

Sim-to-Real Transfer Through Adaptive Post-Training for Autonomous Driving in Unseen Environments

Else Barth

Department of Computer Science, University of North Texas, Denton, TX, USA.
elsebarth80@unt.edu

Abstract

Autonomous driving systems increasingly depend on simulation for scalable training, yet the transfer of learned policies to unseen real-world environments remains a central obstacle. This paper examines sim-to-real transfer through the lens of adaptive post-training, understood not merely as a final fine-tuning step but as a socio-technical capability embedded in the lifecycle of autonomous vehicle deployment. The argument is that unseen environments expose structural limits in simulation-based development because they combine perceptual novelty, behavioral ambiguity, infrastructure irregularity, and governance uncertainty. Adaptive post-training can narrow these gaps by recalibrating representations, control policies, and safety constraints after initial training, using limited field data, targeted simulations, and operational feedback. However, such adaptation also introduces new system-level trade-offs involving architectural complexity, verification burden, data governance, fairness, sustainability, and accountability. The paper develops a conceptual framework for analyzing these trade-offs and connects them to practical concerns in autonomous driving, including operational design domains, edge-case exposure, model auditing, and infrastructure dependence. It also compares autonomous driving with robotics, aviation, and large-scale machine learning systems to clarify how post-training changes reliability engineering and policy expectations. The conclusion is that adaptive post-training is necessary but insufficient by itself; it must be paired with transparent validation, continuous monitoring, and institutional mechanisms that distribute responsibility across manufacturers, operators, regulators, and affected communities. The paper contributes a system-level research agenda for making sim-to-real transfer robust, equitable, and sustainable in unseen environments.

Keywords

sim-to-real transfer; adaptive post-training; autonomous driving; domain shift; robustness; governance; socio-technical systems; sustainability.

1. Introduction

Autonomous driving has become a test case for the broader problem of transferring machine learning systems from controlled simulation into uncertain physical environments. Simulation offers repeatability, low marginal cost, dangerous-scenario generation, and large-scale data synthesis, but it cannot fully reproduce the material, social, and infrastructural complexity of real roads. Early work on domain randomization illustrated how visual and dynamic variability in simulation could improve transfer to real robots and vehicles [1]. Related efforts demonstrated that synthetic data alone could support real-world control tasks when the training distribution was designed carefully [2]. Research on dynamics randomization further showed that variation in physical parameters can produce policies that tolerate real-world

mismatch [3]. Simulators such as CARLA provided open urban environments for training and evaluation, while conditional imitation learning demonstrated how high-level commands and sensory inputs could be combined for end-to-end driving [4], [5]. These advances created a foundation for simulation-heavy development, but they also revealed that transfer failure often appears precisely in unseen environments.

The phrase unseen environments refers not only to unfamiliar streets or weather but also to unfamiliar social practices, maintenance conditions, traffic norms, sensor degradation patterns, and regulatory expectations. A vehicle trained in one city may encounter different lane markings, pedestrian behaviors, construction protocols, and enforcement practices in another. A model trained under clear optical conditions may confront rain, dust, glare, or occlusion. An end-to-end policy that performs well in simulation may fail when perception is corrupted by reflections, when other drivers behave irrationally, or when map data is stale. Recent work on post-training in end-to-end autonomous driving suggests that adaptation after initial pre-training can improve deployment readiness by adjusting models to new tasks, domains, or operational constraints [6]. This paper treats adaptive post-training as a critical bridge between simulation and reality, while arguing that its value depends on how it is integrated into architecture, governance, and infrastructure.

The central claim is that sim-to-real transfer through adaptive post-training should be understood as a system-level process rather than a model-centric intervention. Adaptive post-training can involve continued learning from limited real data, reinforcement from human feedback, safety-constrained recalibration, or targeted simulation that addresses observed failures. Each approach changes the distribution of responsibility among engineers, operators, regulators, and users. It also changes the economics of deployment because adaptation requires data pipelines, validation environments, monitoring systems, and update mechanisms. The paper therefore asks how adaptive post-training can be designed to improve robustness without creating hidden technical debt, unfair performance disparities, or unsustainable operational burdens. It examines architectural trade-offs, safety and governance implications, infrastructure dependencies, fairness concerns, and cross-domain lessons. The goal is not to propose a single algorithm but to clarify the structural conditions under which sim-to-real transfer can become a dependable and accountable capability.

2. Conceptual Foundations of Sim-to-Real Transfer

Sim-to-real transfer rests on the assumption that simulation can approximate the causal structure of the real world closely enough for learned policies to generalize. Domain randomization pursues this goal by varying textures, lighting, object placement, and dynamics so that the real world appears as one sample from a broad training distribution [1]. CAD2RL extended this logic by training from synthetic images without real data, showing that carefully designed visual variation can support real control [2]. Dynamics randomization addressed the physical side of the gap by exposing policies to varied masses, friction, delays, and actuation characteristics [3]. These methods are powerful because they reduce overfitting to a single simulated reality. Yet they do not eliminate the need for real-world validation because randomization can omit rare but consequential factors, such as adversarial behavior, sensor fouling, or legal constraints.

End-to-end driving architectures intensify the transfer problem because they map raw sensory input to control outputs, often without explicit intermediate representations that engineers can inspect. Early end-to-end systems demonstrated the feasibility of learning steering from camera images [7]. Later work used conditional imitation learning to incorporate navigational

commands and improve behavioral structure [5]. Learning by cheating showed that privileged information available in simulation can supervise a student model that relies only on deployable sensors [8]. Data-driven simulation and robust control policies further demonstrated that simulation can support end-to-end autonomy when paired with careful representation learning [9]. Learning from simulation without real-world labels and learning to drive in a day illustrated both the promise and fragility of rapid simulation-to-road pipelines [10], [11]. These works collectively show that transfer is not a single event; it is a continuing negotiation between distributional assumptions, sensor fidelity, and operational context.

The conceptual gap is therefore both epistemic and institutional. Epistemically, simulation cannot prove that a policy will behave safely in unseen environments because the space of possible environments is open-ended. Institutionally, no simulator can certify a vehicle for all roads, weather conditions, and social contexts. This is why formal models of safe and scalable self-driving cars emphasize structured assumptions about perception, planning, and responsibility [14]. It is also why reliability estimates based on miles driven must be interpreted cautiously, because rare events dominate safety outcomes and require enormous exposure to estimate statistically [15]. Adaptive post-training responds to this situation by allowing models to change after deployment or during staged deployment. But adaptation creates a paradox: the system must learn from the very environments that were not represented in training, which means the learning process itself becomes a safety-critical operation.

3. Adaptive Post-Training as a Bridge

Adaptive post-training can be defined as any deliberate process that modifies a pre-trained driving model after initial large-scale training to improve performance in a target domain, task, or operational context. This includes supervised fine-tuning on field data, reinforcement learning with safety constraints, imitation from human interventions, preference-based optimization, and simulator-in-the-loop recalibration. The required reference on post-training in end-to-end autonomous driving highlights that post-training is not an afterthought but a distinct phase with its own data requirements, evaluation challenges, and failure modes [6]. In unseen environments, post-training is especially important because the mismatch between training and deployment may be localized. A model may perform well on highways but poorly at unsignalized intersections, or well in daylight but poorly during heavy rain. Adaptive post-training can target these localized weaknesses without retraining the entire system from scratch.

The advantage of post-training is its efficiency. Full retraining is expensive, risks catastrophic forgetting, and may require data that are difficult to collect. Adaptive methods can use smaller datasets, focus on failure cases, and preserve previously learned capabilities. They can also incorporate human feedback from safety drivers, remote operators, or fleet managers. In this sense, post-training resembles continuous improvement in software operations, where patches address discovered defects. However, the analogy is incomplete because driving models are statistical and their behavior is distributed across many parameters. A targeted update can improve one scenario while degrading another. This creates a verification problem: engineers must determine not only whether the post-trained model improves on the target distribution but also whether it preserves safety on the original distribution.

Another advantage is that adaptive post-training can align models with operational design domains. A vehicle deployed in a geofenced retirement community has different priorities

from one deployed in a dense commercial district. Post-training can adjust speed profiles, pedestrian interaction, and risk tolerance to local norms. Yet this localization raises governance questions. If every fleet operator adapts models independently, then performance and safety may vary across jurisdictions. Regulators may struggle to audit systems that are continuously updated. Manufacturers may face liability questions when a post-trained model behaves differently from the certified baseline. Adaptive post-training therefore cannot be treated as a purely technical optimization; it must be embedded in a lifecycle of documentation, approval, monitoring, and rollback.

4. Architecture and System-Level Trade-Offs

The architecture of an autonomous driving system determines how easily adaptive post-training can be performed. Modular architectures separate perception, prediction, planning, and control, allowing engineers to identify which component fails in an unseen environment. End-to-end architectures may offer smoother optimization and fewer interface losses, but they are harder to diagnose and update safely. A hybrid architecture can combine learned components with explicit rules, maps, or safety monitors. This hybrid approach reflects the insight that machine learning systems accumulate hidden technical debt when data dependencies, configuration, and feedback loops are not managed explicitly [19]. In autonomous driving, technical debt can appear as undocumented scenario libraries, fragile sensor calibration, or opaque post-training datasets that are not representative of deployment conditions.

Adaptive post-training adds new architectural layers, including data selection, scenario mining, simulation replay, model updating, validation, deployment, and rollback. Each layer introduces latency and governance overhead. If post-training occurs in the cloud, the system must manage data upload, privacy, bandwidth, and version control. If it occurs on the vehicle, the system must operate under compute, memory, and thermal constraints. Edge adaptation may improve responsiveness and reduce data transmission, but it can also create a fragmented fleet in which different vehicles run different models. Fleet-level consistency is important for safety certification, maintenance, and incident investigation. These trade-offs resemble those in large-scale machine learning operations, where the cost of maintaining a model in production often exceeds the cost of initial training [19].

Safety architecture must also account for the possibility that post-training introduces unexpected behavior. A safety monitor can enforce hard constraints, such as collision avoidance or speed limits, independent of the learned policy. Formal models of safe self-driving provide conceptual foundations for such monitors, though they depend on assumptions about perception and other agents [14]. AI safety research similarly warns about specification gaming, negative side effects, and reward hacking [20]. In an adaptive driving context, a post-training objective that rewards passenger comfort might unintentionally produce timid behavior that increases risk at merges. An objective that rewards assertiveness might increase collisions. The architecture must therefore balance optimization targets with invariant constraints and human oversight.

5. Robustness, Safety, and Governance in Unseen Environments

Unseen environments are a governance problem as much as a technical one. Regulators must decide what evidence is sufficient to permit deployment when the operational domain cannot be fully enumerated. Accident reports involving autonomous vehicles show that failures often arise from complex interactions among technology, human behavior, and road design [21]. A

model that is robust in simulation may still fail when a construction worker waves a vehicle forward, when a police officer gives conflicting signals, or when a sensor is blocked by mud. Adaptive post-training can help by incorporating these cases after they are observed, but it also means that safety improves through experience. The ethical question is whether the public should bear the risk of generating that experience.

Governance mechanisms must therefore distinguish between adaptation that reduces known risk and adaptation that expands operational scope. A post-training update that improves rain performance within a geofenced area is different from one that enables nighttime highway driving for the first time. The former can be managed through routine validation, while the latter requires new certification. This distinction suggests a tiered approval system based on operational design domain, model change magnitude, and failure impact. Model cards and similar documentation practices can support such governance by recording intended use, performance characteristics, limitations, and evaluation conditions [18]. Internal algorithmic auditing frameworks can further define who reviews updates, how evidence is collected, and how affected communities are represented [17].

Accountability is complicated by the distributed nature of adaptive post-training. A manufacturer may provide the base model, a fleet operator may fine-tune it, a cloud provider may host the update pipeline, and a city may supply infrastructure data. If an incident occurs, responsibility may be diffused across these actors. Fairness and abstraction in sociotechnical systems warn that technical fixes cannot resolve problems that are fundamentally institutional [16]. In autonomous driving, fairness concerns include whether post-training improves performance equally across neighborhoods, whether data collection burdens some communities more than others, and whether accessibility features are preserved after adaptation. A model that performs well in wealthy, well-mapped areas and poorly in informal settlements reflects structural inequality, not merely a data deficiency.

6. Infrastructure, Sustainability, and Operational Economics

Adaptive post-training depends on infrastructure that is often invisible in model-centric discussions. Simulation clusters, data storage, labeling services, over-the-air update networks, and roadside sensors all have costs. Simulators such as CARLA require high-fidelity assets and compute resources [4]. Data-driven simulation and robust control pipelines require scenario generation and validation infrastructure [9]. Diverse scenario platforms such as MetaDrive illustrate the value of composable environments for generalization, but they also require maintenance and curation [12]. If every region requires its own post-training pipeline, the environmental and economic sustainability of autonomous driving may decline. Large-scale training and adaptation consume energy, and continuous fleet learning can multiply those costs across vehicles.

Sustainability also has a social dimension. If adaptive post-training relies on continuous data upload from customer vehicles, it may shift costs onto users through bandwidth, privacy loss, and vehicle wear. If it relies on dedicated test fleets, it concentrates benefits in regions where operators choose to invest. Infrastructure equity matters because post-training for unseen environments may require local mapping, connectivity, and maintenance. Regions with poor roads, inconsistent signage, or limited digital infrastructure may become harder to serve, even though they may benefit most from improved mobility. A sustainable approach would prioritize reusable representations, selective data collection, and adaptation methods that do not require exhaustive local retraining. It would also align with broader policy goals for transportation access, energy use, and urban planning.

The economics of post-training are further shaped by uncertainty. A fleet operator must decide how much data to collect, how often to update, and how to validate changes. Too little post-training leaves performance gaps; too much creates instability and cost. This is a classic trade-off in maintenance and reliability engineering. Aviation addresses similar problems through strict configuration control, redundancy, and incident reporting. Autonomous driving can learn from aviation, but it operates in a more open environment with less centralized control. The challenge is to create update mechanisms that are fast enough for local adaptation but slow enough for meaningful oversight. This balance cannot be resolved by technology alone; it requires standards, insurance models, and regulatory capacity.

7. Fairness, Policy, and Socio-Technical Accountability

Fairness in autonomous driving is often reduced to bias in perception, such as unequal detection of pedestrians with different skin tones or clothing. While important, fairness also includes access, service quality, and risk distribution. Adaptive post-training can either mitigate or amplify disparities depending on data selection and deployment incentives. If post-training data come disproportionately from affluent areas with well-maintained roads, models may perform worse in underrepresented areas. If post-training focuses only on efficiency metrics, it may neglect the needs of pedestrians, cyclists, and people with disabilities. Sociotechnical frameworks remind us that fairness is not a property of a model alone but of the system in which it operates [16]. Policy must therefore require participatory evaluation and disaggregated performance reporting.

Accountability frameworks such as internal algorithmic auditing can help by assigning responsibility for data provenance, model changes, and incident review [17]. Model cards can document the intended use and limitations of each post-trained model version [18]. These instruments are necessary but not sufficient. They must be connected to enforcement, liability, and compensation. If a post-trained model causes harm in an unseen environment, affected parties need a clear path to remedy. This may require new insurance products, mandatory incident reporting, and independent review boards. It also requires transparency about when adaptation occurs, because a vehicle that updates overnight may behave differently the next day. Users, regulators, and the public should not discover such changes only after an accident.

Policy should also address the boundary between simulation and reality. Simulators are not neutral; they encode assumptions about roads, agents, and norms. If simulation becomes the primary basis for certification, then the quality and representativeness of simulation become matters of public interest. Governments may need to audit simulation scenarios, require coverage of vulnerable road users, and mandate testing in diverse real-world conditions. Adaptive post-training can complement these efforts by providing a structured way to learn from deployment, but it must not become a substitute for pre-deployment validation. The goal is a layered safety case in which simulation, closed-course testing, supervised deployment, and continuous monitoring each play a defined role.

8. Case Illustrations and Cross-Domain Comparisons

Several case patterns illustrate the value and limits of adaptive post-training. In robotics, synthetic training without real data showed that visual diversity can support real control, but it also showed that performance depends on task-specific assumptions [2]. Dynamics randomization demonstrated that policies can transfer across physical parameter shifts, yet real-world wear and tear can change those parameters over time [3]. In autonomous driving, conditional imitation learning provided a way to incorporate commands, but it also revealed

that high-level intent does not guarantee safe low-level control in unfamiliar traffic [5]. Learning from simulation without real-world labels and learning to drive in a day demonstrated rapid deployment potential, while also highlighting the need for safety drivers and constrained operational domains [10], [11].

Cross-domain comparison with aviation suggests that adaptation should be accompanied by configuration control and mandatory reporting. Aviation does not permit continuous, opaque model updates in safety-critical systems without certification. Medical devices offer another comparison, where post-market surveillance is required but changes must be validated. These domains show that adaptive post-training is not inherently unsafe; the risk lies in unmanaged adaptation. Autonomous driving can adopt a similar posture by treating post-training as a controlled change process. This means versioning data, models, and scenarios; defining rollback criteria; and monitoring for distributional drift. It also means recognizing that some updates may require re-certification because they alter the operational design domain.

Another comparison is with large-scale recommendation systems, where continuous adaptation is routine but failures are often reversible and low-stakes. Autonomous driving cannot accept the same tolerance for error because failures can be irreversible. This difference justifies a more conservative approach to adaptation. Post-training should be allowed to improve safety and fairness, but it should not be used to silently expand capabilities. The required work on post-training in end-to-end driving is important because it directs attention to this phase, yet its implications must be interpreted within a safety-critical governance framework [6]. The technology can enable adaptation, but the institution must decide what adaptation is permissible, how it is verified, and who is accountable.

9. Future Directions and Research Agenda

Future research should develop methods for selective adaptation that preserve safety invariants. This includes uncertainty-aware data selection, where the system identifies scenarios in which it is likely to fail and requests targeted supervision. It also includes simulation-in-the-loop post-training, where real-world failures are reproduced in simulation and used to generate diverse training variants. Research on diverse driving scenarios provides a foundation for such work by enabling composable environments that test generalization [12]. Reinforcement learning with implicit affordances offers another direction for learning robust urban driving policies [13]. However, these methods must be evaluated not only by task success but also by fairness, interpretability, and governance readiness.

A second direction is the development of infrastructure for accountable post-training. This includes secure data pipelines, privacy-preserving federated learning, and standardized model documentation. Model cards and auditing frameworks provide starting points, but they must be adapted to the continuous update cycles of autonomous fleets [17], [18]. Safety research should also address specification gaming and negative side effects in adaptive policies [20]. If a post-training objective rewards smooth driving, it may learn to avoid difficult situations rather than handle them, which could disadvantage certain passengers or neighborhoods. Evaluation must therefore include distributional analysis across locations, weather, time of day, and road-user types.

A third direction is policy experimentation. Regulators could create sandboxes for adaptive post-training in limited operational domains, with strict reporting and independent evaluation. They could require disaggregated safety and performance data, similar to environmental monitoring. They could also fund public-interest simulation assets so that safety testing is not

controlled solely by private firms. International coordination will be necessary because vehicles and data cross borders. The broader lesson from sociotechnical fairness research is that accountability must be designed into the system from the beginning [16]. Adaptive post-training should not be a hidden layer of engineering; it should be a documented, governed, and contestable process.

10. Conclusion

Sim-to-real transfer through adaptive post-training is a promising but demanding pathway for autonomous driving in unseen environments. Simulation provides scale, safety, and controllability, but it cannot fully represent the open-ended complexity of real roads. Adaptive post-training can close specific gaps by recalibrating models with field data, human feedback, and targeted simulation. It can improve robustness, localize behavior, and support continuous improvement. Yet it also introduces new risks: catastrophic forgetting, fragmented fleets, opaque updates, unfair performance disparities, and diffused accountability. These risks are not incidental; they are structural consequences of embedding learning systems in physical and social infrastructures.

The paper has argued that adaptive post-training must be treated as a socio-technical capability rather than a model-centric technique. Its success depends on architecture that supports diagnosis and rollback, governance that distinguishes improvement from capability expansion, infrastructure that is sustainable and equitable, and policy that ensures accountability across manufacturers, operators, regulators, and communities. Cross-domain comparisons with robotics, aviation, and medical devices suggest that controlled change is possible, but only when adaptation is documented, validated, and bounded. The required work on post-training in end-to-end autonomous driving points toward a future in which adaptation is central to deployment [6]. The task for researchers and policymakers is to ensure that this future is robust, fair, and democratically accountable.

References

1. Tobin, J., Fong, R., Ray, A., Schneider, J., Zaremba, W., & Abbeel, P. (2017). Domain randomization for transferring deep neural networks from simulation to the real world. In 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 23–30). IEEE. <https://doi.org/10.1109/IROS.2017.8202133>
2. Sadeghi, F., & Levine, S. (2016). CAD2RL: Real single-image flight without a single real image. arXiv preprint arXiv:1611.04201.
3. Peng, X. B., Andrychowicz, M., Zaremba, W., & Abbeel, P. (2018). Sim-to-real transfer of robotic control with dynamics randomization. In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 3803–3810). IEEE. <https://doi.org/10.1109/ICRA.2018.8460528>
4. Dosovitskiy, A., Ros, G., Codevilla, F., Lopez, A., & Koltun, V. (2017). CARLA: An open urban driving simulator. In Proceedings of the 1st Annual Conference on Robot Learning (CoRL) (pp. 1–16).
5. 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. <https://doi.org/10.1109/ICRA.2018.8460487>

6. Favarò, F., Nader, N., Eurich, S., Tripp, M., & Varadaraju, N. (2017). Examining accident reports involving autonomous vehicles in California. *PLoS ONE*, 12(9), e0184952. <https://doi.org/10.1371/journal.pone.0184952>
7. 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*.
8. Chen, D., Zhou, B., Koltun, V., & Krähenbühl, P. (2020). Learning by cheating. In *Proceedings of the 4th Conference on Robot Learning (CoRL)* (pp. 66–75).
9. Amini, A., Paull, L., Balakrishnan, H., & 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. <https://doi.org/10.1109/LRA.2020.2966414>
10. Bewley, A., Rigley, J., Liu, Y., Hawke, J., Shen, R., Lam, R., & Kendall, A. (2019). Learning to drive from simulation without real world labels. In *2019 International Conference on Robotics and Automation (ICRA)* (pp. 4818–4824). IEEE. <https://doi.org/10.1109/ICRA.2019.8793668>
11. Kendall, A., Hawke, J., Janz, D., Mazur, P., Reda, M., Allen, J., Lam, V.-D., Qureshi, A., & Shah, A. (2019). Learning to drive in a day. In *2019 International Conference on Robotics and Automation (ICRA)* (pp. 8248–8254). IEEE. <https://doi.org/10.1109/ICRA.2019.8793742>
12. Xu, Q., Zhang, S., & Shen, S. (2022). MetaDrive: Composing diverse driving scenarios for generalizable reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3), 3461–3475. <https://doi.org/10.1109/TPAMI.2022.3197916>
13. Toromanoff, M., Wirbel, E., & Moutarde, F. (2020). End-to-end model-free reinforcement learning for urban driving using implicit affordances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 7153–7162).
14. Shalev-Shwartz, S., Shammah, S., & Shashua, A. (2017). On a formal model of safe and scalable self-driving cars. *arXiv preprint arXiv:1708.06374*.
15. 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. <https://doi.org/10.1016/j.tra.2016.09.010>
16. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)* (pp. 59–68). ACM. <https://doi.org/10.1145/3287560.3287598>
17. Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D., & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT)* (pp. 33–44). ACM. <https://doi.org/10.1145/3351095.3372873>
18. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)* (pp. 220–229). ACM. <https://doi.org/10.1145/3287560.3287596>

19. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems (NeurIPS)* (Vol. 28, pp. 2503–2511).
20. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.