysis positions it within a broader socio-technical architecture involving data pipelines, simulation infrastructure, safety governance, validation regimes, and operational oversight. The paper argues that post-training can improve robustness when it addresses rare interactive scenarios, reward misspecification, and long-horizon credit assignment, but that these benefits depend on careful separation between online adaptation and safety certification. Dynamic traffic conditions are analyzed as compound stressors that combine perceptual ambiguity, social negotiation, adversarial behavior, and infrastructure variability. The discussion considers architectural trade-offs among modular and monolithic policies, infrastructure requirements for replay and simulation, deployment constraints, sustainability costs, fairness concerns, and policy implications. Evaluation is framed as a multi-dimensional problem spanning stress testing, counterfactual simulation, edge-case discovery, and human factors. The conclusion is that reinforcement learning-based post-training is promising but not sufficient by itself; robust autonomy requires layered assurance, traceable adaptation, and governance institutions capable of managing continual learning in public spaces.

Keywords

reinforcement learning; autonomous driving; post-training; robust control; dynamic traffic; socio-technical systems; safety governance; deployment.

1. Introduction

End-to-end autonomous driving represents a significant departure from modular autonomy stacks by training a unified policy to map sensory inputs directly to driving actions or trajectories [1], [2]. Early neural approaches demonstrated that vehicles could learn lane following and basic control from demonstrations, while later conditional imitation learning showed that high-level commands and navigational intent could be integrated into end-to-end policies [3]. The appeal of such architectures lies in their capacity to reduce hand-engineered

interfaces, exploit large-scale data, and optimize driving behavior holistically rather than through disconnected perception, prediction, and planning modules. However, this appeal also introduces a central difficulty: end-to-end policies are often opaque, data-hungry, and sensitive to distributional shift, especially when deployed in dynamic traffic environments that differ from training conditions.

Dynamic traffic conditions intensify these difficulties because they are not simply visual perturbations. They involve social negotiation, uncertain intent, interacting agents, temporary rule changes, adverse weather, construction zones, emergency vehicles, and mixed human and automated behavior. A policy trained primarily through imitation may reproduce average human behavior without understanding when to deviate from it, while a policy optimized only for nominal simulation may fail when real-world agents behave irrationally or strategically [15]. Reinforcement learning offers a possible remedy because it can optimize long-horizon outcomes through interaction, reward signals, and exploration [8], [9]. Yet applying reinforcement learning directly to safety-critical driving is difficult due to sample inefficiency, reward design, exploration risk, and the impossibility of unrestricted real-world trial and error [14].

Reinforcement learning-based post-training has therefore emerged as a promising middle path. In this framing, a model is first trained through supervised learning, imitation, or self-supervised representation learning, and then refined through reinforcement learning objectives that target robustness, rare events, comfort, or rule compliance [18]. Post-training is not a single algorithm but a family of system practices, including offline reinforcement learning, simulation-based fine-tuning, preference optimization, reward-model-guided adaptation, and constrained policy improvement. The central claim of this paper is that the value of post-training depends less on any single learning algorithm than on the surrounding architecture of data generation, safety constraints, validation, deployment, and governance. Robustness under dynamic traffic is thus a system property, not merely a model property.

2. Foundations of End-to-End Autonomous Driving

End-to-end driving systems typically combine perception, representation learning, and control within a differentiable or jointly optimized pipeline. The historical trajectory from ALVINN to modern deep driving policies illustrates a shift from small neural controllers to large-scale multimodal architectures trained on massive datasets [1], [2]. Contemporary systems often use convolutional networks, recurrent layers, transformers, or hybrid designs to extract spatial and temporal features from cameras, lidar, radar, and map inputs [16], [17]. These representations are then decoded into steering, acceleration, trajectory waypoints, or discrete driving decisions. The advantage is that the system can learn features that are directly useful for control rather than optimizing intermediate tasks that may not align with driving performance.

Public datasets and simulators have played a foundational role in this development. KITTI provided early benchmarks for vision-based autonomous driving, while nuScenes and the Waymo Open Dataset expanded the scale and diversity of multimodal data [5], [6], [7]. Simulators such as CARLA enabled controlled experimentation with urban driving scenarios, sensor configurations, and repeatable evaluation [4]. These infrastructures make it possible to train and compare end-to-end policies, but they also shape the kinds of robustness that can be measured. A dataset collected in one city, weather pattern, or traffic culture may encode biases that are invisible in benchmark performance but consequential in deployment. Simulation can reduce cost and risk, yet the sim-to-real gap remains a persistent challenge

because real traffic contains social cues, sensor degradation, and long-tail interactions that are difficult to reproduce faithfully.

The architectural debate between modular and end-to-end systems is also relevant to post-training. Modular stacks can localize failures and enforce explicit rules, but they may suffer from interface loss and brittle hand-offs between perception, prediction, and planning. End-to-end systems can optimize the whole driving objective but are difficult to interpret and certify. Post-training does not resolve this tension automatically. Instead, it can either reinforce monolithic opacity or be embedded within a layered architecture that preserves interpretable safety monitors, fallback controllers, and runtime verification. The more defensible position is that end-to-end learning should be paired with external constraints and governance mechanisms that bound its behavior even when internal representations are not fully understood [15].

3. Reinforcement Learning and Post-Training as a Systems Intervention

Reinforcement learning provides a formal framework for sequential decision making under uncertainty, where an agent learns from rewards obtained through interaction with an environment [9]. Deep reinforcement learning demonstrated that high-dimensional inputs could be connected to control policies through neural networks, beginning with breakthroughs in game playing and continuous control [8], [11], [13]. Algorithms such as proximal policy optimization and soft actor-critic improved stability, sample efficiency, and exploration in complex domains [10], [12]. These advances are relevant to autonomous driving because driving is inherently sequential, interactive, and long-horizon. A vehicle must not only recognize the current scene but also anticipate how its actions will influence other agents and future traffic states.

Post-training uses reinforcement learning after an initial supervised or imitation phase. This ordering matters because pure reinforcement learning from scratch in driving is generally infeasible due to safety and sample complexity. A pre-trained policy can provide a reasonable prior, and post-training can then refine behavior in simulation, replay, or limited real-world settings [18]. The objectives may include reducing collisions, improving comfort, maintaining lane discipline, respecting right-of-way, or recovering from rare events. Offline reinforcement learning can use logged data without online exploration, while simulation-based post-training can generate controlled interactions with scripted or learned agents. Preference-based methods can incorporate human judgments about safety and comfort, and constrained reinforcement learning can penalize unsafe actions during optimization. The system-level question is how these methods are integrated with data provenance, reward validation, safety envelopes, and deployment gates.

Reward design is a critical vulnerability. A reward that emphasizes progress may encourage aggressive behavior, while a reward that emphasizes caution may produce overly timid vehicles that disrupt traffic. Sparse collision penalties may fail to shape behavior before failure, and dense comfort penalties may suppress necessary evasive action. In dynamic traffic, the same action can be safe in one context and dangerous in another, so reward functions must be context-sensitive and validated against normative driving rules [15]. Post-training therefore cannot be treated as an objective engineering exercise alone. It requires institutional processes for defining acceptable behavior, auditing reward models, and detecting specification gaming. The presence of human oversight, red-team testing, and scenario-based validation becomes part of the learning architecture rather than an external compliance layer.

4. Dynamic Traffic Conditions and Robustness Challenges

Dynamic traffic conditions are best understood as compound stressors rather than isolated environmental variations. Perceptual stressors include rain, fog, glare, occlusion, sensor noise, and degraded lane markings. Behavioral stressors include cut-ins, aggressive drivers, jaywalkers, cyclists, and emergency vehicles. Infrastructural stressors include temporary signage, construction barriers, inconsistent lane geometry, and communication failures. Social stressors include ambiguous right-of-way, negotiation at intersections, and informal norms that vary across regions. A robust end-to-end policy must handle combinations of these stressors, not merely each in isolation. This combinatorial complexity makes exhaustive testing impossible and increases reliance on simulation, scenario generation, and adaptive safety margins.

Reinforcement learning post-training can improve robustness by exposing policies to adversarial or rare scenarios during training. For example, simulation can generate cut-ins, sudden braking, pedestrian occlusions, and noncompliant behavior that are underrepresented in naturalistic driving data [4]. Learned traffic agents can also create more realistic interactions than scripted actors, enabling policies to develop negotiation strategies [19]. However, simulated adversarial scenarios may overfit to the simulator’s physics or behavioral models. If the simulation is too benign, post-training provides false confidence; if it is too adversarial, the policy may become excessively defensive or unstable. The challenge is to construct a curriculum that balances realism, diversity, and safety-criticality while preserving performance in ordinary traffic.

Robustness also has temporal and social dimensions. A policy that is safe for a single decision may create risk over a sequence of interactions, such as edging into an intersection and blocking emergency access. A policy that is locally polite may be globally inefficient if it yields too often and causes congestion. Reinforcement learning can optimize long-horizon returns, but credit assignment over long traffic episodes remains difficult. Moreover, robustness must account for other agents who adapt to the autonomous vehicle. Humans may behave differently around automated vehicles, sometimes testing them or exploiting their caution. Post-training should therefore include opponent modeling, population-based training, or game-theoretic considerations rather than assuming a fixed environment. The resulting system is not merely a controller but a participant in a dynamic social system.

5. Architecture and Infrastructure for Post-Training

The infrastructure required for reinforcement learning-based post-training is substantial. It includes data ingestion from fleets, scenario mining, simulation clusters, replay buffers, reward computation, policy training, validation suites, and deployment pipelines. Real-world driving logs provide valuable coverage but are biased toward safe and routine conditions. Simulation provides scalability but requires maps, sensor models, agent behavior models, and physics engines that approximate reality [4]. A robust post-training architecture should combine offline data, closed-loop simulation, and selective on-road testing under supervision. The goal is not to replace one data source with another but to create a feedback loop in which failures discovered in deployment become new training scenarios.

Architecturally, post-training can be centralized, decentralized, or hybrid. Centralized training allows consistent reward definitions, safety constraints, and version control, but it may struggle with regional variation and latency. Decentralized adaptation can tailor policies to local traffic norms, but it risks fragmentation and inconsistent safety outcomes. A hybrid

approach may use a shared foundation model with localized post-training layers that are bounded by global safety constraints. This mirrors broader debates in machine learning about foundation models, fine-tuning, and edge adaptation [16], [17]. For autonomous driving, the key is traceability: every adaptation should be linked to data, objectives, validation results, and approval decisions. Without such traceability, post-training can become an uncontrolled drift process that undermines accountability.

Compute and energy infrastructure also matter. Reinforcement learning post-training often requires extensive simulation and repeated policy evaluation, which can be costly and carbon-intensive. If adaptation is continuous, the environmental footprint may grow with fleet size and scenario complexity. Sustainable deployment therefore requires efficient sampling, prioritized replay, model reuse, and simulation fidelity that is sufficient without being wasteful. The trade-off is between robustness gained through massive scenario exposure and the energy, cost, and time required to generate that exposure. A systems perspective should evaluate not only final driving performance but also the lifecycle costs of data collection, training, validation, and updates.

6. Safety, Governance, and Policy Implications

Safety assurance for post-trained end-to-end driving cannot rely solely on aggregate test metrics. It requires layered governance that includes runtime monitors, operational design domain limits, fallback behaviors, incident reporting, and independent audit. Formal models of safe self-driving can provide normative constraints, such as responsibility-sensitive safety, that bound permissible actions even when the learned policy is uncertain [15]. These constraints can be enforced outside the neural policy or integrated as shielding mechanisms during post-training. The purpose is not to eliminate learning but to ensure that learning cannot override basic safety invariants. In dynamic traffic, such invariants must be context-sensitive and legally defensible.

Policy and regulation must also address continual learning. If a vehicle’s policy can be updated after deployment, then regulators need mechanisms for version approval, change impact assessment, and recall. The analogy to aviation software is instructive: updates are controlled, documented, and tested before release. Autonomous driving, however, may require faster adaptation to local conditions and new scenarios. This creates a tension between responsiveness and stability. A governance framework could require that post-training occurs in a sandbox, that updates are signed and auditable, and that safety monitors remain independent of the learned policy. It could also mandate transparency about the conditions under which post-training data was collected and the distributional assumptions of validation.

Fairness and accountability add another layer. If post-training prioritizes scenarios common in wealthy or well-mapped regions, vehicles may perform worse in underrepresented neighborhoods, rural roads, or informal settlements. If reward functions encode aggressive driving norms, they may disadvantage pedestrians, cyclists, or older adults. Governance should therefore include distributional impact assessment, community consultation, and disaggregated performance reporting. The socio-technical system must be designed so that responsibility is not diffused across data providers, model developers, fleet operators, and regulators. Clear liability chains and incident investigation procedures are necessary for public trust.

7. Sustainability, Fairness, and Deployment

Deployment of post-trained autonomous driving systems involves trade-offs among safety, efficiency, equity, and environmental impact. Reinforcement learning can improve traffic flow by reducing unnecessary braking and smoothing acceleration, but it can also encourage policies that optimize individual vehicle progress at the expense of system-level congestion. If all vehicles are post-trained to be assertive, intersection throughput may improve in simulation but conflict risk may rise in mixed traffic. Conversely, if all vehicles are excessively cautious, deadlocks and delays may increase. The design of reward functions and safety constraints therefore has collective consequences that cannot be evaluated only at the level of a single vehicle.

Equity concerns arise from data and infrastructure asymmetry. Regions with high-quality maps, stable connectivity, and well-maintained roads are easier to support. Regions with poor infrastructure, informal traffic rules, or limited digital connectivity may be excluded from the benefits of autonomy or exposed to higher risk. Post-training on local data can help, but it requires local data collection, maintenance, and validation capacity. Without deliberate investment, reinforcement learning post-training may reinforce existing inequalities in mobility and safety. A fair deployment strategy should include scenario libraries for diverse environments, multilingual and culturally aware interaction design, and accessible incident reporting.

Sustainability is also tied to system architecture. Cloud-based training and continuous simulation can consume significant energy, while edge inference must balance latency, power, and robustness. Post-training may be periodic rather than continuous to reduce cost, but periodic updates may lag behind changing traffic conditions. A more sustainable approach could use targeted scenario mining, active learning, and simulation reuse so that only informative data triggers retraining. The goal is to achieve robustness without unbounded computation. This requires measuring the marginal safety benefit of additional training against its energy and financial cost.

8. Evaluation and Case Illustrations

Evaluation of post-trained end-to-end driving should combine simulation, closed-course testing, on-road trials, and post-deployment monitoring. Simulation is essential for rare events and counterfactual analysis, but it must be validated against real-world distributions [4]. Datasets such as KITTI, nuScenes, and Waymo Open Dataset provide benchmarks and diversity, yet they are not substitutes for deployment-specific validation [5], [6], [7]. A robust evaluation suite should measure collision avoidance, rule compliance, comfort, progress, interaction quality, and recovery from uncertainty. It should also test for negative transfer, where post-training improves one scenario while degrading another.

Case illustrations from learning-based driving research show both promise and limitation. Learning to drive in a day demonstrated rapid adaptation in limited settings, but such demonstrations often rely on constrained routes and favorable conditions [19]. Learning by cheating showed how privileged information can train a sensorimotor policy that generalizes to deployment, but this also raises questions about the fidelity of privileged signals and the sim-to-real gap [20]. These examples suggest that post-training is most effective when it is embedded in a pipeline that includes representation learning, simulation, privileged supervision, and safety filtering. The lesson is not that one method solves robust autonomy, but that architectural integration determines whether learning gains translate into dependable behavior.

In dynamic traffic, evaluation must also consider human-robot interaction. Drivers may behave differently when they recognize an autonomous vehicle, and pedestrians may rely on eye contact or gesture that sensors do not fully capture. Post-training should therefore include human-in-the-loop simulation and field studies that examine how communication, predictability, and vehicle behavior affect cooperation. Metrics should capture not only whether a collision was avoided but whether the interaction was legible to others. A vehicle that is safe but unpredictable may still create risk by forcing humans into sudden evasive actions.

9. Conclusion

Reinforcement learning-based post-training offers a credible pathway for improving end-to-end autonomous driving under dynamic traffic conditions, but it is not a standalone solution. Its benefits depend on high-quality pre-training, realistic simulation, carefully governed reward design, constrained optimization, and layered safety assurance. The most important system-level insight is that robustness emerges from the interaction of models, data, infrastructure, institutions, and operational practices. Post-training can refine policies for rare and interactive scenarios, yet it can also introduce drift, reward gaming, and uneven performance if not bounded by traceable governance.

For deployment, the field should move toward architectures that separate adaptable learned behavior from non-negotiable safety constraints. Runtime monitors, fallback controllers, scenario-based validation, and independent audit should remain active throughout the lifecycle. Policy frameworks should require transparency about post-training data, objectives, and update procedures, while also supporting local adaptation and fairness across regions. Sustainability requires efficient use of simulation and compute, and evaluation must include social interaction, equity, and long-term system effects. The future of robust autonomous driving will therefore depend as much on governance and infrastructure as on algorithmic innovation.

References

1. Pomerleau, D. A. (1988). ALVINN: An autonomous land vehicle in a neural network. Carnegie Mellon University.
2. 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.
3. 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 (pp. 4693-4700). IEEE.
4. 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). PMLR.
5. Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? The KITTI vision benchmark suite. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 3354-3361). IEEE.
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. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., Vasudevan, V., Han, W., Ngiam, J., Zhao, H., Timofeev, A., Ettinger, S., Krivokon, M., Gao, A., Joshi, A., Zhang, Y., Shlens, J., Chen, Z., & Anguelov, D. (2020). Scalability in perception for autonomous driving: Waymo open dataset. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2446-2454). IEEE.
 8. 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.
 9. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.
 10. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347.
 11. Lillicrap, T. P., Hunt, J. J., Pritzel, A., Heess, N., Erez, T., Tassa, Y., Silver, D., & Wierstra, D. (2015). Continuous control with deep reinforcement learning. arXiv preprint arXiv:1509.02971.
 12. Haarnoja, T., Zhou, A., Abbeel, P., & Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In Proceedings of the 35th International Conference on Machine Learning (pp. 1861-1870). PMLR.
 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. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
 15. Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2017). On a formal model of safe and scalable self-driving cars. arXiv preprint arXiv:1708.06374.
 16. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in Neural Information Processing Systems (pp. 5998-6008). Curran Associates.
 17. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444.
 18. Chen, D., Zhou, B., Koltun, V., & Krähenbühl, P. (2020). Learning by cheating. In Proceedings of the Conference on Robot Learning (pp. 66-75). PMLR.
 19. Kendall, A., Hawke, J., Janz, D., Mazur, P., Reda, D., Allen, J., Lam, V. D., Qiao, K., & Shah, A. (2019). Learning to drive in a day. In 2019 International Conference on Robotics and Automation (pp. 8248-8254). IEEE.