Archives available at journals.mriindia.com

Open Access International Journal of Science and Engineering

ISSN: 2456-3293

Volume 9 Issue 03, 2026

The People's AI Watchdog: A Citizen-Led Platform for Transparent and Accountable AI Governance

[1]Shital Namdev Zurade, [2]Dr. Ankita Karale, [3]Dr. Balkrishna K. Pati, [4]Dr. Naresh Thoutam
[1]Student, Department of Computer Engineering, SITRC, Nashik-422213, India
[2,3,4]Dr. Department of Computer Engineering, SITRC, Nashik-422213, India[2]
Email: [1]shitalnzurade@gmail.com, [2]ankita.karale@sitrc.org, [3]balkrishnapatileng@gmail.com, [4]naresh.thoutam@sitrc.org

Peer Review Information

Submission: 10 Feb 2026

Revision: 26 Feb 2026

Acceptance: 11 March 2026

Keywords

Artificial Intelligence, Citizen-Led Auditing, AI Governance, Natural Language Processing, Data Anonymization, Transparency, Bias Visualization, Dashboard Analytics, Public Sector AI, Accountability

Abstract

The rapid integration of Artificial Intelligence into public-sector decision-making has improved efficiency but has also introduced challenges related to transparency and accountability. Citizens often receive automated outcomes without clear explanations, making it difficult to understand or question these decisions. This work presents a system-focused solution in the form of a Citizen-Led AI Policy Audit Platform, designed to capture, process, and visualize real-world experiences of individuals affected by AI-driven systems.The platform allows users to submit their experiences through a simple interface, where the data is securely processed using anonymization techniques and Natural Language Processing methods. The system extracts key metadata from user narratives and stores it in a structured format for further analysis. An interactive dashboard is then used to visualize trends, patterns, and potential biases across different domains.The implementation demonstrates a complete workflow from data collection to visualization, supported by user interfaces and administrative dashboards. Through this approach, the system transforms individual experiences into meaningful insights, enabling better understanding of AI-driven decision patterns. The platform emphasizes usability, privacy, and transparency, providing a practical framework for participatory AI auditing in governance systems.

Introduction

Artificial Intelligence is increasingly used in public systems to support decision-making across multiple domains such as housing, welfare, healthcare, recruitment, and financial services. These systems are designed to process large volumes of data and produce outcomes quickly, which helps improve efficiency and scalability. However, as the use of automated systems grows, concerns related to transparency and fairness are also becoming more visible. [11]

In many real-world situations, citizens receive decisions such as approvals or rejections without any explanation. The reasoning behind these outcomes is not clearly communicated, making it difficult for users to understand why a decision was made. This lack of clarity creates a gap between system functionality and user trust. When individuals cannot question or interpret automated decisions, it leads to reduced confidence in digital governance systems. [3]

Existing approaches to AI accountability are mostly controlled by institutions. These include internal audits, compliance checks, and expert

© 2026 The Authors. Published by MRI INDIA.
This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

evaluations. While such methods are important, they often fail to capture real-world experiences of users. Many issues remain unnoticed because there is no structured way for citizens to report their experiences or contribute to the evaluation process. [6]

The need for a participatory approach becomes clear in this context. A system that allows users to share their experiences can help in identifying recurring patterns and potential issues in automated decision-making. By collecting and analyzing these experiences, it becomes possible to generate meaningful insights that reflect actual system behavior rather than theoretical assumptions.

This work introduces a Citizen-Led AI Policy Audit Platform that addresses these challenges. The system enables users to submit their experiences, processes the data using automated techniques, and presents the results through visual dashboards. The focus is on system design and implementation, ensuring that the platform remains practical, accessible, and easy to use. [9]

The proposed approach shifts the role of citizens from passive users to active contributors in the auditing process. By integrating user input with automated analysis and visualization, the system provides a structured way to understand and evaluate AI-driven decisions in real-world scenarios.

Objectives Of the System

The primary objective of the system is to design and implement a practical platform that allows citizens to report, analyze, and visualize experiences related to AI-driven decision-making in public services. The system focuses on transforming user-reported narratives into structured insights while maintaining privacy and usability.

The specific objectives of the system are as follows:

1. To provide a user-friendly interface that enables citizens to submit their experiences related to automated decisions across different domains such as housing, healthcare, education, and public services.
2. To ensure data privacy by applying automated anonymization techniques that remove sensitive and personally identifiable information from user submissions before processing.
3. To extract meaningful metadata from unstructured text using Natural Language Processing methods, including decision type, reason, location, and timeline.
4. To store processed data in a structured and normalized database, enabling efficient retrieval and analysis of large volumes of submissions.
5. To develop an analytical dashboard that visualizes patterns, trends, and potential bias in AI-driven decisions using charts and graphical representations.
6. To support cross-domain analysis by aggregating data from multiple sectors, helping identify recurring issues across different applications of AI systems.
7. To create a scalable and modular system architecture that can be extended with additional features and datasets in future implementations.
8. To shift the audit process from a system-controlled approach to a citizen-driven model, where users actively contribute to transparency and accountability.

System Overview

The proposed system is designed as an end-to-end platform that captures user experiences, processes them securely, and presents meaningful insights through an interactive interface. The overall system integrates multiple functional modules that work together to convert raw citizen input into structured and visualized data. [10]

At the entry level, the system provides a user interface where individuals can submit their experiences related to AI-driven decisions. These inputs are collected in natural language format, allowing users to describe their situations without any strict structure. This improves usability and ensures that a wide range of users can interact with the system easily.

Once the data is submitted, it moves into the processing layer. In this stage, the system performs data cleaning and anonymization to remove sensitive information. This ensures that user privacy is protected before any analysis is performed. After anonymization, Natural Language Processing techniques are applied to extract key information such as decision type, reason, location, and timeline. This step converts unstructured text into structured data that can be stored and analyzed efficiently. [5]

The structured data is then stored in a database designed to support efficient querying and retrieval. The database maintains both the anonymized narrative and extracted metadata, enabling detailed analysis across different

parameters. This organized storage plays a key role in supporting further analytical operations.

The system also includes a visualization module that presents the processed data in the form of dashboards. These dashboards display trends, patterns, and distributions using charts and graphical elements. Users and stakeholders can easily interpret the data through these visual outputs, which highlight recurring issues and potential bias in decision-making systems.

In addition, the system supports cross-domain analysis, allowing data from different sectors to be analyzed together. This helps identify common patterns across multiple applications of AI systems, providing a broader understanding of how automated decisions impact users.[15]

Overall, the system operates as a continuous pipeline where data flows from user input to visualization. Each module performs a specific function, and together they form a cohesive framework for participatory AI auditing. The design ensures that the system remains scalable, secure, and user-friendly while delivering meaningful insights.

System Architecture

The system follows a modular architecture where each component performs a clearly defined role in transforming raw user input into structured and visual insights. The architecture is designed as a pipeline, ensuring smooth data flow between modules while maintaining privacy, scalability, and analytical capability.

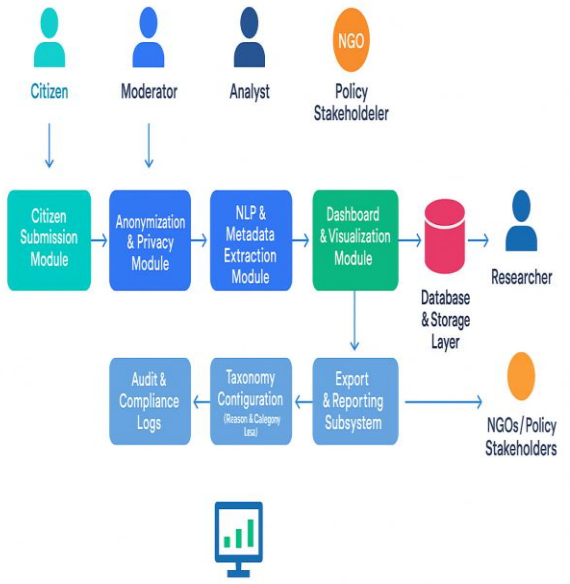

Figure 1: System Architecture

At a high level, the system consists of five major layers: Input Layer, Processing Layer, Storage Layer, Analytics Layer, and Visualization Layer. These layers interact sequentially to ensure that each stage prepares the data for the next stage without loss of information or privacy.

A. Input Layer

The input layer is responsible for collecting user submissions. It provides a user-friendly interface where citizens can enter their experiences related to AI-based decisions. The data is captured in natural language format, which allows flexibility and avoids strict input constraints. This layer ensures accessibility and encourages participation from diverse users.

B. Processing Layer

The processing layer performs the core data transformation tasks. It includes two main components:

1. Anonymization Module
   This module detects and removes personally identifiable information from the input text. It ensures that sensitive user data is not exposed during further processing or storage.
2. NLP-Based Metadata Extraction Module
   After anonymization, Natural Language Processing techniques are applied to extract structured information such as decision type, reason, location, and timeline. This step converts unstructured text into machine-readable data.

This layer is critical because it ensures both privacy protection and meaningful data extraction.

C. Storage Layer

The storage layer manages the structured data generated from the processing stage. A normalized database schema is used to store:

- Anonymized user narratives
- Extracted metadata fields

This structured storage enables efficient querying, filtering, and aggregation of data across different parameters such as domain, location, and time.

D. Analytics Layer

The analytics layer processes stored data to identify patterns and trends. It performs operations such as:

- Aggregation of similar cases
- Detection of recurring decision patterns
- Cross-domain comparisons

This layer transforms stored data into actionable insights, helping identify systemic issues and

potential biases in AI-driven systems.

E. Visualization Layer

The visualization layer presents analytical results through interactive dashboards. It includes:

- Charts showing frequency of decisions
- Maps indicating geographic distribution
- Timelines representing trends over time

This layer makes complex data understandable and accessible to both technical and non-technical users. It plays a key role in communicating insights effectively.

F. Module Interaction Flow

The architecture follows a sequential interaction model:

User Input → Anonymization → NLP Extraction → Database Storage → Analytics → Visualization

Each module passes processed data to the next stage, ensuring a continuous and reliable workflow. This structured interaction enables the system to handle large volumes of data while maintaining consistency and performance.

Overall, the system architecture ensures that data is collected securely, processed intelligently, and presented effectively. The modular design allows easy extension and maintenance, making the system adaptable for future improvements and additional functionalities.

Methodology

The system follows a structured workflow that converts user-submitted experiences into meaningful and visual insights. The methodology is designed to ensure smooth data flow across all modules while maintaining privacy, accuracy, and usability. Each step in the workflow plays a specific role in transforming raw input into structured and analyzable output.

Step-by-Step Workflow

1. User Input Submission
   The process begins when a user submits their experience through the system interface. The input is provided in natural language, allowing flexibility in describing the situation.
2. Input Validation and Preprocessing
   The system checks whether the submitted data is valid and complete. Basic cleaning is performed to remove unnecessary symbols or inconsistencies in the text.
3. Anonymization Stage
   The cleaned input is passed through an anonymization module. This module detects and removes sensitive information such as names, identifiers, or contact details. This step ensures that privacy is maintained before any further processing.
4. Metadata Extraction using NLP
   The anonymized text is processed using Natural Language Processing techniques. The system extracts key information such as:
   - Decision type
   - Reason for decision
   - Location
   - Timeline

   This step converts unstructured input into structured metadata.
5. Structured Data Formation
   The extracted metadata is combined with the anonymized narrative to form a structured record. This record is ready for storage and analysis.
6. Database Storage
   The structured record is stored in a normalized database. This allows efficient retrieval and supports large-scale data handling.
7. Data Aggregation and Analysis
   The stored data is processed to identify patterns, trends, and recurring issues. The system performs grouping and comparison across different parameters such as domain and location.
8. Visualization Generation
   The final step involves presenting the processed data through dashboards. Charts, graphs, and timelines are generated to help users understand patterns and insights easily.

Workflow Flow Representation

The overall workflow can be summarized as:

User Submission → Validation → Anonymization → NLP Extraction → Structured Storage → Analysis → Visualization

This linear pipeline ensures that each stage builds upon the output of the previous stage, maintaining consistency and reliability throughout the system.

Key Characteristics of Workflow

- Privacy-First Approach: Data is anonymized before any analysis, ensuring secure handling of user information.
- Flexible Input Handling: The system accepts natural language input, making it user-friendly.

- Automated Processing: Most steps are automated, reducing manual intervention.
- Scalability: The pipeline can handle increasing volumes of data without structural changes.
- Continuous Insight Generation: As more data is added, the system continuously updates analytical results.

Workflow Significance

This methodology ensures that the system remains practical and efficient in real-world scenarios. It bridges the gap between raw human input and structured analytical output, enabling better understanding of AI-driven decision-making systems. The workflow design supports both technical processing and user accessibility, making it suitable for large-scale deployment.

System Implementation

The implementation of the system focuses on translating the designed architecture into a functional and user-interactive platform. It includes user interfaces for data submission, processing modules for handling input, and dashboards for visualization. The system is implemented in a way that each module is clearly visible through the interface, making it easy to understand the flow of data from input to output.

User Input Interface

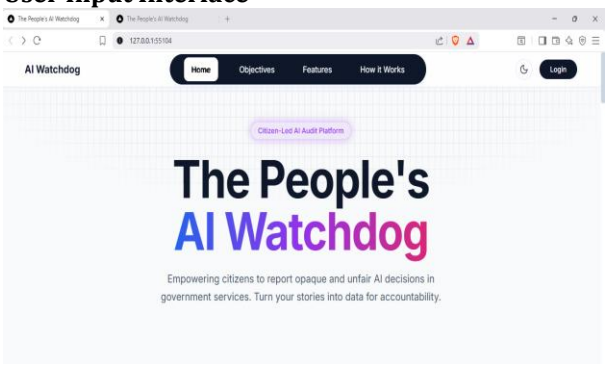

Figure 2: User Submission Interface

This screen represents the primary entry point of the system where users can submit their experiences. The interface allows users to describe their interaction with AI-driven systems in natural language. It is designed to be simple and accessible so that users from different backgrounds can easily provide input without technical knowledge. This step is important because it captures real-world data directly from affected individuals.

Input Processing and Validation

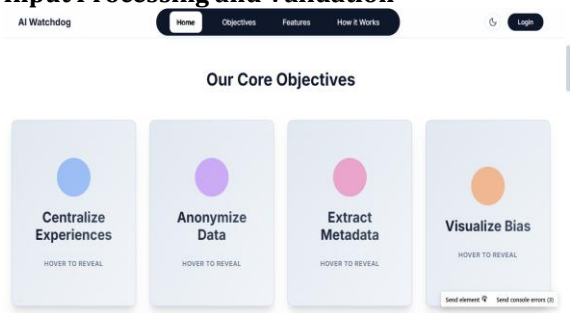

Figure 3: Input Processing and Validation Screen

This stage shows how the system handles submitted data before further processing. The interface reflects validation checks and preprocessing steps applied to the input. It ensures that only meaningful and complete data is passed to the next stage, maintaining the quality and consistency of the dataset.

Anonymization Module

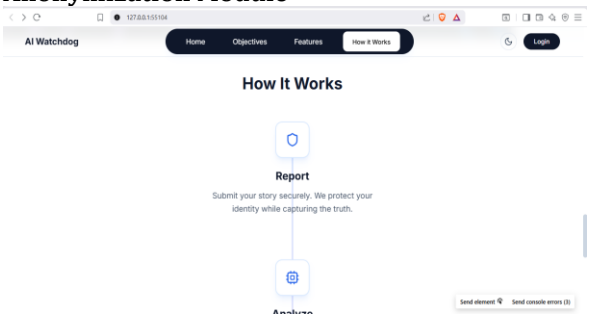

Figure 4: Anonymization Process Output

This screen demonstrates how sensitive information is removed from user submissions. Personal identifiers are automatically detected and masked, ensuring that privacy is preserved. This module is critical because it allows the system to process real user data without exposing confidential details.

Metadata Extraction Module

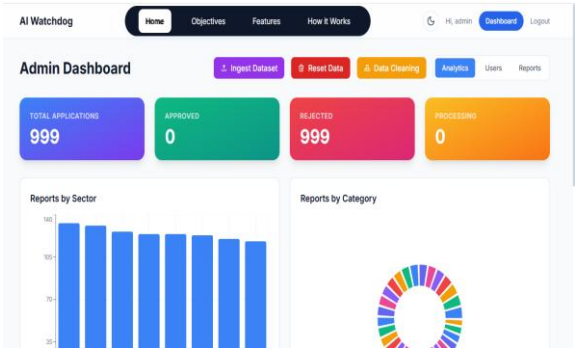

Figure 5: NLP-Based Metadata Extraction

This interface shows the output of the Natural Language Processing module. The system extracts structured information such as decision type, reason, location, and timeline from the input text. This transformation is essential because it converts unstructured narratives into structured data that can be analyzed efficiently.

Dashboard Visualization

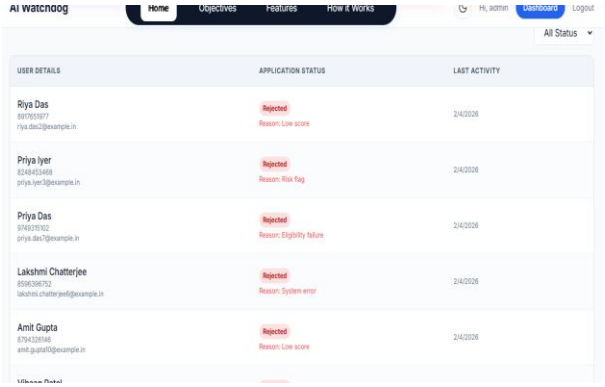

Figure 6: Analytical Dashboard Interface

The dashboard presents aggregated insights derived from processed data. It includes charts and graphical representations that highlight patterns and trends. Users can observe how decisions vary across different domains and identify recurring issues. This visualization improves interpretability and supports better understanding of system behavior.

Trend and Pattern Analysis View

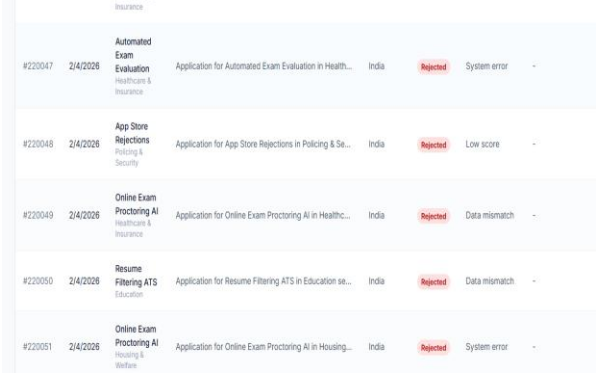

Figure 7: Trend Analysis Visualization

This screen focuses on temporal and domain-based trends. It helps in identifying how certain types of decisions occur over time or within specific regions. Such analysis is important for detecting patterns that may indicate systemic issues or bias.

System Output Summary

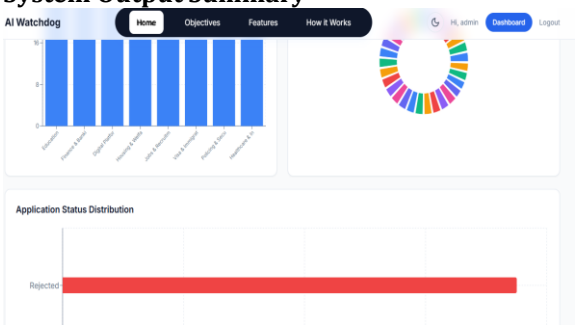

Figure 8: Final Output / Report View

This view summarizes the processed results and presents them in a structured format. It acts as a consolidated output of the system, combining extracted data and analytical insights. This final stage ensures that users and stakeholders can easily interpret the findings generated by the platform.

Implementation Insights

The implementation demonstrates a smooth integration of user interface, data processing, and visualization modules. Each component is connected in a way that ensures continuous data flow without interruption. The use of automated processing reduces manual effort, while the visual outputs make the system easy to interpret.

The inclusion of multiple interface screens highlights the practical usability of the system and shows how each module contributes to the overall functionality. This implementation validates that the proposed design can be effectively translated into a working system that supports participatory AI auditing.

Discussion

The developed system demonstrates a practical approach to addressing the challenges of transparency and accountability in AI-driven decision-making. By allowing citizens to directly contribute their experiences, the platform shifts the focus from purely institutional evaluation to a more inclusive and participatory model. This approach helps capture real-world issues that are often missed in traditional audit mechanisms.

One of the key strengths of the system is its usability. The interface is designed in a simple and intuitive manner, making it accessible to users without technical expertise. The ability to submit data in natural language removes the barrier of structured input formats, encouraging wider participation. This design choice ensures that the system can collect diverse and

meaningful data from different user groups.

The integration of anonymization plays a critical role in maintaining user trust. By automatically removing sensitive information before processing, the system ensures that privacy is preserved at all stages. This makes the platform suitable for handling real-world data where confidentiality is important. The privacy-first approach also supports ethical data handling practices, which are essential in systems dealing with citizen-generated content.

Another important aspect is the transformation of unstructured data into structured insights. The use of Natural Language Processing enables the system to extract meaningful information from user narratives without requiring predefined templates. This capability allows the platform to scale effectively while maintaining flexibility in input handling. It also ensures that valuable contextual information is not lost during processing.

The visualization module enhances the interpretability of the system. By presenting data through dashboards and graphical elements, the platform makes complex information easy to understand. Users and stakeholders can quickly identify patterns, trends, and potential issues without needing deep technical knowledge. This improves the overall effectiveness of the system in communicating insights.

From a practical perspective, the system can support multiple applications. It can be used by policymakers to understand the impact of AI systems, by researchers to study decision patterns, and by organizations to improve transparency in automated processes. The ability to perform cross-domain analysis further strengthens its usefulness by revealing patterns that extend across different sectors.

However, the system also has certain limitations. Its effectiveness depends on the quality and volume of user submissions. Incomplete or vague inputs may affect the accuracy of extracted metadata. Additionally, while the system provides insights based on collected data, it does not directly modify or control the underlying AI systems. These limitations highlight areas for future improvement.

Overall, the system offers a balanced combination of usability, privacy, and analytical capability. It provides a practical way to understand and evaluate AI-driven decisions using real-world data. By enabling citizen participation and integrating automated processing, the platform contributes toward building more transparent and accountable AI systems.

Conclusion

The presented system offers a practical and user-focused approach to improving transparency in AI-driven decision-making systems. By enabling citizens to submit their experiences and transforming those inputs into structured and visual insights, the platform creates a clear connection between real-world outcomes and analytical understanding. This helps bridge the gap between automated systems and user awareness.

The implementation demonstrates that it is possible to build a complete workflow that includes data collection, anonymization, metadata extraction, storage, and visualization within a single integrated system. Each component works together to ensure that data is handled securely while still being useful for analysis. The use of Natural Language Processing allows flexible input handling, while dashboards provide an intuitive way to interpret results.

A key contribution of this work is the shift toward a citizen-driven audit model. Instead of relying only on institutional processes, the system allows individuals to actively participate in identifying and understanding issues in AI-based decisions. This approach improves inclusiveness and helps capture real-world scenarios that may not be visible through traditional methods.

The system also demonstrates strong usability and scalability. Its modular design allows it to be extended with additional features or adapted to different application domains. The ability to perform cross-domain analysis further enhances its usefulness by providing broader insights into system behavior across multiple sectors.

In summary, the platform provides a structured and effective solution for participatory AI auditing. It combines user interaction, automated processing, and visualization to deliver meaningful insights while maintaining privacy and simplicity. This work contributes toward building more transparent, accountable, and user-aware AI systems in governance and public services.

References

J. Wihbey and D. L. McGuinness, “Public attitudes toward AI transparency in government,” Policy & Internet, 2025.

S. Huang, Y. Chen, and K. Zhang, “A framework for fairness auditing under differential privacy,” arXiv preprint arXiv:2501.12345, 2025.

M. Janssen and E. Estevez, "Trustworthy automated decision-making requirements in the public sector," Technological Forecasting & Social Change, vol. 205, p. 123456, 2025.

Center for Democracy & Technology (CDT), Spectrum of Audit Approaches: Policy & Legal Perspectives, Washington, DC, 2025.

OECD, Advancing Accountability in AI, OECD Publishing, 2023.

National Telecommunications and Information Administration (NTIA), AI Accountability Policy Report, U.S. Department of Commerce, 2024.

I. D. Raji and J. Buolamwini, "Assurance audits of algorithmic systems," in Proc. ACM Conf. on Fairness, Accountability, and Transparency (FAccT), 2024.

B. Goodman and J. Powles, Algorithmic Auditing and Audit-Washing: Risks and Standards, German Marshall Fund of the United States (GMFUS), 2022.

S. Barocas, A. D. Selbst, and M. Raghavan, "Legal, ethical, and technical survey of algorithmic auditing," in Responsible AI in Practice, Springer, 2023.

AI Now Institute, Algorithmic Impact Assessments (AIA) Framework, New York University, 2018.

S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual explanations without opening the black box: Automated decisions and the GDPR," Harvard Journal of Law & Technology, vol. 31, no. 2, pp. 841–887, 2018.

M. Veale and L. Edwards, "Transparency and accountability in AI: A cross-domain review," Frontiers in Human Dynamics, vol. 6, p. 123456, 2024.

S. Williams and A. Narayanan, "Algorithmic discrimination: Types, remedies, and legal frameworks," Policy & Society, vol. 43, no. 2, pp. 234–250, 2024.

K. Lum and W. Isaac, Predictive Policing: Bias Feedback Loops in Criminal Justice, Human Rights Data Analysis Group (HRDAG), 2019.

N. McIntyre, "UK visa algorithm bias and policy response," The Guardian / UK Government Policy Studies, 2020.

J. Drèze and R. Khera, "Aadhaar-related biometric failures and ration denials in welfare distribution," Journal of Medical Internet Research (JMIR), vol. 19, no. 5, p. e160, 2017.

OECD, Governing with AI: Public-Sector Case Studies, OECD Publishing, 2024.

V. Eubanks, "Criminal justice algorithms: Values and risks," in Automating Inequality, Springer, 2020.