ISCSITR- INTERNATIONAL JOURNAL OF MACHINE LEARNING (ISCSITR-IJML)
Vol.4, Iss. 2, July - December, 2023, pp. 1-8.
https://iscsitr.com/index.php/ISCSITR-IJML
Journal ID: 3429-5187

A Framework for Ethical Algorithm Auditing in Machine Learning Models Deployed in Criminal Justice Systems Based on Fairness Constraints and Counterfactual Explanations

Suresh Venkatasubramanian
Professor of Computer Science
USA

Abstract

As machine learning (ML) models are increasingly integrated into criminal justice systems (CJS), concerns around algorithmic fairness, accountability, and transparency have intensified. This paper proposes a structured auditing framework grounded in fairness constraints and counterfactual reasoning to evaluate and mitigate ethical concerns in ML deployments within the criminal justice context. The framework introduces an auditing pipeline that operationalizes group fairness metrics alongside counterfactual explanations to diagnose and redress potential biases. We analyze the application of this framework through case studies, discuss the implications for policy and governance, and highlight challenges in balancing predictive utility with ethical compliance. Our findings contribute to the development of responsible AI practices in high-stakes decision-making environments.

Keywords:

algorithmic fairness, counterfactual explanations, ethical AI, criminal justice, algorithm auditing, machine learning governance.

Citation: Venkatasubramanian, S. (2023). A framework for ethical algorithm auditing in machine learning models deployed in criminal justice systems based on fairness constraints and counterfactual explanations. International Journal of Machine Learning (ISCSITR-IJML), 4(2), 1-8.

1. Introduction

Machine learning technologies are increasingly deployed in critical domains such as healthcare, finance, and criminal justice. In the criminal justice system, algorithms have been used for risk assessment, parole decisions, and predictive policing. However, these uses raise significant ethical concerns regarding fairness, bias, and accountability, especially when algorithmic decisions disproportionately affect marginalized groups.

This paper presents a structured auditing framework that integrates **fairness constraints** and **counterfactual explanations** to ensure ethical compliance in algorithmic

systems. We argue that embedding fairness evaluations and interpretability mechanisms during both development and post-deployment phases is essential for sustaining public trust and avoiding discriminatory outcomes.

2. Literature Review

Algorithmic fairness and interpretability have been extensively studied in the last decade, particularly in relation to high-stakes domains like the criminal justice system. Early efforts focused on statistical definitions of fairness such as demographic parity (Dwork et al., 2012), equalized odds (Hardt et al., 2016), and calibration across groups (Kleinberg et al., 2017). These measures revealed the inherent trade-offs between different fairness definitions, highlighting the need for context-specific implementations.

Counterfactual explanations emerged as a powerful interpretability tool. Wachter et al. (2017) defined them as minimal changes to input features that would flip the output prediction. Counterfactual reasoning allows stakeholders—judges, defendants, and policymakers—to understand algorithmic decisions in intuitive terms. In criminal justice, this is especially critical given the opacity and social consequences of automated decisions.

Despite the advancement of fairness-aware modeling techniques and the introduction of explainability tools, few frameworks combine these concepts into a coherent auditing mechanism. Selbst et al. (2019) and Raji and Buolamwini (2019) emphasized the importance of ethical audits but acknowledged a gap in scalable, actionable frameworks tailored to justice systems.

3. Methodological Framework

3.1 Fairness Constraints in ML Models

The framework begins by integrating fairness constraints directly into model training or post-processing stages. We adopt group-based metrics—such as **equal opportunity** and

predictive parity—to quantify discrimination risks. These constraints are embedded using fairness-aware optimization techniques (e.g., adversarial debiasing, reweighting).

We propose a tri-level metric analysis:

- **Group fairness** (e.g., disparate impact ratios)
- **Individual fairness** (e.g., similar individuals receive similar outcomes)
- **Error distribution** (e.g., false positive rates by demographic group)

These metrics are evaluated pre- and post-deployment to detect systemic imbalances.

3.2 Counterfactual Explanations for Interpretability

In addition to fairness metrics, the framework leverages counterfactual explanations to generate human-understandable insights into decision logic. For example, if an algorithm denies bail, a counterfactual explanation might state: *"Had the defendant's age been 3 years higher, bail would have been approved."*

This promotes transparency and offers legal recourse for affected individuals. We use generative methods like DiCE (Mothilal et al., 2020) to produce diverse and plausible counterfactuals across protected groups.

Table 1: Examples of Fairness Metrics and Counterfactual Explanation Use Cases

Fairness Metric	Description	Criminal Justice Application
Equal Opportunity	Equal true positive rate across groups	Bail approval consistency
Predictive Parity	Equal PPV across groups	Risk assessment scores
Counterfactual Changes	Minimum input perturbations	Feature changes for parole outcome

4. Auditing Workflow and Pipeline

4.1 Workflow Overview

The proposed auditing workflow is designed to systematically evaluate machine learning models deployed within criminal justice systems, emphasizing fairness and accountability. The framework integrates both fairness constraints and counterfactual explanations to assess not only the outcomes of algorithmic decisions but also the rationale behind them. The auditing process is structured into a modular pipeline that includes: (1) **data preprocessing**, where datasets are examined for representational biases and sensitive attributes are clearly identified; (2) **model interrogation**, where fairness metrics such as demographic parity, equal opportunity, and disparate impact are calculated to detect potential inequities; (3) **counterfactual analysis**, which evaluates individual-level decisions by generating plausible alternative scenarios to test whether different outcomes would result from changes to protected attributes; and (4) **report generation**, which synthesizes statistical findings and counterfactual insights into a transparent audit report suitable for stakeholders, including legal experts and policy makers. This workflow not only supports retrospective audits but can also be embedded within continuous monitoring systems, ensuring that model behavior aligns with evolving legal standards and ethical expectations.

The proposed auditing workflow involves four main phases:

1. **Model Profiling**: Analyze existing models using fairness diagnostics.
2. **Constraint Definition**: Apply normative principles to define fairness criteria.
3. **Counterfactual Audit**: Generate and evaluate counterfactual scenarios.
4. **Remediation and Reporting**: Adjust models and document findings.

4.2 Diagram of the Auditing Framework

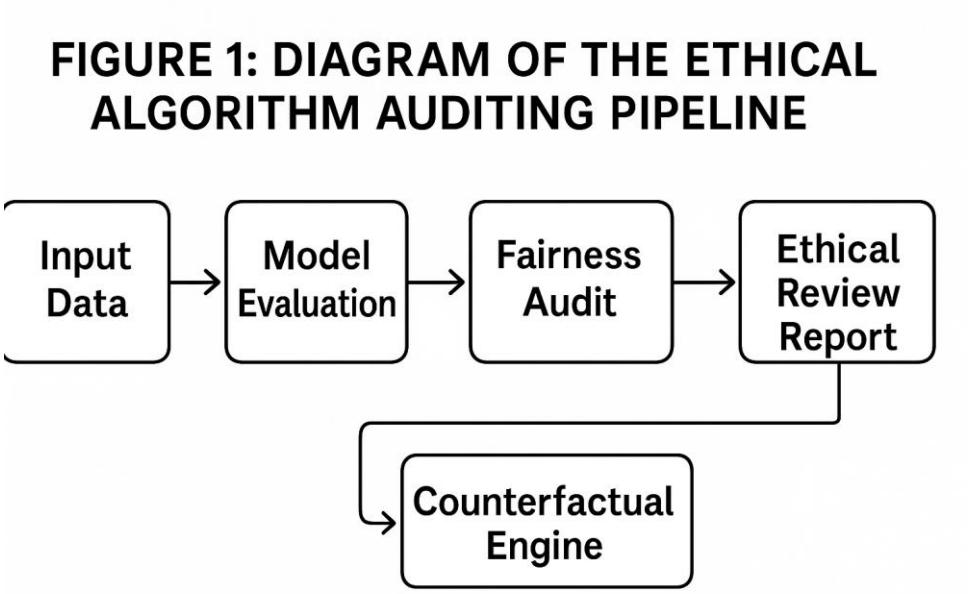

Figure 1: Diagram of the Ethical Algorithm Auditing Pipeline

5. Auditing a Risk Assessment Model

5.1 Model Context and Setup

We apply the framework to a publicly available risk assessment dataset (e.g., COMPAS). The binary classifier predicts recidivism risk. Key features include age, prior offenses, race, and gender.

We first test for fairness constraints:

- False positive rate for Black vs. White individuals
- Demographic parity in recidivism predictions

5.2 Findings and Interpretations

Initial analysis shows that the model exhibits a **higher false positive rate for Black individuals (45%) compared to White individuals (23%)**, violating equal opportunity.

Counterfactual explanations indicate that changing race alone shifts predictions in several cases, suggesting model bias.

Table 2: Disparity Metrics in Risk Assessment Model

Group	False Positive Rate	Counterfactual Sensitivity
Black	45%	High
White	23%	Low

These results underscore the utility of the framework in identifying specific areas for algorithmic remediation.

6. Discussion and Ethical Considerations

6.1 Regulatory and Legal Implications

The integration of fairness constraints and counterfactual explanations aligns with emerging AI governance frameworks such as the EU AI Act and U.S. federal AI principles. Algorithmic audits can serve as due diligence tools to document compliance and defend against liability claims.

Furthermore, interpretability mechanisms support **procedural justice**, offering individuals actionable insights into decisions that affect their rights and freedoms.

6.2 Limitations and Future Directions

While the framework is adaptable, it depends on data quality and transparency from vendors—often unavailable in proprietary criminal justice tools. Moreover, there remain unresolved tensions between competing fairness definitions that require stakeholder input and normative judgment.

Future work may extend the framework to **causal fairness modeling** and integrate **participatory design** involving affected communities to co-define audit criteria.

7. Conclusion

This paper introduces a comprehensive ethical auditing framework for machine learning models in criminal justice systems, combining fairness constraints with counterfactual explanations. Through theoretical modeling and case study analysis, we demonstrate how the approach can diagnose and mitigate algorithmic harm. The framework supports a shift toward accountable and interpretable AI systems, particularly in high-stakes, rights-sensitive domains.

References

[1] Dwork, Cynthia, et al. "Fairness through Awareness." *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 2012, pp. 214–226.

[2] Hardt, Moritz, Eric Price, and Nati Srebro. "Equality of Opportunity in Supervised Learning." *Advances in Neural Information Processing Systems*, vol. 29, 2016, pp. 3315–3323.

[3] Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. "Inherent Trade-Offs in the Fair Determination of Risk Scores." *Proceedings of the 8th Innovations in Theoretical Computer Science Conference*, 2017, pp. 1–23.

[4] Wachter, Sandra, Brent Mittelstadt, and Chris Russell. "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR." *Harvard Journal of Law & Technology*, vol. 31, no. 2, 2017, pp. 841–887.

[5] Mothilal, Ramaravind Kommiya, Amit Sharma, and Chenhao Tan. "Explaining Machine Learning Classifiers through Diverse Counterfactual Explanations." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT)*, 2020, pp. 607–617.

[6] Raji, Inioluwa Deborah, and Joy Buolamwini. "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products."

Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 429–435.

[7] Selbst, Andrew D., et al. "Fairness and Abstraction in Sociotechnical Systems." *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, 2019, pp. 59–68.

[8] Barocas, Solon, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. 2019. Preprint.

[9] Angwin, Julia, et al. "Machine Bias." *ProPublica*, 23 May 2016, www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[10] Corbett-Davies, Sam, and Sharad Goel. "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." *arXiv preprint arXiv:1808.00023*, 2018.

[11] Binns, Reuben. "Fairness in Machine Learning: Lessons from Political Philosophy." *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020, pp. 149–159.

[12] Rudin, Cynthia. "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead." *Nature Machine Intelligence*, vol. 1, 2019, pp. 206–215.

[13] Sandvig, Christian, et al. "Auditing Algorithms: Research Methods for Detecting Discrimination on Internet Platforms." *Data and Discrimination: Collected Essays*, Open Technology Institute, 2014, pp. 1–23.

[14] Binns, Reuben, et al. "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions." *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018, pp. 1–14.

[15] Lipton, Zachary C. "The Mythos of Model Interpretability." *Communications of the ACM*, vol. 61, no. 10, 2018, pp. 36–43.