

Multi-Agent Learning for Post-Trained Autonomous Driving Decision Making

Abdirahman Helgesson

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

ahelgesson@ku.edu

Simo Huhtala

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

simo1996@buffalo.edu

Abstract

Autonomous driving decision making is increasingly framed as a socio-technical problem in which perception, prediction, planning, and control must operate under uncertainty, regulatory scrutiny, and heterogeneous human expectations. This paper examines multi-agent learning for post-trained autonomous driving decision making, treating post-training as the stage in which pretrained representation models are adapted toward instruction following, preference alignment, safety constraints, and interactive coordination. The central argument is that multi-agent learning is not merely a technical extension of single-vehicle reinforcement learning but a system-level architecture that redistributes intelligence across vehicles, roadside infrastructure, fleet operations, simulation environments, and governance institutions. Such redistribution creates important trade-offs involving latency, robustness, interpretability, fairness, energy consumption, liability, and long-term maintainability. The paper synthesizes foundational work on multi-agent reinforcement learning, end-to-end driving, imitation learning, and post-training alignment, then discusses deployment architectures that combine centralized training with decentralized execution, mixed cooperative-competitive interaction, and human-in-the-loop oversight. It also considers validation, edge infrastructure, sustainability, and policy implications. The analysis suggests that post-trained multi-agent driving systems can improve adaptability and coordination, but only when their value conflicts, failure modes, and accountability pathways are designed explicitly rather than treated as downstream engineering concerns. The conclusion outlines research directions for safe, equitable, and sustainable deployment.

Keywords

multi-agent learning, post-training, autonomous driving, decision making, socio-technical systems, safety, governance, sustainability.

1. Introduction

Autonomous driving decision making has moved from modular pipelines toward learned systems that integrate perception, prediction, and planning in increasingly unified representations. Early safety models emphasized formal rules and verifiable constraints, while later deep learning approaches demonstrated that large-scale data could support end-to-end control in constrained settings [1]. Reinforcement learning further showed that sequential decision making could be optimized through interaction, reward design, and simulation at scale [2]. Self-play methods then revealed that competitive and cooperative dynamics could

produce superhuman behavior in complex games, suggesting analogous possibilities for traffic interaction [3]. Multi-agent reinforcement learning extended these insights to settings where multiple learners influence one another, creating non-stationary environments and coordination problems that single-agent methods cannot fully capture [4]. In autonomous driving, these developments converge on a practical question: how should post-trained decision systems coordinate with other vehicles, pedestrians, cyclists, roadside units, fleet managers, and regulators without sacrificing safety or accountability.

The phrase post-training refers to the adaptation of pretrained models after initial representation learning. In language and multimodal systems, post-training often includes instruction tuning, preference optimization, reinforcement learning from human feedback, and safety filtering [14]. Similar techniques are now being explored for driving, where a foundation model may be pretrained on large corpora of sensor data, driving logs, maps, and behavioral traces, then post-trained to follow traffic norms, respond to rare events, and align with fleet or jurisdictional policies [13]. This paper treats post-training not as a single algorithm but as a socio-technical layer that mediates between general capabilities and context-specific obligations. When combined with multi-agent learning, post-training becomes a mechanism for shaping interaction protocols, negotiation styles, risk tolerances, and escalation behavior across a population of vehicles.

A central challenge is that autonomous driving is not a closed game. Roads are shared infrastructures populated by humans, machines, animals, weather, construction, emergency vehicles, and legal regimes. Multi-agent learning can improve coordination, but it can also amplify strategic behavior, inequitable access, and opaque failure modes. If one fleet learns aggressive merging while another learns conservative yielding, the resulting equilibrium may be unstable or socially undesirable. If roadside infrastructure participates as an agent, questions arise about who controls it, who pays for it, and who bears liability when its recommendations fail. Therefore, the study of multi-agent post-trained driving must include architecture, governance, deployment, sustainability, robustness, and fairness from the outset. This paper develops that system-level perspective and proposes that post-training should be understood as a governance-sensitive design practice rather than only a model optimization step.

2. Foundations of Multi-Agent Learning and Post-Training

Multi-agent learning concerns systems in which multiple decision makers learn concurrently, often with partial observability and mixed incentives. The mixed cooperative-competitive setting is particularly relevant to traffic because drivers sometimes cooperate to merge, sometimes compete for right of way, and sometimes coordinate implicitly through signals, motion, and social convention. Multi-agent actor-critic methods demonstrated that centralized training with decentralized execution could stabilize learning in such environments [4]. Counterfactual policy gradients further addressed credit assignment by asking how one agent’s action contributed to a joint outcome relative to alternatives [5]. Value factorization methods such as QMIX introduced monotonic decomposition so that decentralized policies could be derived from joint value estimates [6]. These methods are technically important, but for autonomous driving they must be interpreted through infrastructure and governance, because the meaning of a reward is not given by the road alone.

Large-scale multi-agent environments have provided evidence that coordination can emerge under pressure. StarCraft II required long-horizon planning, partial observability, and real-time adaptation, and multi-agent reinforcement learning achieved grandmaster-level play [7].

Dota 2 similarly showed that large-scale deep reinforcement learning could support complex team coordination over extended matches [8]. These cases are not direct templates for driving, but they clarify that multi-agent learning can handle rich interaction when simulation, compute, and reward design are carefully engineered. They also expose limitations: game environments have clear boundaries, resettable episodes, and measurable victory conditions, whereas road systems have open-ended moral, legal, and economic stakes. The transfer from games to streets therefore requires more than scaling; it requires institutional embedding.

End-to-end driving research offers complementary lessons. Conditional imitation learning showed that driving policies could be conditioned on navigational commands and visual inputs [9]. Learning by cheating demonstrated that privileged information during training could improve a student policy that later relies only on onboard sensors [10]. Early demonstrations of learning to drive in a day illustrated the appeal and fragility of rapid end-to-end adaptation [11]. CARLA and similar simulators became essential infrastructures for reproducible experimentation, scenario generation, and safety testing [12]. Post-training builds on these foundations by taking pretrained driving representations and aligning them with preferences, instructions, and constraints [13]. In language models, reinforcement learning from human feedback and related alignment methods showed how pretrained capabilities could be steered toward helpful and harmless behavior [14], [15], [16]. The analogy for driving is powerful but imperfect, because human preferences about driving are distributed across pedestrians, passengers, insurers, regulators, and communities, not merely individual users.

3. Architectural Patterns and Decision Layers

A multi-agent post-trained driving architecture typically separates representation learning, interaction modeling, policy adaptation, and runtime arbitration. The pretrained component may encode sensor data, maps, traffic rules, and behavioral priors. The post-training component may then refine decision policies through demonstrations, preferences, simulations, and constrained optimization. Multi-agent interaction can be handled through centralized training with decentralized execution, where agents share a critic during training but act independently at deployment [4]. Alternatively, value factorization can allow each vehicle to select actions based on local observations while preserving some global coordination signal [6]. These patterns reduce communication dependence but do not eliminate the need for shared conventions, because decentralized actors must still predict one another and resolve ambiguity.

At the system level, the architecture must decide where learning occurs and where authority resides. A fully onboard system offers low latency and resilience to communication failure, but it limits fleet-level coordination. A cloud-trained fleet model can incorporate diverse data and update globally, but it introduces versioning, privacy, and update-consistency problems. Roadside edge infrastructure can provide local coordination, shared perception, and traffic optimization, yet it creates dependencies on public or private infrastructure that may be unevenly maintained [23]. Mixed architectures are likely, with safety-critical decisions remaining onboard and coordination functions distributed across fleet, edge, and regional services. The key design question is not whether to centralize or decentralize, but which decisions require which level of shared context and which failure modes are acceptable.

Post-training adds another layer of arbitration. A vehicle may have a base driving policy, a post-trained interaction style, and a runtime safety shield that overrides unsafe actions. In multi-agent settings, the interaction style of one vehicle affects the learning environment of

others. If fleets post-train on their own preference data, they may develop incompatible norms. If they post-train on shared data, they risk homogenizing behavior in ways that fail under local customs or unusual conditions. Governance mechanisms may therefore need to define minimum interoperability standards, conflict-resolution protocols, and audit trails. The architecture must record not only what action was taken but also which policy version, which post-training objective, and which external recommendation influenced the decision. Without such traceability, accountability becomes impractical after an incident [17], [18].

4. System-Level Trade-offs, Infrastructure, and Deployment

The deployment of multi-agent post-trained driving systems involves trade-offs among latency, bandwidth, compute, reliability, and cost. Onboard inference must operate under strict timing constraints, especially for braking, evasive maneuvers, and pedestrian protection. Cloud services can support heavy simulation, map updates, and fleet analytics, but they cannot be in the critical path for every millisecond-level decision. Edge computing can reduce latency for local coordination, but it requires roadside units, power, maintenance, and cybersecurity. The resulting infrastructure is not neutral; it shapes which neighborhoods receive enhanced coordination and which rely on less capable onboard systems [19]. Infrastructure equity therefore becomes part of the technical design space, not an external policy afterthought.

Sustainability concerns also enter the architecture. Larger models and more frequent post-training cycles increase energy consumption during training and over-the-air updates. Fleet-wide deployment multiplies inference costs across vehicles, and edge infrastructure adds continuous operational energy use. Efficiency techniques such as model compression, distillation, pruning, and event-triggered computation can reduce burden, but they may also reduce robustness in rare scenarios. The trade-off is not simply accuracy versus energy; it is also adaptability versus maintainability. A system that is too costly to update may become brittle, while a system that updates too often may create instability and validation debt [23]. Sustainable deployment therefore requires lifecycle thinking that includes data collection, training, validation, deployment, monitoring, and decommissioning.

The infrastructure layer also raises questions about interoperability and technical debt. Multi-agent systems depend on shared protocols for negotiation, priority, and intent sharing. If each manufacturer implements proprietary interaction styles, the road becomes a patchwork of incompatible conventions. If standards are too rigid, they may suppress innovation or fail to accommodate local conditions. Hidden technical debt accumulates when data dependencies, configuration choices, feedback loops, and monitoring gaps are left unmanaged [23]. Post-training can exacerbate this debt because preference data, reward models, and safety filters may be updated independently of the base driving stack. A disciplined deployment process must treat these components as versioned infrastructure with clear ownership, testing, and rollback paths.

5. Robustness, Safety, and Validation

Safety in multi-agent driving cannot rely solely on average performance metrics. Rare events, adversarial interactions, sensor degradation, and communication loss can produce catastrophic failures. Formal safety models emphasize verifiable constraints and conservative behavior, but they often struggle with open-world perception and social negotiation [1]. Learned systems can generalize across complex patterns, but they may fail unpredictably when other agents behave outside the training distribution. Post-training can improve alignment with

desired behavior, yet it can also overfit to preference proxies that do not capture true safety. Validation must therefore combine simulation, closed-course testing, shadow mode deployment, and carefully staged real-world trials [12]. The multi-agent dimension makes validation harder because the system under test changes when other agents adapt to it.

Robustness also depends on how uncertainty is represented and acted upon. A vehicle must distinguish between uncertainty about perception, uncertainty about another agent's intent, and uncertainty about its own policy. In multi-agent settings, these uncertainties interact. A pedestrian may hesitate because the vehicle hesitates, creating a feedback loop that neither agent can resolve without a convention. Post-training can encode conventions through demonstrations or preferences, but conventions vary by culture, weather, road type, and time of day. A robust system must detect when its learned conventions are inappropriate and default to conservative, legible behavior. This is both a technical and a communicative requirement, because other drivers and pedestrians need to interpret the vehicle's intentions [17].

Human oversight remains essential, but it must be designed realistically. Remote operators cannot supervise every vehicle continuously. Instead, oversight may focus on edge cases, fleet-level anomalies, and post-incident review. Post-training can support escalation policies by teaching vehicles when to request help, when to pull over, and when to prioritize minimal-risk maneuvers. However, oversight introduces its own failure modes: communication delay, operator overload, ambiguous authority, and moral distress. A safety case must specify who is responsible for which decision, how the system detects that it is outside its operational design domain, and how it degrades gracefully. Without such clarity, multi-agent coordination may create diffuse responsibility that weakens accountability [18].

6. Fairness, Policy, and Sociotechnical Governance

Autonomous driving systems embody values through their design choices, even when those choices appear purely technical [20]. A policy that minimizes travel time may disadvantage pedestrians; a policy that maximizes caution may impose delays on other drivers; a policy that prioritizes fleet efficiency may externalize risk onto non-users. Multi-agent learning makes these value conflicts explicit because different agents may optimize different objectives. Post-training can incorporate fairness constraints, but fairness itself is contested. Does fairness mean equal risk across road users, equal access to mobility, equal treatment by enforcement, or equitable distribution of infrastructure investment? These are political and ethical questions that cannot be resolved by model architecture alone [21], [22].

Governance must therefore operate across multiple levels. At the vehicle level, regulators may require logging, explainability, fail-safe behavior, and cybersecurity standards. At the fleet level, operators may need to demonstrate non-discriminatory dispatch, transparent incident reporting, and fair labor practices for remote support. At the infrastructure level, public agencies may need to decide whether roadside coordination units are public utilities, regulated monopolies, or private services. At the international level, interoperability and liability regimes must be negotiated so that vehicles crossing borders do not face incompatible expectations. The history of technology shows that artifacts can have politics because they distribute power and constrain behavior [19]. Multi-agent driving infrastructures are no exception.

Fairness also involves data and participation. Post-training depends on preference data, demonstration data, and simulation scenarios. If these data reflect only certain driving cultures,

weather conditions, or socioeconomic contexts, the resulting policies may perform unevenly. If communities are not included in scenario design, their safety concerns may be overlooked. If fleet data are proprietary, independent researchers and regulators may struggle to audit behavior. A more legitimate approach would combine privacy-preserving data sharing, public scenario libraries, third-party auditing, and participatory governance. Such arrangements are slower and more complex than purely corporate deployment, but they may produce systems that are more robust, trusted, and sustainable over time.

7. Sustainability, Economic Viability, and Maintenance

The economic viability of multi-agent post-trained driving depends on more than model accuracy. Fleet operators must manage capital expenditure, insurance, energy, maintenance, connectivity, data storage, and regulatory compliance. Post-training adds costs for preference collection, simulation, red-teaming, validation, and continuous monitoring. These costs may be justified if they reduce incident rates, improve public trust, or enable higher utilization. However, they may also create barriers to entry that concentrate the market among a few large firms. Smaller operators may struggle to meet safety and reporting standards, while public agencies may lack the expertise to evaluate proprietary systems. The result could be an infrastructure that is efficient for some users but exclusionary for others.

Maintenance is a particularly underappreciated challenge. Multi-agent systems depend on shared protocols, map updates, simulation environments, and model versioning. When one component changes, the behavior of the whole system can shift. Post-training updates may improve one scenario while degrading another. Fleet-wide over-the-air updates can synchronize behavior, but they can also synchronize failure. A sustainable approach requires canary deployments, staged rollouts, rollback capability, and continuous monitoring of interaction effects. It also requires long-term support for older vehicles, because a mixed fleet of different generations will operate on the same roads. The technical debt literature warns that such dependencies become unmanageable without deliberate ownership and documentation [23]. Autonomous driving governance must treat maintenance as a first-class safety and equity concern.

8. Case Illustrations and Cross-Domain Comparisons

Robotaxi fleets provide a useful case because they combine centralized fleet management with decentralized street-level decision making. A robotaxi must negotiate with human drivers, pedestrians, cyclists, and other robotaxis. Multi-agent post-training could improve coordination by teaching vehicles to predict fleet behavior and to adopt consistent pickup and drop-off protocols. However, it could also produce collusion-like effects if fleet vehicles coordinate to reduce wait times at the expense of non-fleet traffic. Regulatory oversight would need to distinguish beneficial coordination from anti-competitive behavior. The case illustrates that multi-agent learning is not only a safety problem but also a market and governance problem.

Autonomous trucking offers a different configuration because routes are more structured, platooning is feasible, and operational domains are often limited to highways. Multi-agent methods can support platoon formation, spacing, fuel efficiency, and coordinated merging [7], [8]. Yet trucking also involves labor, port operations, warehouse interfaces, and long-haul infrastructure. Post-training must align with hours-of-service rules, cargo constraints, and maintenance schedules. A platoon that optimizes fuel may create risks for surrounding traffic if it forms overly long blocking patterns. Cross-domain comparison with air traffic control

suggests that centralized separation assurance can coexist with decentralized execution, but only when roles and fallback procedures are explicit. The analogy highlights the need for layered authority rather than monolithic autonomy.

Warehouse and logistics robotics provide a more controlled environment where multi-agent learning has already shown benefits in coordination and task allocation [14]. These systems operate in bounded spaces with known maps and limited human presence. Their success depends on infrastructure such as markers, sensors, and scheduling systems. By contrast, road systems lack such uniformity. The comparison suggests that multi-agent post-training is easiest where the environment is instrumented and rule-bound, and hardest where it is open, mixed, and political. This does not mean road deployment is impossible, but it does mean that infrastructure investment and governance must precede or accompany learning advances. The most successful deployments may therefore be those that treat the road as a partially managed system rather than an unconstrained game.

9. Future Directions

Future research should integrate post-training with continual learning, so that fleets can adapt to new conditions without forgetting safety-critical behavior. Federated and privacy-preserving approaches could allow multiple operators to share interaction insights without sharing raw data. Simulation must become more socially realistic, including pedestrians, cyclists, emergency vehicles, and non-compliant actors [12]. Evaluation should move beyond collision rates to include comfort, accessibility, environmental impact, and distributive fairness. Human factors research is needed to understand how people interpret and trust multi-agent vehicles, especially when those vehicles coordinate in ways that are invisible to human observers [15], [16]. Safety cases should be dynamic, updated as policies, infrastructure, and traffic compositions change.

Governance innovation is equally important. Regulators may need new tools for auditing learned policies, certifying post-training pipelines, and investigating multi-agent incidents [17], [18], [22]. Standards bodies should define interoperability protocols for intent sharing, priority negotiation, and fallback behavior. Insurers and courts will need frameworks for allocating liability when several autonomous agents, infrastructure providers, and human actors contribute to an outcome. Public participation can help surface value conflicts before they become accidents. The broader lesson from socio-technical systems is that technical capability and social legitimacy must co-evolve. Multi-agent post-trained driving will be safest and most sustainable when it is designed as public infrastructure with private participation, not as a private capability deployed on public roads [19], [20].

Conclusion

Multi-agent learning for post-trained autonomous driving decision making offers a powerful way to coordinate vehicles, infrastructure, and fleets under uncertainty. It can improve adaptability, support preference alignment, and enable richer interaction than single-agent pipelines. However, its benefits depend on architecture choices that balance onboard autonomy with shared coordination, validation that covers rare and adversarial interactions, and governance that clarifies responsibility across manufacturers, operators, regulators, and users. Post-training should not be treated as a final optimization step that solves safety and ethics after the fact. It is a governance-sensitive layer that shapes norms, conflicts, and accountability. The most credible path forward combines robust multi-agent methods with interoperable infrastructure, transparent evaluation, fair access, sustainable maintenance, and

inclusive policy design. Without these elements, coordination gains may be offset by fragility, inequity, and eroding public trust. With them, multi-agent post-training can contribute to driving systems that are not only intelligent but also legitimate, resilient, and socially sustainable.

References

1. Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2017). On a formal model of safe and scalable self-driving cars. arXiv preprint arXiv:1708.06374.
2. Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529-533.
3. Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., ... & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419), 1140-1144.
4. Lowe, R., Wu, Y., Tamar, A., Harb, J., Abbeel, P., & Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in Neural Information Processing Systems*, 30, 6379-6390.
5. Foerster, J., Farquhar, G., Afouras, T., Nardelli, N., & Whiteson, S. (2018). Counterfactual multi-agent policy gradients. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1), 2974-2982.
6. Rashid, T., Samvelyan, M., Schroeder de Witt, C., Farquhar, G., Foerster, J., & Whiteson, S. (2018). QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. *Proceedings of the 35th International Conference on Machine Learning*, 80, 4295-4304.
7. Vinyals, O., Babuschkin, I., Czarnecki, W. M., Mathieu, M., Dudzik, A., Chung, J., ... & Silver, D. (2019). Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782), 350-354.
8. Berner, C., Brockman, G., Chan, B., Cheung, V., Debiak, P., Dennison, C., ... & Zaremba, W. (2019). Dota 2 with large scale deep reinforcement learning. arXiv preprint arXiv:1912.06680.
9. 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*, 4693-4700.
10. Chen, D., Zhou, B., Koltun, V., & Krähenbühl, P. (2020). Learning by cheating. *Proceedings of the Conference on Robot Learning*, 100, 66-75.
11. Kendall, A., Hawke, J., Janz, D., Mazur, P., Reda, D., Allen, J. M., ... & Shah, A. (2019). Learning to drive in a day. *2019 International Conference on Robotics and Automation*, 8248-8254.
12. 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.
13. 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.

14. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730-27744.
15. Christiano, P. F., Leike, J., Brown, T., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 4299-4307.
16. Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., ... & Kaplan, J. (2022). Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
17. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.
18. Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.
19. Winner, L. (1980). Do artifacts have politics? *Daedalus*, 109(1), 121-136.
20. Nissenbaum, H. (2001). How computer systems embody values. *Computer*, 34(3), 120-127.
21. Friedman, B., & Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems*, 14(3), 330-347.
22. Barocas, S., Hardt, M., & Narayanan, A. (2019). *Fairness and machine learning: Limitations and opportunities*. fairmlbook.org.
23. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., ... & Dennison, D. (2015). Hidden technical debt in machine learning systems. *Advances in Neural Information Processing Systems*, 28, 2503-2511.
24. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25, 1097-1105.
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.