Towards Auditing AI Systems in the Wild

Aditya T. Vadlamani
vadlamani.12@osu.edu
Ohio State University
Columbus, Ohio, USA

Anutam Srinivasan
asrinivasan350@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia, USA

Srinivasan Parthasarathy
srini@cse.ohio-state.edu
Ohio State University
Columbus, Ohio, USA

Abstract

AI systems are increasingly deployed in real-world settings where their behavior is shaped by dynamic environments, evolving data distributions, and complex interactions with users and infrastructure. Traditional machine learning evaluation focuses on benchmarks and operates within sandboxed environments, providing only a limited view of the true system behavior *in the wild*. We argue for the development of principled auditing frameworks that monitor deployed AI systems throughout their lifecycle. We further propose framing auditing as a statistical problem of monitoring constraint violations under uncertainty, where desired properties (e.g., fairness and safety) are treated as risk-controlled constraints that must be continuously evaluated as systems evolve through iterative feedback. This perspective highlights the need for uncertainty-aware monitoring methods, socio-technical specifications of audit criteria, and auditing infrastructures that enable ongoing oversight of AI systems in the wild.

CCS Concepts

• **Social and professional topics** → **Technology audits**; • **Computing methodologies** → **Machine learning**; *Uncertainty quantification.*

Keywords

AI Auditing, Risk Control, Deployed ML Systems, Fairness, Safety

ACM Reference Format:
Aditya T. Vadlamani, Anutam Srinivasan, and Srinivasan Parthasarathy. 2026. Towards Auditing AI Systems in the Wild. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD '26), August 09–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3770855.3818648

1 Auditing AI: A Clear and Present Need

Recent advances in machine learning have led to rapid development and deployment of AI systems across many real-world domains. Today, AI and ML systems shape information access [33] and decision-making in areas ranging from recommendation systems [1] and financial services [21, 66] to AI assistants in healthcare [12], autonomous agents [41], and embodied agents [24]. As these systems are increasingly embedded in social and institutional processes, ensuring their trustworthiness has become a core challenge.

Despite the growing societal impact, the dominant evaluation paradigm in machine learning is mainly *offline* and *pre-deployment*. In practice, many of the most consequential behaviors of AI systems only emerge after deployment. These challenges are amplified by the rapid pace of innovation in AI [11]. Organizations face strong incentives to deploy increasingly capable models quickly, with the ever-evolving landscape and competition between frontier AI players (e.g., DeepMind [61], Anthropic [6], & OpenAI [55]). This dynamic creates a "race" in which organizations prioritize immediate deployment over careful evaluation and long-term risk assessment. Organizations that largely ignore or pay lip service to safety precautions may win the race in the short term, but at a cost to society in the long term.

Researchers have highlighted the vulnerabilities in deployed AI systems, including risks related to safety, legality, and security, as well as their potential for discriminatory effects [29, 37]. As AI systems continue to expand in capability and influence, public trust in such systems is at a tipping point. Mechanisms that enable independent verification of claims about system performance, fairness, and reliability are increasingly important [35]. Effective auditing of AI systems offers a promising approach to addressing this challenge by assessing an organization's claims about algorithmic efficacy, safety, and security, while helping organizations detect failures, mitigate fraud, and reduce discrimination [25]. Building such trust is the essential fabric of modern society [8].

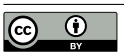
Effective and efficient auditing mechanisms enable the careful deployment of AI systems and their democratic governance [23]. In short, auditing is essential for achieving trustworthy AI, but the challenges are daunting. First, the proprietary nature of many AI systems makes auditing them challenging. Second, auditing requires a quantifiable specification (from a legal, governance, safety, and security standpoint, grounded in societal norms and specific silos (e.g., health, finance)). Third, the socio-technical ability to assess whether this specification continues to hold once such systems are deployed at scale in the wild is non-trivial. Fourth, audits may be public (black-box audits) or private (grey- and white-box audits conducted by trusted third parties). Relatedly, audits are required not only for the final product but also for the development process. Finally, to ensure efficient deployment, audits should be seamless, fast, secure, and accurate–enabling organizations to serve their clients and/or the public effectively.

2 Dimensions of AI Auditing

Modern AI deployments involve complex interactions among models, data pipelines, infrastructure, and human users, and as a result, auditing cannot be treated as a single evaluation task; it must instead consider several complementary dimensions.

2.1 Lifecycle

2.1.1 Pre-Deployment AI Evaluations. The dominant paradigm for evaluating AI systems has primarily focused on **pre-deployment**



KDD '26, Jeju Island, Republic of Korea

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3818648

evaluations, where models are trained on historical data and evaluated on benchmark datasets or held-out test sets prior to deployment. This paradigm enabled rapid progress in algorithmic development and comparisons of model performance, further motivating innovations by organizations and research labs. However, as AI systems are increasingly deployed in dynamic real-world environments, benchmarking often provides only a narrow view of system behavior [46]. Importantly, such benchmarks can be manipulated and engineered (see, for example, Pendragon's case against Sun Microsystems on the CaffeineMark Benchmark in the 1990s [65] or the more recent Volkswagen "dieselgate" scandal). In practice, the reliability, safety, and societal impact of AI systems are determined not only by their performance on static benchmarks but also by their interactions with evolving environments, infrastructure, and human users. Several features of modern AI systems limit the effectiveness of a purely pre-deployment evaluation.

Limited Visibility into Complex AI Systems. Modern AI systems rarely consist of a single model operating in isolation, but instead include several models embedded within complex pipelines that include data collection, model training, deployment infrastructure, and downstream decision-making processes [54]. These systems, in turn, interact with external components (e.g., data sources, other automated systems), raising questions about the *provenance* of AI systems. As a result, evaluating a model on a benchmark dataset provides a limited view of the broader system behavior. Furthermore, AI system internals and decisions made during development are not well understood (or even shared), resulting in sources of uncertainty that are ignored by evaluation or auditing. For example, a key component of model development that is commonly overlooked concerns how data is collected (i.e., data provenance) for training, as well as whether data from diverse contexts (e.g., different socio-cultural groups) is well-represented. Finally, system failures may encompass issues in data collection or system integration, highlighting the incompleteness of existing model evaluation mechanisms for assessing the *entire* AI system pipeline.

Real-World Dynamics. AI systems that are deployed in real-world settings must operate in inherently dynamic environments–data distributions may shift over time [2], user behavior may evolve in response to system outputs [13], and new or unintended use cases may emerge after deployment (see, for example, Microsoft's Tay chatbot from 2016). These dynamics can produce behaviors that are difficult to anticipate during development. For example, model performance can degrade due to distribution shift [34] or amplify biases through human-AI feedback loops [27]. This is exacerbated for embodied AI, where distribution shifts can lead to unreliable performance and dangerous behaviors (i.e., obstacle collisions) [57]. Pre-deployment testing can help identify certain classes of failures, but it cannot fully anticipate how systems will behave in complex social and technical environments.

Incentives for Rapid Deployment. The aforementioned challenges are compounded due to the incentive to rapidly develop and deploy AI systems. Frontier AI companies operate in a competitive environment in which rapid innovation and deployment are required to remain relevant and deliver significant advantages. As a result, organizations compete by prioritizing incremental improvements and rapid release cycles to remain competitive. The emphasis shifts to achieving state-of-the-art performance on widely used *existing* benchmarks and product metrics, rather than developing robust mechanisms for long-term understanding of behavior and risks. While this perspective has led to remarkable AI capabilities, it disincentivizes a focus on comprehensive evaluation and oversight. The gap between technological advancement and mechanisms for monitoring, governing, and auditing continues to widen as we advance in the former while neglecting the latter.

The Need for Auditing Across the AI Lifecycle. Altogether, these factors highlight the limitations of evaluation frameworks that solely focus on model performance prior to deployment. Ensuring AI systems remain reliable and trustworthy requires mechanisms that extend beyond traditional benchmark-based evaluations and encompass the *entire lifecycle* of AI systems. In particular, there is a need for approaches that enable **ongoing auditing of AI systems during development and deployment**, allowing researchers, organizations, and regulators to monitor system behavior, detect emerging risks, and intervene when necessary. Such approaches can help bridge the gap between static (model) evaluation and the dynamic environments in which the AI systems actually operate. This leads us to critically understand what **post-deployment evaluations (or auditing)** entails.

2.1.2 Post-Deployment AI Auditing. Auditing comprehensive AI systems after deployment poses additional challenges and questions. Incorrect usage of AI models can lead to catastrophic performance [44], thereby flagging a model as unsafe post-deployment. However, the scope of the audit may be focused on compliance when the model is used *within reason*. Thereby, introducing ambiguity, i.e., "what is *within reason*?", into the auditing problem and requiring careful treatment in scoping [32]. Furthermore, large-scale audits will require automated data filtration of data collected during deployment to align with the scope of the auditing problem, thereby increasing the scope for errors in the auditing pipeline.

A well-known example of the complexity of post-deployment auditing arose in the case of the COMPAS recidivism risk assessment tool used in U.S. criminal justice systems. An investigation by ProPublica argued that the system exhibited racial disparities in false positive rates when predicting recidivism risk, suggesting discriminatory outcomes against Black defendants [5]. In response, the system's developer contested these findings, arguing that the model satisfied an alternative fairness notion based on calibration across groups [20]. The resulting debate highlighted that post-deployment auditing is not merely a technical task but also depends critically on the specification of auditing criteria and the interpretation of statistical evidence. Without clear specifications for fairness or risk thresholds, different auditing analyses may reach conflicting conclusions despite relying on the same underlying data.

Post-deployment auditing further encounters data heterogeneity, with the ubiquitous use of the same AI models, e.g., how educators use AI to create lesson plans, vastly differing from how software developers use them. Audits that focus on global performance across all tasks (e.g., toxicity in LLMs) will require careful handling to accurately reflect the model's behavior, even when certain tasks are prone to eliciting non-compliant behavior.

Unlike pre-deployment auditing, correcting compliance issues in deployed models also poses several challenges: corrections must be made on the fly under expedited timelines while addressing the audit results. These constraints preclude full-model retraining and require us to consider approaches that seamlessly transition model audits to model updates.

2.2 Access Levels and Institutional Roles

Another important dimension of AI auditing concerns both the level of access that auditors have to the system and the institutional actors responsible for conducting the audit. In practice, these two aspects are closely related: different stakeholders typically operate under different levels of system visibility. For example, external researchers often rely on black-box interaction with deployed systems, whereas internal teams may have full (white-box) access to model internals and training pipelines.

As more and more AI companies compete to develop and deploy their own models, to maintain a competitive advantage, companies are becoming less transparent and more restrictive about their data, model architectures, and other system components, making black-box auditing techniques more desirable [35]. That said, certain companies do (partially) open-source these details (e.g., Llama 4, Qwen, DeepSeek), resulting in wider adoption since end-users can independently evaluate model performance.

Table 1 summarizes common auditing actors and the levels of system access typically available to them. These categories are not mutually exclusive, but they illustrate the range of possible auditing arrangements in practice and highlight the importance of a diverse auditing ecosystem. No single actor has complete visibility into the behavior and risks of complex AI systems. Internal teams possess the deepest technical access, while external researchers and users often observe real-world behaviors that may not emerge in controlled testing environments.

2.3 Targets for AI Auditing

Auditing AI systems requires examining multiple aspects of system behavior and impact. A recent report from NIST highlights six categories for post-deployment monitoring [50], which are useful for categorizing AI auditing while allowing interactions between categories (e.g., human-AI interactions can cause safety failures).

Functional Audits. Functional auditing can be seen as an extension of benchmark evaluations, but focused on whether an AI system continues to perform reliably as intended after deployment. This includes evaluating model performance under real-world conditions, identifying degradation due to distribution shifts, and detecting unanticipated inputs or use cases.

Operational Audits. Operational auditing focuses on the broader system, including the infrastructure that underlies the AI deployment (e.g., data collection pipelines, logging mechanisms, and service reliability (uptime)). Failures in AI systems aren't necessarily due to the model; they can also come from other parts of the system or from the integration of different subsystems. Operational audits can be seen as evaluating system-level consistency and reliability.

Human-AI Interaction Audits. In many cases, AI systems operate as part of human-AI teams, where humans interpret and act on model outputs. Auditing these interactions requires evaluating how users understand, trust, and respond to model recommendations–inspecting the feedback loops and the potential for automation bias or misuse. This is particularly challenging as there are no good benchmarks, and it is unclear where to start for evaluation.

Safety and Security Audits. Safety and security auditing focuses on whether AI systems are resilient to misuse, adversarial inputs, and malicious attacks. This includes evaluating robustness to adversarial manipulation and assessing guardrails against deceptive or unsafe system behavior. These audits are particularly important for systems deployed in high-impact or adversarial environments. Recent work has also explored safety protocols designed to prevent models from intentionally subverting safeguards. For example, Greenblatt et al. [28] proposes "control evaluations," a methodology for evaluating safety protocols by simulating adversarial models that attempt to bypass monitoring and auditing mechanisms. However, in such settings, when we use a separate AI model as the auditor, it is natural to ask *who audits the auditor?* For this reason, we benefit from considering auditing as a statistical framework, which further motivates our proposed perspective of treating auditing as a risk-control framework.

Compliance Audits. Compliance auditing assesses whether AI systems adhere to relevant laws, regulations, and organizational policies. This may include verifying that systems meet requirements related to fairness, privacy, safety standards, or domain-specific regulatory obligations. As regulatory frameworks for AI continue to evolve, compliance audits play an important role in ensuring that organizations deploy AI systems responsibly.

Large-scale Impact Audits. Large-scale impact auditing considers the broader effects of AI systems once deployed. This includes identifying unintended harms, discriminatory outcomes, or systemic risks that may emerge across populations or institutions. This is especially important in the current AI climate, with frontier AI companies deploying models that garner millions of monthly users[1] from people across the world. Such audits often require longitudinal analysis and interdisciplinary approaches, as the societal consequences of AI systems may evolve and vary across contexts.

3 Blue-Sky Vision and Challenges

Position Statement: We argue for the development of principled auditing tools that assess AI systems in the wild and provide feedback mechanisms to maintain compliance with desired properties. We view the auditing of AI systems as a statistical problem of *monitoring constraint violations under uncertainty*. Modern AI systems are deployed in dynamic environments where data distributions shift, user behavior evolves, and ground truth labels may be delayed or unavailable. Consequently, auditing cannot rely solely on deterministic evaluation metrics or static benchmarks; instead, it must quantify the *risk* that deployed systems violate important constraints such as fairness, safety, or regulatory compliance.

Under this perspective, auditing becomes a problem of continuously monitoring whether deployed systems satisfy specified constraints as they interact with real-world environments. Because

[1]There were a reported 18.9 million monthly active Claude AI users in early 2026.

Table 1: Institutional actors involved in AI auditing and the levels of access typically available to them. Different actors contribute complementary forms of oversight, ranging from internal technical audits to independent public scrutiny.

Auditor Type	Access Level	Role	Example Audit Tasks
Internal Organizational Teams [31, 47]	White-box	Internal oversight and risk management	Audit training pipelines and datasets; conduct internal safety evaluations and red-teaming
Trusted Third-Party Auditors [53]	Grey-box	Independent technical evaluation	Assess model documentation and evaluation protocols; test for safety or fairness issues under controlled access
Regulators and Government Agencies [60]	Grey-box / White-box	Regulatory oversight and compliance checks	Verify safety adherence, privacy, or consumer protection requirements; review system documentation or logs
External Researchers and Civil Society [36, 53]	Black-box	Independent scrutiny of deployed systems	Probe systems by testing inputs and outputs; detect biases, unsafe behaviors, or unintended model responses
User Audits and Participatory Governance [18, 42, 59, 63]	Black-box	Real-world feedback and incident reporting	Surface unexpected failures or harmful outputs during everyday system use; report incidents and emerging risks

observations of system behavior may be noisy, incomplete, or influenced by human–AI interactions and surrounding infrastructure, auditing must reason about uncertainty in both the data and evaluation processes. This framing highlights several core challenges for AI auditing: specifying measurable constraints, estimating system behavior under uncertainty, and designing auditing mechanisms that operate continuously as systems evolve over time.

Lack of Trusted Specifications. Algorithmic fairness in machine learning [9] has been a sandbox for developing auditing tools [26, 35, 68], where notions of fairness have been specified in regulations (e.g., Four-Fifths Rule [22]). For example, figure 2 provides an example of fairness auditing as it pertains to the COMPAS recidivism dispute, where AVOIR [35] is used to validate both ProPublica's and NorthPointe's claims about the recidivism system's fairness, which depended on the specific fairness specification. However, the notion of safety in AI systems is less well-defined, with many safety and alignment works operating under the perspective that "we know unsafe behavior when we see it" [16, 40]. One of the key challenges and initiatives the community is starting to pursue is defining and specifying safety more precisely as something we can monitor and audit against. Individual domains, such as environmental and financial law, have defined their own specifications for specific use cases, suggesting that specifications need to be very siloed to make them quantifiable or evaluable, but also suggesting that we can pull from these other domains to come up with a way to define socio-technical specifications for AI safety.

Functional Auditing Challenges. Auditing models deployed in real-world, dynamic environments introduces new challenges, including drift, missing ground truth, and unexpected user behavior, which complicate evaluation [39]. A risk-based perspective would enable probabilistic measurement of constraints such as fairness, safety, and compliance, providing a statistical foundation for auditing. Recent works have explored methods to incorporate fairness constraints into uncertainty quantification frameworks [51, 58, 62]. Conformal Fairness (CF) is one such work that provides a method for performing fairness auditing in settings where the i.i.d. assumption does not hold post-deployment. Figure 1 illustrates the end-to-end pipeline for auditing post-deployment. A similar risk-based perspective can be applied to safety, with clearly defined specifications that accommodate context and uncertainty. While these methods use Conformal Prediction [64] as the underlying statistical framework, other statistical frameworks can be used to give guarantees on model performance [3, 10].

Operational, Data, and Compliance Challenges. AI infrastructure, data pipelines, and provenance tracking are fragmented in practice, complicating auditing [39]. In addition to evaluating and auditing the learning models, understanding the data used to train them is just as crucial [52], but is less well explored. There has been some work on auditing data membership [30]. There are also ways to characterize properties of data that determine the effectiveness, accuracy, and scalability of machine learning models, including the **5 Vs of big data**–volume, velocity, variety, veracity, and value [19]. Data audits often examine these dimensions to assess dataset quality and identify issues that may affect model performance. Data audits are further complicated by federated deployments, data sovereignty, and heterogeneous data types and modalities [14]. A similar risk-based auditing perspective to Conformal Fairness can be applied in federated settings [58], but the area remains ripe for exploration. Compliance audits must navigate evolving policy landscapes and ensure consistency across pre- and post-deployment phases. Addressing these challenges requires socio-technical solutions that integrate technical, legal, and organizational perspectives.

Human-AI Teams and Socio-Technical Challenges. Auditing models with humans in the loop, as users of AI (Human-AI teams) or as auditors, obfuscates model behavior with human intent [15], thereby undermining the reliability of the auditing results. The auditing problem is analogous to partial-observation problems in reinforcement learning and classical control theory [38], in which the true state (the audit results) must be distilled from raw, noisy observations. However, this is challenged by the unintelligible nature of human behavior and responses to information from ML models [17, 56]. To redress, we need to develop approaches that are robust to bias and uncertainty stemming from human behavior, draw on insights from the partial observability literature, and balance the socio-technical with humans-in-the-loop, while maintaining data privacy and human safety. Furthermore, the epistemic uncertainty of AI agents, combined with the cognitive uncertainty in human decision-making, suggests that auditing should shift toward quantifying the risk associated with the **joint behavior** of human-AI teams, enabling a more reliable assessment of system compliance

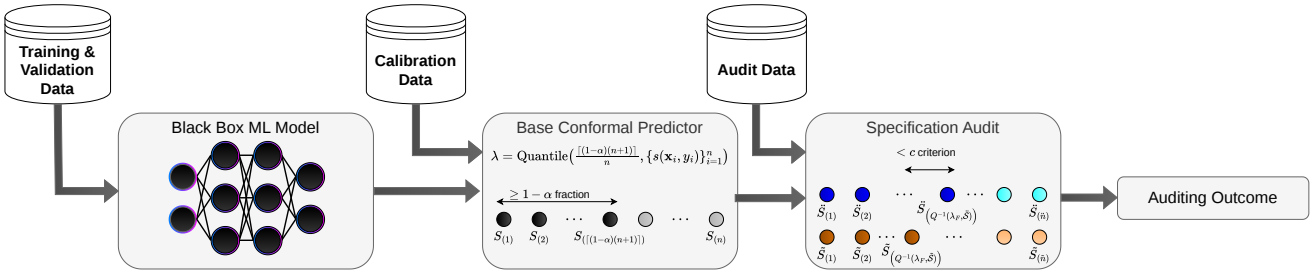


Figure 1: End-to-end Conformal Fairness pipeline. Once a classification model is trained, a conformal predictor is constructed and later audited using a held-out auditing set that is *exchangeable* with the original calibration data. (Image source: [62])

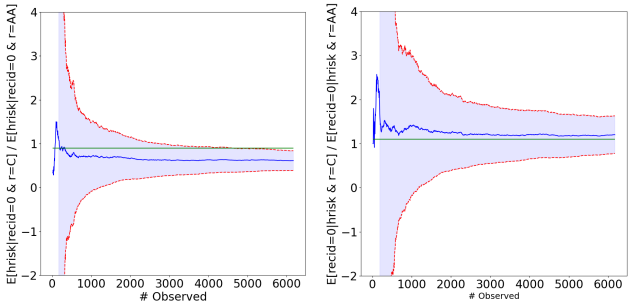


Figure 2: COMPAS dataset case study (Image source: [35]). (a) Analysis done by ProPublica using False Positive Rate Bias specification demonstrating a fairness violation. (b) Analysis done by Northpointe using the False Discovery Rate.

and performance. The increasing prevalence of AI agents for a broad spectrum of tasks necessitates further consideration.

Challenges with Modern AI Agents in Human-AI Teams. In a recent case study, many users recognized the potential value of agents for low-risk, repetitive tasks, but were skeptical of using AI agents for complex medical and financial decisions, particularly when errors could not be corrected [59]. Trust increased when AI agents offered transparency, user control, and step-by-step explanations. For silos, including health or finance, there was strong support for agents requiring explicit consent before completing tasks involving sensitive data. Thus, to increase user trust, a risk-sensitive approach to auditing will be essential for providing concrete guarantees of model performance, particularly in high-risk scenarios. Similarly, closing the development loop by incorporating audit feedback will instill trust in a continually evolving environment. Lastly, [59] elucidates the need for more user case studies to develop concrete audit specifications for agentic models.

Risk-Based Perspective on Constraint Violations Many auditing tasks can be interpreted as verifying whether deployed systems satisfy constraints that capture desirable properties such as fairness, safety, reliability, or regulatory compliance. In practice, however, these properties cannot typically be evaluated deterministically due to noisy observations, incomplete data, and evolving deployment environments. As a result, auditing must reason about the *risk* that a system violates a given constraint. Under this perspective, auditing becomes the task of estimating and monitoring the likelihood of constraint violations as systems interact with real-world environments. This framing naturally emphasizes uncertainty-aware auditing methods and motivates approaches that continuously monitor system behavior and trigger intervention when risks exceed acceptable thresholds. Recent work on uncertainty-aware evaluation methods, such as conformal approaches to fairness auditing [58, 62], illustrates one possible direction for operationalizing this perspective. Broadly, risk-control methods for sequential decision-making (cf. [4, 45, 48, 49, 67]) provide a relevant basis for encoding auditing specifications and for maintaining rigorous guarantees. This perspective provides a method for quantifying uncertainty in the auditing process, which we can then use to inform the feedback loop for model development. This view aligns with causal and mechanism-aware anomaly detection, which treats failures as violations of stable system invariants rather than distributional shifts [7, 43].

4 Concluding Remarks

What would success look like for AI auditing? Under this risk-based perspective, we would be able to quantify the uncertainty or the "risk" associated with an audit criterion, which is then used to inform subsequent model updates. If the information used to perform subsequent development results in a model that passes the audit (or controls the "risk"), we consider the audit a success.

Advances across multiple disciplines–including law, governance, management, economics, and health sciences, as well as computer science, engineering, signal processing, mathematics, and statistics–will be necessary to address the challenges of auditing modern AI systems. These challenges span the specification of measurable constraints (e.g., fairness, safety, and compliance), the development of mechanisms for monitoring deployed systems under uncertainty, and the design of auditing processes that operate throughout the lifecycle of an AI system. Comparable auditing practices already exist in sectors such as finance, healthcare, pharmaceuticals, and environmental regulation, where systems are monitored continuously and assessed against evolving regulatory and safety standards. An important direction for future work is therefore to examine how principles from these established auditing frameworks can inform the development of uncertainty-aware auditing mechanisms for modern AI systems across the AI lifecycle.

Acknowledgments

The authors acknowledge support from National Science Foundation (NSF) grant #2112471 (AI-EDGE). The authors' views and findings do not necessarily reflect those of the funding agencies.

References

[1] Charu C Aggarwal. 2016. *Recommender Systems* (1 ed.). Springer International Publishing, Cham, Switzerland.

[2] Shawqi Al-Maliki, Faissal El Bouanani, Mohamed Abdallah, Junaid Qadir, and Ala Al-Fuqaha. 2024. Addressing Data Distribution Shifts in Online Machine Learning Powered Smart City Applications Using Augmented Test-Time Adaptation. *IEEE Internet of Things Magazine* 7, 4 (July 2024), 116–124. doi:10.1109/iotm.001.2300135

[3] Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. 2023. Prediction-Powered Inference. arXiv:2301.09633 [stat.ML] https://arxiv.org/abs/2301.09633

[4] Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. 2024. Conformal Risk Control. In *The Twelfth International Conference on Learning Representations*.

[5] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine bias. In *Ethics of Data and Analytics*. Auerbach Publications, 254–264.

[6] AI Anthropic. 2025. System card: Claude opus 4 & claude sonnet 4. *Claude-4 Model Card* (2025).

[7] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2020. Invariant Risk Minimization. arXiv:1907.02893 [stat.ML] https://arxiv.org/abs/1907.02893

[8] K.J. Arrow. 1974. *The Limits of Organization*. Norton. https://books.google.com/books?id=_JHZAAAAMAAJ

[9] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2023. *Fairness and machine learning: Limitations and opportunities*. MIT press.

[10] Meshi Bashari, Roy Maor Lotan, Yonghoon Lee, Edgar Dobriban, and Yaniv Romano. 2025. Synthetic-Powered Predictive Inference. arXiv:2505.13432 [cs.LG] https://arxiv.org/abs/2505.13432

[11] Yoshua Bengio, Stephen Clare, Carina Prunkl, Maksym Andriushchenko, Ben Bucknall, Malcolm Murray, Rishi Bommasani, Stephen Casper, Tom Davidson, Raymond Douglas, et al. 2026. International AI Safety Report 2026. *arXiv preprint arXiv:2602.21012* (2026).

[12] Soumitra S Bhuyan, Vidyoth Sateesh, Naya Mukul, Alay Galvankar, Asos Mahmood, Muhammad Nauman, Akash Rai, Kahuwa Bordoloi, Urmi Basu, and Jim Samuel. 2025. Generative Artificial Intelligence Use in Healthcare: Opportunities for Clinical Excellence and Administrative Efficiency. *J Med Syst* 49, 1 (Jan. 2025), 10.

[13] Allison J. B. Chaney, Brandon M. Stewart, and Barbara E. Engelhardt. 2018. How algorithmic confounding in recommendation systems increases homogeneity and decreases utility. In *Proceedings of the 12th ACM Conference on Recommender Systems (RecSys '18)*. ACM, 224–232. doi:10.1145/3240323.3240370

[14] Hongyan Chang, Brandon Edwards, Anindya S Paul, and Reza Shokri. 2024. Efficient privacy auditing in federated learning. In *33rd USENIX Security Symposium (USENIX Security 24)*. 307–323.

[15] You Chen, Ellen Wright Clayton, Laurie Lovett Novak, Shilo Anders, and Bradley Malin. 2023. Human-Centered Design to Address Biases in Artificial Intelligence. *J Med Internet Res* 25 (2023), e43251.

[16] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA) *(NIPS'17)*. Curran Associates Inc., Red Hook, NY, USA, 4302–4310.

[17] Jesse C Cresswell, Bhargava Kumar, Yi Sui, and Mouloud Belbahri. 2025. Conformal Prediction Sets Can Cause Disparate Impact. In *The Thirteenth International Conference on Learning Representations*.

[18] Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The Participatory Turn in AI Design: Theoretical Foundations and the Current State of Practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization* (Boston, MA, USA) *(EAAMO '23)*. Association for Computing Machinery, New York, NY, USA, Article 37, 23 pages. doi:10.1145/3617694.3623261

[19] Yuri Demchenko, Paola Grosso, Cees de Laat, and Peter Membrey. 2013. Addressing big data issues in Scientific Data Infrastructure. In *2013 International Conference on Collaboration Technologies and Systems (CTS)*. 48–55. doi:10.1109/CTS.2013.6567203

[20] William Dieterich, Christina Mendoza, and Tim Brennan. 2016. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc* 7, 4 (2016).

[21] Matthew F Dixon, Igor Halperin, Paul Bilokon, et al. 2020. *Machine learning in finance*. Vol. 1170. Springer.

[22] The U.S. EEOC. 1979. Uniform guidelines on employee selection procedures. (March 1979).

[23] Gregory Falco, Ben Shneiderman, Julia Badger, Ryan Carrier, Anton Dahbura, David Danks, Martin Eling, Alwyn Goodloe, Jerry Gupta, Christopher Hart, et al. 2021. Governing AI safety through independent audits. *Nature Machine Intelligence* 3, 7 (2021), 566–571.

[24] Tongtong Feng, Xin Wang, Yu-Gang Jiang, and Wenwu Zhu. 2025. Embodied ai: From llms to world models. *IEEE CIRCUITS AND SYSTEMS MAGAZINE* (2025).

[25] Iason Gabriel, Arianna Manzini, Geoff Keeling, Lisa Anne Hendricks, Verena Rieser, Hasan Iqbal, Nenad Tomašev, Ira Ktena, Zachary Kenton, Mikel Rodriguez, et al. 2024. The ethics of advanced AI assistants. *arXiv preprint arXiv:2404.16244* (2024).

[26] Bishwamittra Ghosh, Debabrota Basu, and Kuldeep S Meel. 2021. Justicia: A stochastic SAT approach to formally verify fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 7554–7563.

[27] Moshe Glickman and Tali Sharot. 2025. How human-AI feedback loops alter human perceptual, emotional and social judgements. *Nat. Hum. Behav.* 9, 2 (Feb. 2025), 345–359.

[28] Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. 2024. AI Control: Improving Safety Despite Intentional Subversion. In *Proceedings of the 41st International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 235)*, Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (Eds.). PMLR, 16295–16336. https://proceedings.mlr.press/v235/greenblatt24a.html

[29] Dennis Hirsch, Timothy Bartley, Aravind Chandrasekaran, Davon Norris, Srinivasan Parthasarathy, and Piers Norris Turner. 2024. *Business data ethics: Emerging models for governing AI and advanced analytics*. Springer.

[30] Zonghao Huang, Neil Zhenqiang Gong, and Michael K. Reiter. 2024. A General Framework for Data-Use Auditing of ML Models. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*. ACM, 1300–1314. doi:10.1145/3658644.3690226

[31] Sathwik Karnik, Zhang-Wei Hong, Nishant Abhangi, Yen-Chen Lin, Tsun-Hsuan Wang, Christophe Dupuy, Rahul Gupta, and Pulkit Agrawal. 2024. Embodied red teaming for auditing robotic foundation models. *arXiv preprint arXiv:2411.18676* (2024).

[32] Noam Kolt, Nicholas Caputo, Jack Boeglin, Cullen O'Keefe, Rishi Bommasani, Stephen Casper, Mariano-Florentino Cuéllar, Noah Feldman, Iason Gabriel, Gillian K Hadfield, et al. 2026. Legal Alignment for Safe and Ethical AI. *arXiv preprint arXiv:2601.04175* (2026).

[33] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature* 521, 7553 (2015), 436–444.

[34] Jiashuo Liu, Zheyan Shen, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, and Peng Cui. 2023. Towards Out-Of-Distribution Generalization: A Survey. arXiv:2108.13624 [cs.LG] https://arxiv.org/abs/2108.13624

[35] Pranav Maneriker, Codi Burley, and Srinivasan Parthasarathy. 2023. Online Fairness Auditing through Iterative Refinement. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining* (Long Beach, CA, USA) *(KDD '23)*. Association for Computing Machinery, New York, NY, USA, 1665–1676. doi:10.1145/3580305.3599454

[36] Danaë Metaxa, Joon Sung Park, Ronald E. Robertson, Karrie Karahalios, Christo Wilson, Jeff Hancock, and Christian Sandvig. 2021. Auditing Algorithms: Understanding Algorithmic Systems from the Outside In. *Found. Trends Hum.-Comput. Interact.* 14, 4 (Nov. 2021), 272–344. doi:10.1561/1100000083

[37] Jakob Mökander. 2023. Auditing of AI: Legal, ethical and technical approaches. *Digital Society* 2, 3 (2023), 49.

[38] George E Monahan. 1982. State of the art—a survey of partially observable Markov decision processes: theory, models, and algorithms. *Management science* 28, 1 (1982), 1–16.

[39] Victor Ojewale, Ryan Steed, Briana Vecchione, Abeba Birhane, and Inioluwa Deborah Raji. 2025. Towards AI Accountability Infrastructure: Gaps and Opportunities in AI Audit Tooling. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. ACM, 1–29. doi:10.1145/3706598.3713301

[40] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL] https://arxiv.org/abs/2203.02155

[41] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*. 1–22.

[42] Ambreesh Parthasarathy, Aditya Phalnikar, Ameen Jauhar, Dhruv Somayajula, Gokul S Krishnan, and Balaraman Ravindran. 2024. Participatory Approaches in AI Development and Governance: A Principled Approach. arXiv:2407.13100 [cs.CY] https://arxiv.org/abs/2407.13100

[43] Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. 2016. Causal Inference by using Invariant Prediction: Identification and Confidence Intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology* 78, 5 (11 2016), 947–1012. doi:10.1111/rssb.12167

[44] Neoklis Polyzotis, Martin Zinkevich, Sudip Roy, Eric Breck, and Steven Whang. 2019. Data validation for machine learning. *Proceedings of machine learning and systems* 1 (2019), 334–347.

[45] Drew Prinster, Xing Han, Anqi Liu, and Suchi Saria. 2025. WATCH: Adaptive Monitoring for AI Deployments via Weighted-Conformal Martingales. In *Forty-second International Conference on Machine Learning*. https://openreview.net/

forum?id=GMjkK2CKx5
[46] Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. 2021. AI and the Everything in the Whole Wide World Benchmark. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren and S. Yeung (Eds.), Vol. 1. https://datasets-benchmarks-proceedings.neurips.cc/paper_files/paper/2021/file/084b6fbb10729ed4da8c3d3f5a3ae7c9-Paper-round2.pdf
[47] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 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*. ACM, 33–44.
[48] Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. 2023. Game-theoretic statistics and safe anytime-valid inference. *Statist. Sci.* 38, 4 (2023), 576–601.
[49] Aaditya Ramdas, Johannes Ruf, Martin Larsson, and Wouter Koolen. 2022. Admissible anytime-valid sequential inference must rely on nonnegative martingales. arXiv:2009.03167 [math.ST] https://arxiv.org/abs/2009.03167
[50] Anita Rao. 2026. *Challenges to the monitoring of deployed AI systems.* Technical Report. National Institute of Standards and Technology, Gaithersburg, MD.
[51] Yaniv Romano, Rina Foygel Barber, Chiara Sabatti, and Emmanuel Candès. 2020. With Malice Toward None: Assessing Uncertainty via Equalized Coverage. *Harvard Data Science Review* 2, 2 (apr 30 2020). https://hdsr.mitpress.mit.edu/pub/qedrwcz3.
[52] Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "Everyone wants to do the model work, not the data work": Data Cascades in High-Stakes AI. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–15.
[53] Christian Sandvig, Kevin Hamilton, Karrie Karahalios, and Cédric Langbort. 2014. Auditing Algorithms : Research Methods for Detecting Discrimination on Internet Platforms. https://api.semanticscholar.org/CorpusID:15686114
[54] David Sculley, Gary Holt, Daniel Golovin, Eugene Davydov, Todd Phillips, Dietmar Ebner, Vinay Chaudhary, Michael Young, Jean-Francois Crespo, and Dan Dennison. 2015. Hidden technical debt in machine learning systems. *Advances in neural information processing systems* 28 (2015).
[55] Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. 2025. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267* (2025).
[56] Edward Small, Yueqing Xuan, Danula Hettiachchi, and Kacper Sokol. 2023. Helpful, misleading or confusing: How humans perceive fundamental building blocks of artificial intelligence explanations. *arXiv preprint arXiv:2303.00934* (2023).
[57] Anutam Srinivasan, Antoine Leeman, and Glen Chou. 2026. Safety Beyond the Training Data: Robust Out-of-Distribution MPC via Conformalized System Level Synthesis. In *8th Annual Learning for Dynamics & Control Conference*.
[58] Anutam Srinivasan, Aditya T. Vadlamani, Amin Meghrazi, and Srinivasan Parthasarathy. 2026. FedCF: Fair Federated Conformal Prediction. https://openreview.net/forum?id=6rCsaBOQON
[59] Stanford Deliberative Democracy Lab and Center on Democracy, Development and the Rule of Law. 2026. *Industry-Wide Forum: Overall Summary.* Summary Report. Stanford University. In partnership with Meta, Oracle, DoorDash, PayPal, Cohere, and Microsoft.
[60] Elham Tabassi. 2023. *Artificial intelligence risk management Framework (AI RMF 1.0).* Technical Report. National Institute of Standards and Technology (U.S.), Gaithersburg, MD.
[61] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
[62] Aditya T. Vadlamani, Anutam Srinivasan, Pranav Maneriker, Ali Payani, and Srinivasan Parthasarathy. 2025. A Generic Framework for Conformal Fairness. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=xiQNfYl33p
[63] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3174014
[64] Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. 2005. *Algorithmic learning in a random world*. Vol. 29. Springer.
[65] WIRED. [n. d.]. Sun Called on Java Claims. https://www.wired.com/1997/11/sun-called-on-java-claims/
[66] Mario V Wüthrich and Michael Merz. 2023. *Statistical foundations of actuarial learning and its applications.* Springer.
[67] Ziyu Xu, Nikos Karampatziakis, and Paul Mineiro. 2024. Active, anytime-valid risk controlling prediction sets. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. https://openreview.net/forum?id=4ZH48aGD60
[68] Tom Yan and Chicheng Zhang. 2022. Active fairness auditing. In *International Conference on Machine Learning*. PMLR, 24929–24962.