

# **ML Auditing: Designing a Criteria Catalog**

**Nanekar Shravani Thakaji, Prof. K. M. Borude, Prof. L. S. Bhadake, Dr. H. B. Jadhav**

Department of Computer Engineering

Adsul's Technical Campus, Chas, Ahilyanagar, Maharashtra, India

shravaninanekar7@gmail.com, lokeshb.atc@gmail.com

**Abstract:** *Artificial Intelligence (AI) and Machine Learning (ML) are increasingly being used in healthcare and other sectors to improve productivity, decision-making, accuracy, and service quality. However, the use of these technologies also creates challenges related to fairness, transparency, data quality, safety, accountability, and reliability. Therefore, auditing machine learning systems is essential to ensure that they operate responsibly and produce trustworthy results. This study focuses on developing a structured set of criteria for auditing machine learning systems based on existing research and published literature.*

*The study follows a scoping review approach to identify and analyze existing concepts, methods, tools, and practices related to auditable artificial intelligence. The reviewed literature was examined using qualitative content analysis to identify common themes and organize them into a practical auditing framework. The resulting ML Auditing Criteria Catalog contains 30 questions divided into three major categories: Conceptual Basics, Data and Algorithm Design, and Assessment Metrics. These categories address important aspects such as risk management, project methodology, ethical principles, data representativeness, model complexity, documentation, fairness, and performance evaluation.*

*The proposed catalog provides a systematic approach for identifying potential weaknesses and assessing the quality of machine learning systems throughout their development and deployment. It also highlights the importance of qualitative and quantitative evaluation methods, including confusion matrices, precision, sensitivity, specificity, F1-score, and Area Under the Curve (AUC). Furthermore, tools such as LIME and SHAP can help explain model predictions, while continuous monitoring and human oversight support the responsible use of AI. The catalog is intended as a starting point rather than a complete auditing standard and requires further evaluation in real-world or simulated projects. Overall, this study emphasizes that trustworthy machine learning depends not only on prediction accuracy but also on fairness, transparency, suitable data, accountability, and ongoing assessment..*

**Keywords:** Artificial Intelligence, Machine Learning, ML Auditing, Auditable AI, Algorithmic Fairness, Data Quality, Model Evaluation, Transparency, Risk Management, Healthcare AI

## **I. INTRODUCTION**

Artificial Intelligence (AI) and Machine Learning (ML) are increasingly used in healthcare and other sectors to improve productivity, automate tasks, analyze data, and support decision-making. These technologies help organizations identify patterns, generate predictions, and improve the quality of services. However, machine learning systems may also produce incorrect results, show bias, or perform differently for different groups of people. Therefore, it is important to evaluate these systems carefully to ensure their fairness, transparency, safety, and reliability.

Machine learning auditing is the process of examining an AI system to determine whether it behaves as expected and meets its intended requirements. It involves checking data quality, algorithm design, model performance, documentation, risk management, and ethical considerations. Various evaluation methods, including confusion matrices, precision, sensitivity, specificity, F1-score, and AUC, can help measure model performance. Tools such as LIME and SHAP can also support the interpretation of model predictions. Regular monitoring and human oversight are necessary to identify potential problems and maintain the reliability of machine learning systems after deployment.



The study “ML Auditing: Designing a Criteria Catalog” presents a structured framework containing 30 questions for auditing machine learning systems. Based on a review of existing literature, the criteria are organized into three main categories: Conceptual Basics, Data and Algorithm Design, and Assessment Metrics. The framework addresses important aspects such as risk management, data representativeness, algorithm complexity, fairness, documentation, and statistical evaluation. It provides a practical starting point for organizations to identify weaknesses and improve AI auditing practices. The study emphasizes that trustworthy machine learning requires not only accurate predictions but also reliable data, accountability, transparency, fairness, and continuous assessment.

## **II. PROBLEM STATEMENT**

The increasing use of Artificial Intelligence (AI) and Machine Learning (ML) in healthcare and other sectors has created a need for reliable methods to evaluate the safety, fairness, transparency, and accuracy of these systems. However, many organizations lack a complete and standardized approach to auditing machine learning models. Problems such as poor data quality, algorithmic bias, unclear decision-making processes, insufficient documentation, and limited accountability can affect the reliability of model predictions and lead to undesirable outcomes.

The main problem addressed in this study is the absence of a universally accepted auditing process that covers the important technical, ethical, and operational aspects of machine learning systems. Existing auditing approaches may focus on individual areas, making it difficult for organizations to evaluate the overall quality and trustworthiness of an AI system. Therefore, there is a need for a structured framework that helps identify potential risks, assess data and algorithm design, measure model performance, and ensure appropriate human oversight.

## **III. OBJECTIVES**

To study the fundamental concepts and importance of Machine Learning (ML) auditing.

To identify potential risks related to fairness, transparency, safety, and accountability in AI systems.

To examine the quality, representativeness, and reliability of data used in machine learning models.

To evaluate machine learning algorithms using appropriate qualitative and quantitative assessment metrics.

To develop a structured ML Auditing Criteria Catalog containing 30 questions to support the systematic auditing of machine learning systems.

## **IV. LITERATURE SURVEY**

### **1. Schwarz, Hinske, Mansmann and Albashiti (2024)**

#### **Title: ML Auditing Criteria Catalog as a Starting Point for Developing an Auditing Framework**

This study focuses on developing a structured framework for auditing machine learning systems. The authors reviewed existing publications and used qualitative content analysis to organize important auditing concepts into a catalog of 30 questions. The catalog is divided into three categories: Conceptual Basics, Data and Algorithm Design, and Assessment Metrics. It covers risk management, ethical considerations, data quality, algorithm selection, fairness, documentation, and model performance. The study provides a practical starting point for organizations seeking to evaluate machine learning systems systematically. However, the authors emphasize that the catalog requires further testing and improvement through practical applications.

### **2. Berghoff et al. (2021)**

#### **Title: Towards Auditable AI Systems: Current Status and Future Directions**

This paper examines the current challenges and future directions of auditing artificial intelligence systems. The authors discuss important areas such as AI system security, robustness, verification, explainability, and safety. The study highlights the difficulty of establishing reliable auditing procedures for complex machine learning models, particularly when their internal decision-making processes are difficult to understand. It emphasizes the importance of technical testing, appropriate documentation, and continuous evaluation throughout the AI lifecycle. The paper provides a



foundation for understanding the requirements of auditable AI systems and identifies areas where further research and standardized methods are needed.

### 3. Zicari et al. (2021)

#### **Title: Z-Inspection®: A Process to Assess Trustworthy AI**

This research introduces Z-Inspection, a process designed to assess whether artificial intelligence systems can be considered trustworthy. The authors present an approach based on applied ethics and the principles of trustworthy AI. The process can be applied in different domains, including healthcare, business, and the public sector. It considers ethical and societal implications alongside the technical characteristics of AI systems. The study demonstrates the importance of examining AI applications from multiple perspectives and involving relevant stakeholders in the assessment process. Its findings are relevant to ML auditing because they support systematic evaluation of ethical risks, transparency, accountability, and responsible AI development.

### 4. Ribeiro, Singh and Guestrin (2016)

#### **Title: “Why Should I Trust You?”: Explaining the Predictions of Any Classifier**

This paper introduces Local Interpretable Model-Agnostic Explanations (LIME), a method for explaining the predictions of machine learning classifiers. The authors address the difficulty of understanding complex models that function as black boxes. LIME explains an individual prediction by approximating the model's behavior around a specific input using an interpretable model. The research demonstrates the usefulness of explanations for evaluating predictions, comparing models, and identifying potentially unreliable outputs. This work is relevant to machine learning auditing because explainability can help reviewers understand model behavior and investigate possible weaknesses. However, explanations alone cannot establish that a model is fair, safe, or accurate.

### 5. Mitchell et al. (2019)

#### **Title: Model Cards for Model Reporting**

This paper proposes model cards as a method for documenting machine learning models and improving transparency. Model cards provide information about a model's intended purpose, evaluation procedures, performance characteristics, limitations, and suitability for different groups or conditions. The authors explain that reporting performance across relevant demographic and intersectional groups can help users understand where a model performs well and where it may have limitations. The framework supports responsible model development by encouraging clear documentation and transparent communication of model capabilities. This research is closely related to ML auditing because documentation helps auditors evaluate a model's design, intended use, performance, and potential fairness concerns.

## COMPARISON TABLE

| Author & Year          | Method Used                                                                  | Advantages                                                                                        | Limitations                                                                                       |
|------------------------|------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|---------------------------------------------------------------------------------------------------|
| Schwarz et al. (2024)  | Scoping review and qualitative content analysis                              | Provides a structured catalog of 30 auditing questions covering three major areas of ML auditing. | Requires further testing and validation in real-world projects.                                   |
| Berghoff et al. (2021) | Literature review and analysis of auditable AI systems                       | Highlights the importance of security, robustness, verification, and explainability.              | Establishing a common auditing approach for complex AI systems remains challenging.               |
| Zicari et al. (2021)   | Z-Inspection process based on ethical assessment and stakeholder involvement | Supports the evaluation of trustworthy AI, ethical risks, transparency, and accountability.       | Assessment can be complex and depends on the context and involvement of relevant stakeholders.    |
| Ribeiro et al. (2016)  | LIME: Local, model-agnostic explanation of individual predictions            | Helps users understand individual predictions and investigate black-box models.                   | Explains individual predictions but does not independently establish overall fairness, safety, or |



|                        |                                                                          |                                                                                                              |                                                                                                  |
|------------------------|--------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------|
|                        |                                                                          |                                                                                                              | reliability.                                                                                     |
| Mitchell et al. (2019) | Model Cards for documenting model characteristics and evaluation results | Improves transparency by documenting intended use, performance, limitations, and relevant evaluation groups. | The quality of reporting depends on the completeness and accuracy of the documentation provided. |

## V. WORKING OF SYSTEM

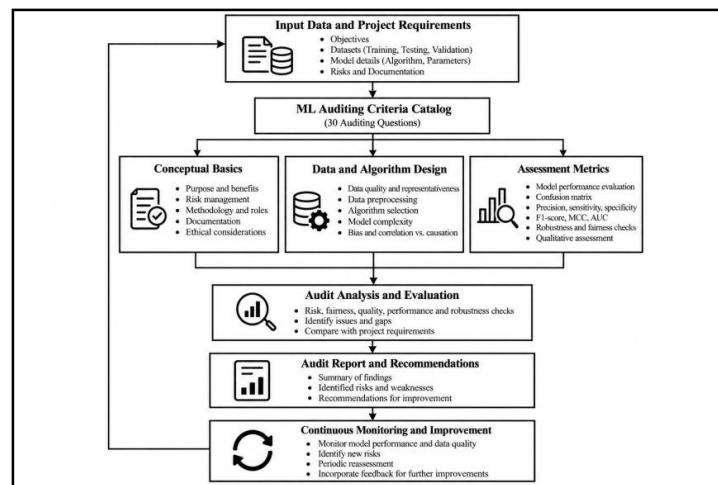


Fig 1: Design of the system

### 1. Problem Definition and Risk Identification

The first step is to define the purpose and expected outcomes of the machine learning system. The development team identifies the benefits, possible risks, and consequences of incorrect predictions. Potential issues such as bias, unsafe behavior, and unreliable results are considered. Responsibilities for managing risks are also established. This stage provides a clear foundation for the auditing process.

### 2. Project Planning and Documentation

The project team selects an appropriate development and auditing methodology. Roles, responsibilities, requirements, and acceptance criteria are clearly defined before implementation. Documentation is prepared to describe the dataset, model characteristics, intended use, and limitations. Data sheets and model cards can help maintain transparent records. Proper documentation supports systematic evaluation and future auditing.

### 3. Data Quality Assessment

The training, testing, and validation datasets are examined to determine their quality and suitability. The assessment considers data accuracy, consistency, completeness, and freshness. The team checks whether the data represents the people or situations in which the model will be used. Missing values, conflicting records, and possible sources of bias are also reviewed. High-quality and representative data helps support reliable and fair model predictions.

### 4. Algorithm Design and Selection

The machine learning algorithm is evaluated according to the problem requirements and expected performance. The team considers model complexity, interpretability, and the relationship between input features and predicted outcomes. Simpler models, such as decision trees and logistic regression, may be considered when suitable. The difference between correlation and causation is also examined when interpreting results. The selected algorithm should meet the required specifications without unnecessary complexity.



### **5. Model Performance Evaluation**

The developed model is tested using appropriate qualitative and quantitative assessment methods. A confusion matrix helps identify true positives, true negatives, false positives, and false negatives. Metrics such as precision, sensitivity, specificity, F1-score, MCC, and AUC are used to evaluate predictive performance. Statistical tests may also be applied to examine data properties and model assumptions. The results are compared with predefined performance requirements to identify possible weaknesses.

### **6. Explainability and Fairness Testing**

The model is examined to understand its predictions and identify possible unfair outcomes. Tools such as LIME and SHAP can help explain individual predictions and estimate the contribution of input features. Fairness measures can be used to compare outcomes and prediction errors across different groups. Robustness testing checks whether small or expected input changes produce unexpected results. These evaluations help identify potential issues, although explanations alone do not guarantee that a model is fair or safe.

### **7. Audit Results and Documentation**

After testing, the audit findings are recorded in a structured manner. The report includes model assumptions, data-handling procedures, evaluation results, identified risks, and limitations. The results are compared with the system specifications and intended user requirements. Any weaknesses discovered during the audit are reviewed so that appropriate corrective actions can be considered. Clear records also help independent reviewers understand how the model was developed and evaluated.

### **8. Continuous Monitoring and Improvement**

Machine learning auditing continues after the model is deployed. The system should be monitored to identify changes in data quality, prediction performance, and operating conditions. Organizations should periodically review fairness, reliability, and compliance with applicable requirements. Human oversight and procedures for responding to unexpected behavior are important, particularly in high-risk applications. Feedback from practical use can help improve the auditing process and maintain the trustworthiness of the machine learning system.

## **VI. SYSTEM DESIGN**

The system design of the ML Auditing Criteria Catalog describes the overall structure and workflow used to evaluate machine learning systems. It organizes the auditing process into three main modules: Conceptual Basics, Data and Algorithm Design, and Assessment Metrics. Each module examines specific aspects of the ML system to identify risks, evaluate model quality, and support trustworthy AI development.

### **1. Input Layer**

The input layer collects information about the machine learning project, including its objectives, training data, testing data, validation data, algorithm specifications, and intended application. It also identifies potential risks, user requirements, and relevant documentation. These inputs provide the information needed to begin the auditing process.

### **2. Conceptual Basics Module**

This module evaluates the purpose, expected benefits, and potential risks of the AI system. It examines project methodology, responsibilities, documentation, ethical principles, and risk management procedures. It also considers whether appropriate human oversight and independent auditing are available.

### **3. Data and Algorithm Design Module**

This module examines the quality, completeness, consistency, and representativeness of the data. It evaluates the suitability and complexity of the selected algorithm and checks whether the problem and expected model behavior are clearly defined. It also considers the difference between correlation and causation and identifies potential sources of bias.

### **4. Assessment Metrics Module**

This module evaluates the machine learning model using qualitative and quantitative methods. Qualitative assessment includes reviewing documentation, assumptions, and testing records. Quantitative assessment uses metrics such as



precision, sensitivity, specificity, F1-score, MCC, and AUC. Robustness and fairness tests help identify unexpected predictions and differences in outcomes between groups.

### 5. Audit Analysis and Reporting

The results from the three modules are reviewed against the project's requirements and acceptance criteria. Identified issues, including data-quality problems, unfair predictions, weak performance, and insufficient documentation, are recorded. The audit report provides evidence that can help developers and organizations decide which improvements are required.

### 6. Monitoring and Feedback

After deployment, the machine learning system should be monitored for changes in performance, data quality, and operating conditions. Feedback from users and audit results can help identify new risks and opportunities for improvement. Periodic reassessment and human oversight support the continued reliability and responsible use of the system.

### Overall System Design

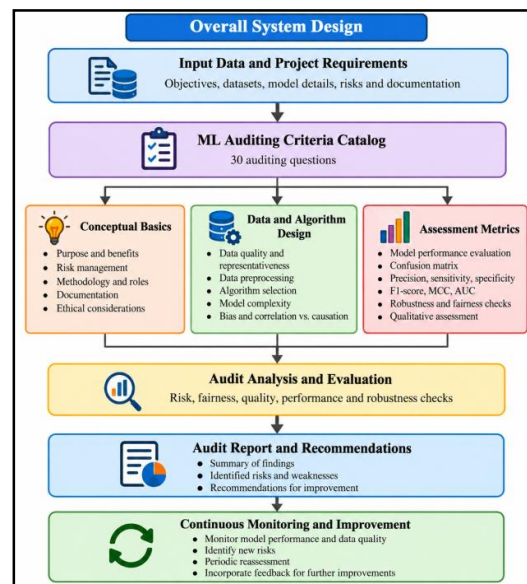


Fig 1: Overall System Design

## VII. RESULTS

The study resulted in the development of an ML Auditing Criteria Catalog containing 30 questions organized into three main categories: Conceptual Basics, Data and Algorithm Design, and Assessment Metrics. The review identified important aspects of machine learning auditing, including risk management, project methodology, data quality, algorithm selection, fairness, documentation, and model evaluation.

The results highlight that machine learning auditing requires both qualitative and quantitative assessment. Qualitative assessment includes reviewing documentation, assumptions, and testing records, while quantitative assessment uses measures such as precision, sensitivity, specificity, F1-score, MCC, and AUC. The study also identifies the importance of robustness testing and fairness measures in evaluating model behavior.

The catalog provides a structured starting point for organizations to identify potential weaknesses and improve their machine learning auditing practices. However, the study does not establish that the catalog has been validated in real-world applications. Further testing using real or simulated projects is required to assess its practical effectiveness and identify areas for improvement.



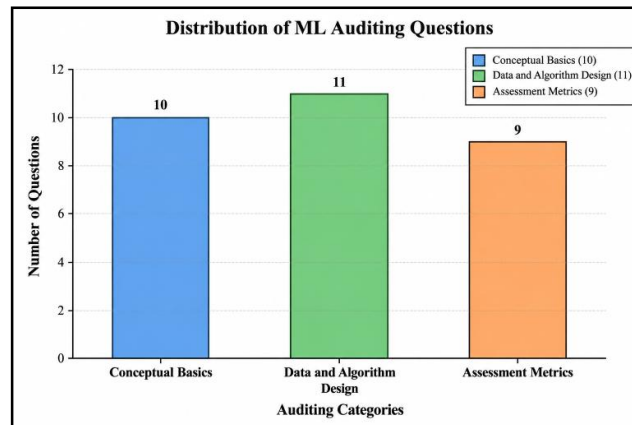


Figure 3: Classification of ML Auditing Criteria

| Sr. No. | Auditing Category         | Main Focus                                               |
|---------|---------------------------|----------------------------------------------------------|
| 1       | Conceptual Basics         | Risks, methodology, responsibilities and ethics          |
| 2       | Data and Algorithm Design | Data quality, representativeness and algorithm selection |
| 3       | Assessment Metrics        | Performance, robustness and fairness evaluation          |

**Interpretation:** The table shows the three main categories of the ML Auditing Criteria Catalog. Each category focuses on different aspects of machine learning auditing. Together, they support the systematic assessment of data quality, algorithm design, model performance, fairness, and risk management. The catalog contains 30 questions in total, but the PDF does not specify the number of questions assigned to each category.

## VIII. CONCLUSION

The study of the ML Auditing Criteria Catalog highlights the importance of systematically evaluating machine learning systems to ensure their reliability, fairness, transparency, and safety. Based on the reviewed literature, the catalog organizes 30 auditing questions into three major categories: Conceptual Basics, Data and Algorithm Design, and Assessment Metrics. These categories cover essential aspects such as risk management, data quality, algorithm selection, ethical considerations, documentation, and model performance evaluation.

The study shows that trustworthy machine learning requires more than high prediction accuracy. Suitable data, fair outcomes, understandable results, proper documentation, human oversight, and continuous monitoring are also essential. Quantitative measures such as precision, sensitivity, specificity, F1-score, MCC, and AUC can support performance evaluation, while tools such as LIME and SHAP can help explain individual predictions. These methods contribute to a better understanding of model behavior but cannot independently guarantee that a system is completely trustworthy.

In conclusion, the ML Auditing Criteria Catalog provides a practical starting point for organizations and project teams to identify potential risks and improve their auditing practices. However, it is not a complete or universally accepted auditing standard. Further evaluation using real-world or simulated projects is necessary to determine its practical effectiveness and improve the framework. The study emphasizes that responsible AI development requires a systematic approach combining technical evaluation, ethical considerations, accountability, and continuous improvement.

## IX. FUTURE SCOPE

The ML Auditing Criteria Catalog has significant potential for further development and practical application in machine learning projects. In the future, the catalog can be tested using real-world and simulated projects to evaluate its usefulness and identify areas requiring improvement. Automated auditing tools can be developed to support data-



quality assessment, model-performance evaluation, fairness testing, and robustness analysis. The framework can also be expanded to improve the identification of algorithmic bias and ensure that machine learning systems produce fair and reliable outcomes for different groups of users. Continuous monitoring mechanisms may help detect changes in model performance and data characteristics after deployment. Furthermore, future research can focus on improving explainability through tools such as LIME and SHAP, strengthening documentation practices, and incorporating relevant ethical and regulatory requirements. Feedback from developers, auditors, and domain experts can help refine the 30-question catalog and make it more practical across different industries. These improvements can support the development of more transparent, accountable, safe, and trustworthy artificial intelligence systems.

## REFERENCES

- Schwarz, M., Hinske, L. C., Mansmann, U., & Albashiti, F. (2024). "Designing an ML Auditing Criteria Catalog as Starting Point for the Development of a Framework." *IEEE Access*, 12, 39953–39967. <https://doi.org/10.1109/ACCESS.2024.3375763>
- Berghoff, C., et al. (2021). Towards Auditable AI Systems: Current Status and Future Directions. Fraunhofer Heinrich Hertz Institute and partner organizations.
- Zicari, R. V., et al. (2021). "Z-Inspection®: A Process to Assess Trustworthy AI." *IEEE Transactions on Technology and Society*, 2(2), 83–97. <https://doi.org/10.1109/TTS.2021.3066209>
- Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "“Why Should I Trust You?”: Explaining the Predictions of Any Classifier." *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144. <https://doi.org/10.1145/2939672.2939778>
- Mitchell, M., et al. (2019). "Model Cards for Model Reporting." *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. <https://doi.org/10.1145/3287560.3287596>
- Gebru, T., et al. (2021). "Datasheets for Datasets." *Communications of the ACM*, 64(12), 86–92. <https://doi.org/10.1145/3458723>
- Barocas, S., Hardt, M., & Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.
- Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys*, 54(6), Article 115. <https://doi.org/10.1145/3457607>
- Suresh, H., & Guttat, J. V. (2021). "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle." *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. <https://doi.org/10.1145/3465416.3483305>
- Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.
- Molnar, C. (2022). *Interpretable Machine Learning* (2nd ed.). <https://christophm.github.io/interpretable-ml-book/>
- Lundberg, S. M., & Lee, S.-I. (2017). "A Unified Approach to Interpreting Model Predictions." *Advances in Neural Information Processing Systems*, 30.
- Doshi-Velez, F., & Kim, B. (2017). "Towards a Rigorous Science of Interpretable Machine Learning." *arXiv:1702.08608*. <https://arxiv.org/abs/1702.08608>
- Amershi, S., et al. (2019). "Software Engineering for Machine Learning: A Case Study." *2019 IEEE/ACM 41st International Conference on Software Engineering: Software Engineering in Practice*, 291–300. <https://doi.org/10.1109/ICSE-SEIP.2019.00042>
- Breck, E., et al. (2017). "The ML Test Score: A Rubric for ML Production Readiness and Technical Debt Reduction." *2017 IEEE International Conference on Big Data*, 1123–1132. <https://doi.org/10.1109/BigData.2017.8258038>



16. Raji, I. D., et al. (2020). "Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing." Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, 33–44. <https://doi.org/10.1145/3351095.3372873>
17. Kroll, J. A., et al. (2017). "Accountable Algorithms." University of Pennsylvania Law Review, 165(3), 633–705.
18. Selbst, A. D., et al. (2019). "Fairness and Abstraction in Sociotechnical Systems." Proceedings of the Conference on Fairness, Accountability, and Transparency, 59–68. <https://doi.org/10.1145/3287560.3287598>
19. 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." Proceedings of the Conference on Fairness, Accountability, and Transparency, 220–229. <https://doi.org/10.1145/3287560.3287596>
20. Organisation for Economic Co-operation and Development (OECD). (2023). Advancing Accountability in AI: Governing and Managing Risks throughout the Lifecycle for Trustworthy AI. OECD Publishing.

