Uncertainty-Aware Post-Training of Autonomous Driving Models for Safety-Critical Decision-Making

Maria Belen Benet
Department of Computer Science, University of Houston, Houston, TX, USA.
mariabelenwork@uh.edu

Pauric Moher
Department of Computer Science, University of Central Florida, Orlando, FL, USA.
pauric2000@ucf.edu

Abstract

Autonomous driving models increasingly rely on large-scale representation learning, end-to-end policy learning, and data-driven perception to operate in open-world traffic environments. Yet safety-critical decision-making cannot be reduced to average-case predictive accuracy. This paper examines uncertainty-aware post-training as a system-level strategy for improving the safety, robustness, fairness, and governability of autonomous driving models after initial pre-training and before or during deployment. The argument is that post-training should not be treated merely as a final optimization stage for task performance, but as an architectural and institutional process that shapes how models represent ambiguity, defer to fallback systems, communicate confidence, and respond to distributional shift. The paper analyzes key sources of uncertainty in driving, including perceptual ambiguity, behavioral indeterminacy, environmental novelty, and socio-legal opacity. It then discusses how uncertainty-aware post-training can support calibrated decision thresholds, selective prediction, risk-sensitive policy adjustment, and human oversight. Particular attention is given to structural trade-offs among latency, compute cost, interpretability, and safety margins, as well as to the governance implications of deploying uncertainty-aware systems in mixed human and autonomous traffic. The paper also considers infrastructure requirements, evaluation regimes, sustainability concerns, and regulatory accountability. It concludes that uncertainty-aware post-training is necessary but insufficient unless integrated into broader assurance ecosystems that combine technical calibration, operational monitoring, ethical governance, and public accountability. Such integration is essential for autonomous driving systems that must act conservatively under ambiguity while remaining viable at scale.

Keywords

uncertainty-aware learning; post-training; autonomous driving; safety-critical decision-making; calibration; robustness; governance; socio-technical systems.

1. Introduction

Autonomous driving has become a central test case for artificial intelligence in safety-critical environments because it combines real-time perception, prediction, planning, and control under conditions that are only partially observable. A driving system must interpret noisy sensor data, infer the intentions of other road users, and select actions whose consequences may be irreversible. The literature on artificial intelligence safety has long warned that systems optimized only for average performance can exhibit brittle behavior in rare or adversarial situations [1]. This concern is especially acute in autonomous driving, where the

cost of a single miscalibrated decision can be measured in human injury, legal liability, and loss of public trust. Consequently, uncertainty-aware methods have emerged as a promising direction for improving model reliability, particularly when models must decide whether to act, slow down, request assistance, or enter a minimal-risk condition.

Uncertainty in machine learning is commonly divided into aleatoric and epistemic forms, where the former reflects irreducible noise in the data-generating process and the latter reflects ignorance about the model or environment [2]. In driving, aleatoric uncertainty may arise from sensor noise, occlusion, or ambiguous social cues, while epistemic uncertainty may arise from unfamiliar road layouts, rare weather conditions, or edge cases absent from training data. Bayesian approximations and approximate inference techniques have been used to estimate such uncertainty in deep models [3], and deep ensembles have provided a practical route to predictive uncertainty estimation without full Bayesian inference [4]. In the autonomous vehicle context, Bayesian deep learning has been proposed as a way to improve safety by identifying when a model is likely to fail and by enabling more cautious behavior under uncertainty [5].

However, uncertainty estimation alone does not guarantee safe decisions. A model may produce a well-calibrated uncertainty score yet still lack a policy for translating that score into a safe action. Formal approaches to self-driving safety have attempted to define rules and responsibility-sensitive safety frameworks that make conservative assumptions about other actors [6]. Reliability demonstrations based on accumulated mileage have also been questioned because the number of miles required to statistically validate rare-event safety may be impractically large [7]. These debates indicate that autonomous driving safety is not simply a perception problem or a control problem; it is an interdisciplinary challenge involving verification, validation, human factors, law, and public policy [8]. The central question of this paper is how uncertainty-aware post-training can be positioned within that broader system to improve safety-critical decision-making.

2. Safety-Critical Autonomy as a Socio-Technical System

Autonomous driving systems are best understood as socio-technical infrastructures rather than isolated software artifacts. Their behavior depends on sensors, maps, communication networks, compute platforms, maintenance regimes, regulatory frameworks, and the expectations of human road users. The operational design domain concept formalizes the conditions under which an automated driving system is intended to function, including road type, weather, speed range, and geographic constraints [9]. A model that performs well inside a narrow operational design domain may fail when the domain boundary is crossed, and post-training must therefore address not only nominal performance but also domain awareness and boundary detection. This is a system-level requirement because the detection of domain exit may depend on perception, localization, weather sensing, and map metadata working together.

End-to-end driving models further complicate the picture because they map sensor inputs directly to driving actions, often through learned representations rather than explicit modular interfaces [10]. Early demonstrations of end-to-end learning showed that neural networks could learn lane following and basic control from human driving data [11]. Subsequent work introduced conditional imitation learning to incorporate navigational commands and higher-level intent, enabling more structured behavior [10]. While end-to-end architectures can reduce engineering complexity and capture subtle correlations, they also make it harder to inspect intermediate reasoning and to assign responsibility when failures occur. Post-training

must therefore balance the flexibility of end-to-end learning with the need for interpretable safety constraints and fallback mechanisms.

Testing such systems is equally challenging. Rare-event simulation has been proposed as a way to generate safety-critical scenarios more efficiently than real-world mileage accumulation [12]. Yet simulated scenarios are only as good as the assumptions embedded in them, and a model that learns to satisfy a simulator may not transfer to the open world. From a socio-technical perspective, safety emerges from the interaction of technical performance, operational procedures, human oversight, and regulatory expectations. A vehicle that is uncertain may need to communicate that uncertainty to a remote operator, to passengers, or to nearby human drivers. The design of those communication channels is not merely a user interface issue; it affects the allocation of responsibility and the feasibility of graceful degradation. Thus, uncertainty-aware post-training must be aligned with the social and institutional architecture in which the vehicle operates.

3. Sources of Uncertainty in Driving Models

Uncertainty in autonomous driving arises from multiple layers of the perception-action pipeline. At the perceptual layer, object detection and tracking are affected by occlusion, lighting, weather, sensor degradation, and class ambiguity. Probabilistic object detection studies have shown that representing detection uncertainty can improve downstream reasoning, but they also reveal the difficulty of calibrating uncertainty across diverse conditions [15]. Anomaly detection research similarly emphasizes that out-of-distribution objects, unusual road debris, or atypical pedestrian behavior can fall outside the training distribution and produce overconfident predictions [16]. These are not merely technical edge cases; they are predictable features of open-world driving, which means that post-training must explicitly target uncertainty representation rather than relying on pre-training data alone.

At the behavioral layer, uncertainty arises because other road users are intentional agents whose future actions depend on social norms, local customs, and moment-to-moment negotiation. A driver may wave a pedestrian across, hesitate at an intersection, or violate a traffic rule to avoid a collision. These behaviors are not fully captured by deterministic trajectory prediction, and a model that treats them as noise may underestimate risk. Post-training can help by exposing models to diverse interaction scenarios and by optimizing for risk-sensitive objectives that penalize confident but dangerous predictions. Recent work on post-training in end-to-end autonomous driving is relevant here because it suggests that the stage after initial pre-training can be used to refine policy behavior, align representations with safety constraints, and improve decision-making under uncertainty [13]. The value of such post-training depends on whether it preserves calibration and whether it avoids overfitting to narrow safety benchmarks.

At the environmental and infrastructural layer, uncertainty is produced by changing road geometry, construction zones, sensor blind spots, and communication failures. Even high-definition maps can become stale, and localization can degrade in tunnels or urban canyons. The operational design domain concept is useful precisely because it makes these dependencies explicit [9]. However, a static operational design domain definition is insufficient for dynamic conditions; the system must monitor whether it remains inside the intended domain and adjust its behavior when it does not. Post-training can support this by teaching models to recognize conditions associated with elevated uncertainty and by coupling those estimates to conservative driving policies. The challenge is that uncertainty estimates

themselves can be miscalibrated, especially under distributional shift, so post-training must include evaluation methods that test calibration, not just accuracy [14].

4. Post-Training as an Architecture of Calibration

Post-training is often described as a final stage of model development, but in safety-critical autonomy it should be understood as an architecture of calibration. Calibration refers to the correspondence between predicted confidence and actual correctness, and it is essential when a system must decide whether to trust its own output. A model that is accurate on average but overconfident in rare situations is dangerous because it may fail to trigger fallback behavior. Conversely, a model that is overly conservative may become unusable because it constantly slows down or disengages. Post-training can adjust decision thresholds, temperature parameters, ensemble weighting, and policy constraints to achieve a better balance between assertiveness and caution. These adjustments are not merely numerical; they encode institutional risk preferences about what kinds of errors are acceptable and who bears the cost of those errors.

One important function of uncertainty-aware post-training is selective prediction, in which the model abstains or defers when uncertainty exceeds a threshold. In driving, abstention does not mean doing nothing; it means transitioning to a safer mode, such as reducing speed, increasing following distance, or requesting remote assistance. The design of such transitions requires coordination between the learned model and a rule-based or classical fallback controller. If the fallback is too abrupt, it may create new hazards; if it is too slow, it may fail to prevent an accident. Post-training can help by shaping the model to produce uncertainty signals that are timely and actionable, rather than merely statistically interesting. This requires training objectives that reward early recognition of degrading conditions and penalize sudden, unsafe handovers.

A second function is risk-sensitive policy adjustment. Standard imitation learning tends to mimic average human behavior, but safety-critical driving may require deviation from average behavior when risk is elevated. For example, a model might learn to proceed through an intersection when most human drivers would, yet a safety-oriented system should hesitate when visibility is poor or when another vehicle's behavior is ambiguous. Post-training can incorporate risk penalties, conservative bounds, or adversarial scenario generation to make the policy more cautious in high-uncertainty states. However, excessive conservatism can cause congestion, unpredictable behavior, and new safety risks. The structural trade-off is therefore between safety margins and traffic efficiency, and it cannot be resolved purely by technical optimization. It requires governance decisions about acceptable risk, operational constraints, and public expectations.

5. Safety, Robustness, and Distributional Shift

Distributional shift is a central threat to autonomous driving safety because the world is non-stationary. Weather changes, road construction, seasonal vegetation, pedestrian clothing, vehicle designs, and social norms all evolve. A model trained on historical data may encounter situations that are statistically rare but operationally critical. Uncertainty-aware post-training can improve robustness by identifying when the model is operating outside its competence and by adapting its behavior accordingly. However, robustness is not the same as uncertainty estimation. A model can be robust to noise while still being uncertain about semantics, or it can produce low uncertainty while being wrong. Post-training must therefore

combine uncertainty quantification with stress testing, adversarial evaluation, and scenario-based validation.

Safety also depends on the ability to handle compound failures. A sensor may degrade at the same time that a communication link fails and a pedestrian behaves unexpectedly. In such cases, the system must prioritize among competing risks and degrade gracefully. Post-training can simulate compound failure modes and train policies that remain stable under partial observability. Yet simulation cannot cover every combination, and overfitting to simulated failures may create a false sense of security. The literature on rare-event simulation highlights the need for statistically rigorous scenario generation, but it also warns that simulation fidelity is a limiting factor [12]. A system-level approach must therefore combine simulation, closed-course testing, on-road monitoring, and incident reporting.

Robustness is also tied to interpretability and verification. If a model's uncertainty estimate cannot be explained or audited, regulators and safety engineers may be unable to determine whether the system is safe enough for deployment. This is particularly important because end-to-end models often lack transparent intermediate representations [11]. Post-training can incorporate architectural constraints, such as attention maps, concept bottlenecks, or rule-based guards, that make uncertainty signals more inspectable. However, adding such constraints may reduce flexibility and performance. The trade-off between interpretability and capability is a recurring theme in safety-critical AI, and it cannot be resolved by a single technical fix. It requires a design philosophy that treats transparency as a system requirement rather than an optional feature.

6. Fairness, Governance, and Regulatory Infrastructure

Uncertainty-aware post-training also has fairness implications. If a model is less certain about certain types of pedestrians, vehicles, or environments because they are underrepresented in training data, then its cautious behavior may be distributed unevenly. For example, a system might be more hesitant around cyclists, children, or people in wheelchairs if those categories are rare in the training set. While hesitation may be safer in some cases, it can also lead to discriminatory traffic outcomes, such as longer wait times at crossings or unequal access to mobility. Research on algorithmic bias in autonomous systems has shown that fairness concerns arise not only from explicit labels but also from data collection, feature representation, and deployment context [20]. Post-training must therefore include fairness audits that examine whether uncertainty is correlated with protected attributes or vulnerable road user groups.

Governance is equally important. Ethical governance frameworks for robotics and artificial intelligence emphasize the need for accountability, transparency, and mechanisms for redress [19]. In autonomous driving, these principles translate into requirements for incident reporting, data retention, model versioning, and regulatory oversight. Standards such as the SAE taxonomy for driving automation provide a common vocabulary for levels of automation, but they do not by themselves ensure safety [17]. Regulatory bodies must define how uncertainty-aware claims are validated, how fallback behavior is certified, and how updates to post-trained models are approved. Without such infrastructure, manufacturers may face incentives to deploy models that appear safe in limited tests while obscuring residual uncertainty.

Case investigations illustrate the stakes. The National Transportation Safety Board report on a fatal collision involving a developmental automated driving system showed how gaps in perception, operational design, and safety culture can combine with tragic consequences [18].

Such incidents demonstrate that safety is not achieved by a single algorithm but by an integrated system of design, testing, monitoring, and organizational responsibility. Uncertainty-aware post-training can contribute to this system by making models more aware of their limits, but it cannot replace regulatory oversight, ethical review, or public accountability. Governance must therefore treat uncertainty as a first-class object of regulation, requiring documentation of how uncertainty is estimated, how thresholds are set, and how decisions are audited after deployment.

7. Infrastructure, Deployment, and Sustainability

Deploying uncertainty-aware models at scale requires significant infrastructure. Uncertainty estimation methods such as deep ensembles and Bayesian approximations often increase computation and memory requirements [4]. In a vehicle, these costs must be balanced against latency constraints, energy consumption, and hardware reliability. A model that takes too long to compute uncertainty may be unable to react in time, while a model that is too lightweight may produce poorly calibrated estimates. Post-training must therefore consider the deployment platform, including edge accelerators, sensor fusion pipelines, and over-the-air update mechanisms. These are not merely engineering details; they determine whether uncertainty-aware safety can be delivered consistently across vehicle fleets.

Sustainability is another dimension. Large-scale training and post-training consume energy, and frequent over-the-air updates increase network traffic and fleet management complexity. If uncertainty-aware post-training is used to continuously fine-tune models on operational data, it may create a cycle of perpetual retraining that raises costs and carbon footprints. A more sustainable approach may involve selective updates, modular adaptation, and rigorous change management. The goal should be to improve safety without generating unnecessary technical debt or environmental burden. This requires lifecycle thinking that considers data collection, model training, validation, deployment, monitoring, and eventual decommissioning.

Infrastructure also includes the human and organizational systems that support autonomous vehicles. Remote operators, fleet managers, maintenance technicians, and safety investigators all interact with uncertainty signals. If a model reports high uncertainty, a remote operator must understand what that means and what action to take. If uncertainty is systematically ignored or overridden, the safety benefit is lost. Post-training should therefore be accompanied by training for human operators, clear escalation protocols, and user interfaces that communicate uncertainty without causing alarm or confusion. The socio-technical system must be designed so that uncertainty awareness improves collective decision-making rather than merely shifting responsibility to a human who lacks sufficient context.

8. Evaluation, Assurance, and Operational Feedback

Evaluation of uncertainty-aware post-training cannot rely solely on standard accuracy metrics. It must assess calibration, sharpness, selective risk, and the downstream consequences of uncertainty-triggered actions. A model may have excellent calibration on average but fail in safety-critical subgroups. Evaluation should therefore include stratified testing across weather, geography, traffic density, road user type, and sensor condition. It should also include adversarial and rare-event scenarios that test whether the model recognizes its own limitations. The literature on probabilistic object detection and anomaly detection provides useful methods, but these methods must be integrated into a broader assurance case that connects technical metrics to operational risk [15] [16].

Assurance cases for autonomous driving typically combine evidence from simulation, closed-course testing, on-road trials, and incident analysis. Uncertainty-aware post-training adds a new layer of evidence: documentation of how uncertainty is represented, how thresholds are chosen, and how fallback behavior is triggered. Regulators may require that such evidence be traceable to design decisions and risk assessments. This is challenging because post-training can be iterative and data-dependent, making it difficult to reconstruct why a particular model behaves as it does. Version control, reproducibility, and audit trails are therefore essential. Without them, uncertainty-aware systems may be difficult to certify even if their technical performance is strong.

Operational feedback is also crucial. After deployment, real-world data can reveal new edge cases and calibration failures. However, continuous learning from operational data raises risks of feedback loops, privacy violations, and uncontrolled model drift. Post-training updates should be governed by strict change management, including regression testing, safety validation, and rollback capability. The goal is not to freeze models forever but to create a controlled learning cycle that improves safety without introducing new hazards. This requires collaboration among engineers, safety officers, regulators, and ethicists, and it reinforces the view that uncertainty-aware post-training is as much an institutional process as a technical one.

9. Conclusion

Uncertainty-aware post-training offers a promising path toward safer autonomous driving models, but its benefits depend on how it is embedded in the broader socio-technical system. Estimating uncertainty is necessary because safety-critical decisions require more than average-case accuracy; models must know when they do not know and must translate that knowledge into conservative action. Post-training can calibrate confidence, shape risk-sensitive policies, support selective prediction, and improve robustness under distributional shift. Yet uncertainty estimates can be wrong, fallback mechanisms can fail, and governance can be inadequate. The technical and institutional dimensions of safety are deeply intertwined.

This paper has argued that uncertainty-aware post-training should be treated as an architecture of calibration rather than a final optimization step. It must address perceptual, behavioral, environmental, and infrastructural sources of uncertainty. It must navigate trade-offs among latency, compute cost, interpretability, fairness, and traffic efficiency. It must be supported by standards, regulatory oversight, incident reporting, and public accountability. It must also be sustainable, both computationally and organizationally, and it must include operational feedback that improves safety without creating uncontrolled model drift. Ultimately, safe autonomous driving will not emerge from a single algorithm or a single training stage. It will emerge from an integrated system that combines uncertainty-aware learning, robust engineering, transparent governance, and a sustained commitment to public safety. Post-training is an important part of that system, but it is not a substitute for it.

References

1. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.

2. Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? Advances in Neural Information Processing Systems, 30, 5574–5584.

3. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. Proceedings of the 33rd International Conference on Machine Learning, 48, 1050–1059.

4. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30, 6402–6413.

5. McAllister, R., Gal, Y., Kendall, A., van der Wilk, M., Shah, A., Cipolla, R., & Weller, A. (2017). Concrete problems for autonomous vehicle safety: Advantages of Bayesian deep learning. Proceedings of the 26th International Joint Conference on Artificial Intelligence, 4745–4753.

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

7. Kalra, N., & Paddock, S. M. (2016). Driving to safety: How many miles of driving would it take to demonstrate autonomous vehicle reliability? Transportation Research Part A: Policy and Practice, 94, 182–193.

8. Koopman, P., & Wagner, M. (2017). Autonomous vehicle safety: An interdisciplinary challenge. IEEE Intelligent Transportation Systems Magazine, 9(1), 90–96.

9. Czarnecki, K. (2018). Operational design domain for automated driving systems. Waterloo Intelligent Systems Engineering Lab, University of Waterloo.

10. Codevilla, F., Müller, M., López, A., Koltun, V., & Dosovitskiy, A. (2018). End-to-end driving via conditional imitation learning. IEEE International Conference on Robotics and Automation, 4693–4700.

11. Bojarski, M., Del Testa, D., Dworakowski, D., Firner, B., Flepp, B., Goyal, P., ... & Zieba, K. (2016). End to end learning for self-driving cars. arXiv preprint arXiv:1604.07316.

12. O'Kelly, M., Sinha, A., Namkoong, H., Tedrake, R., & Duchi, J. C. (2018). Scalable end-to-end autonomous vehicle testing via rare-event simulation. Advances in Neural Information Processing Systems, 31, 9827–9838.

13. Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. Proceedings of the 26th International Joint Conference on Artificial Intelligence, 4691–4697.

14. Michelmore, R., Kwiatkowska, M., & Gal, Y. (2018). Evaluating uncertainty quantification in end-to-end autonomous driving control. arXiv preprint arXiv:1811.06817.

15. Feng, D., Harakeh, A., Waslander, S. L., & Dietmayer, K. (2021). A review and comparative study on probabilistic object detection in autonomous driving. IEEE Transactions on Intelligent Transportation Systems, 23(8), 9961–9980.

16. Bogdoll, D., Nitsche, M., & Zöllner, J. M. (2022). Anomaly detection in autonomous driving: A survey. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, 4487–4498.

17. SAE International. (2021). Taxonomy and definitions for terms related to driving automation systems for on-road motor vehicles (J3016_202104). SAE International.

18. National Transportation Safety Board. (2020). Collision between vehicle controlled by developmental automated driving system and pedestrian. National Transportation Safety Board.

19. Winfield, A. F., & Jirotka, M. (2018). Ethical governance is essential to building trust in robotics and AI systems. Philosophical Transactions of the Royal Society A, 376(2133), 20180085.