Fair, Explainable, and Accountable AI: Human Oversight in Public-Sector Decision Systems

Johannes Ojanen [1]
[1]Demos Helsinki, FINLAND

ABSTRACT

Public-sector artificial intelligence can shape access to services, inspections, benefits, taxation, healthcare, mobility, and other consequential decisions. Its legitimacy therefore depends on more than predictive performance. Systems must avoid unjustified discrimination, provide explanations suitable for affected people and reviewers, assign responsibility, support challenge, and preserve meaningful human control. This review analyses fairness, explainability, accountability, and participation as interconnected dimensions of AI governance. It draws on the evidence base of a systematic synthesis that retained 95 high-quality studies and governance frameworks from 2020 to mid-2025. The source literature shows that ethics and transparency are frequently discussed, yet bias mitigation, accountability mechanisms, citizen participation, and implementation procedures remain uneven. The article examines how bias enters through problem definition, data, labels, modelling, thresholds, deployment, and institutional practice. It distinguishes transparency from useful explanation and describes accountability as a chain of ownership, documentation, review, appeal, and correction. It also considers the limitations of nominal human-in-the-loop arrangements and proposes conditions for effective oversight. A practical assurance model is presented that connects principles to controls, metrics, and evidence. The review concludes that fair and accountable AI requires institutional capacity, stakeholder participation, and continuous monitoring, not a one-time technical test or public statement.

Keywords: Algorithmic Fairness; Explainable AI; Accountability; Human Oversight; Public-Sector AI; Bias Governance; Participation

Introduction

Artificial intelligence used by public institutions carries a special burden of legitimacy. Citizens may have little choice about interacting with a tax authority, health service, transport system, benefits agency, regulator, or local government. When an AI system influences these services, people need confidence that decisions are fair, understandable, reviewable, and subject to responsible human authority.

Many governance frameworks acknowledges these values. Fairness, transparency, and accountability appear repeatedly in policy documents and academic literature. Yet implementation often remains weak. Fairness may be reduced to one metric, even though different metrics can conflict. Transparency may be interpreted as publishing technical information that ordinary users cannot understand. Accountability may be declared without identifying who has authority to correct errors. Human oversight may exist formally while reviewers lack time, training, evidence, or power.

These shortcomings are connected. Bias cannot be addressed without documentation and data. Explanations cannot support accountability if there is no appeal process. Human reviewers cannot intervene if decisions are made too quickly or the interface hides uncertainty. Participation cannot improve design if it occurs only after deployment. This

review treats fairness, explainability, accountability, human oversight, and participation as a single governance system.

Review Scope and Context

The article is based on the supplied systematic review of AI governance and cybersecurity frameworks. The underlying study retained 95 high-quality sources and mapped governance themes across public administration, smart cities, infrastructure, and related domains. Its findings showed that ethics and accountability or transparency were among the most visible themes, while citizen participation and stakeholder engagement remained peripheral. Framework clustering also revealed a divide between human-centred approaches and technical or compliance-centred approaches.

The present review reinterprets this evidence for public-sector decision systems. It does not present a new systematic search. Instead, it provides a complete thematic analysis of how fairness and accountability can be operationalised throughout the AI lifecycle. Particular attention is given to decisions that affect rights, opportunities, access, or public trust.

How Bias Enters AI Systems

Bias can begin before any data are collected. Problem-definition bias occurs when an institution frames a social issue in a way that makes some outcomes visible and others invisible. A model designed to predict non-compliance may reflect enforcement patterns rather than underlying behaviour. A system built to maximise efficiency may ignore unequal burdens placed on vulnerable groups.

Data bias arises when training data do not represent the population or context of use. Historical data may contain discrimination, under-service, or selective recording. Some groups may appear less often or be measured less accurately. Missingness itself can reflect unequal access. Using more data does not necessarily solve the problem if the data reproduce the same structural patterns.

Label bias occurs when the target variable is an imperfect or unfair proxy. Arrest, complaint, expenditure, or previous service use may be influenced by institutional practice rather than true need or risk. Proxy variables can also reproduce sensitive attributes. Location, language, employment history, or purchasing patterns may function as indirect indicators of race, class, disability, or migration status.

Model and threshold choices affect how errors are distributed. A single decision threshold may create different false-positive or false-negative rates across groups. Optimising overall accuracy can hide poor performance for minorities. Deployment bias appears when a model developed for one setting is used in another, or when staff interpret recommendations in inconsistent ways. Feedback loops may then reinforce earlier decisions, creating the appearance that the model was correct.

Bias governance must therefore extend beyond technical debiasing. It should examine the purpose, institutional history, data generation process, model design, workflow, and consequences.

Fairness as a Governance Process

Fairness cannot be reduced to a universal formula. Statistical parity, equal opportunity, equalised error rates, calibration, and individual fairness express different ideas. In many settings, they cannot all be satisfied at once. The appropriate approach depends on the decision, the harm being prevented, legal duties, affected groups, and the consequences of different errors.

A fair-governance process begins by identifying stakeholders and protected or vulnerable groups. It defines the decision context, likely benefits, and possible harms. It then selects metrics that reflect those concerns and explains why they are suitable. Results should be examined across relevant subgroups and intersections rather than only broad categories.

Thresholds for unacceptable disparity should be agreed before testing where possible. If a threshold is exceeded, the organisation needs a remediation workflow. Options may include improving data, changing features, adjusting thresholds, redesigning the decision process, limiting use, or stopping deployment. Each action can create trade-offs and should be documented.

Monitoring is essential because fairness can change after deployment. Population patterns, service access, data collection, and institutional behaviour may shift. A model that initially performs acceptably may later produce unequal outcomes. Complaint patterns and qualitative feedback can reveal harms that statistical dashboards miss.

Transparency and Explainability

Transparency is a broad condition of openness, while explainability is the provision of understandable reasons for a system's operation or a specific outcome. Publishing model architecture or source code may increase technical transparency but does not automatically help an affected person understand a decision. Useful explanation is audience-specific.

Citizens generally need clear information about whether AI was used, what role it played, which factors were important, what limitations exist, and how to request review. Operators need information about confidence, uncertainty, data quality, and conditions where the model may fail. Auditors require detailed documentation, test results, lineage, thresholds, change history, and access to relevant records. Regulators may need evidence of compliance and risk management.

Explanations should not create false certainty. Many AI outputs are probabilistic, and post-hoc explanation methods may only approximate internal reasoning. Governance should require teams to state the limits of an explanation. It should also test whether users understand the information and whether explanations help them challenge errors.

Transparency must be balanced with privacy, security, and legitimate confidentiality. Detailed disclosure can expose personal information or make a system easier to attack. Layered access provides a practical compromise. Public explanations remain understandable, while sensitive technical material is available to authorised reviewers under controlled conditions.

Accountability and Meaningful Human Oversight

Accountability means that responsibility can be traced and exercised. It includes ownership before deployment, supervision during operation, and correction after harm. A responsibility matrix should identify the senior accountable owner, system owner, model developer, data steward, privacy and security roles, operational users, and independent reviewers. Vendor involvement should be explicit because outsourcing does not remove public responsibility.

Approval gates should match risk. High-impact systems need documented review of purpose, data, fairness, security, privacy, and human oversight. Decisions to accept residual risk should be made by people with appropriate authority. Records should show the evidence considered and any conditions attached to approval.

Human oversight is meaningful only when the reviewer can understand the system, recognise uncertainty, challenge the output, and take a different action without penalty. A human who merely confirms recommendations under severe time pressure provides little protection. Oversight design should address workload, training, interface quality, escalation, and organisational culture.

Different forms of oversight may be appropriate. Human-in-the-loop arrangements require approval for individual decisions. Human-on-the-loop models allow automated operation but maintain active supervision and intervention. Human-in-command governance gives people authority over objectives, limits, and shutdown decisions. Consequential public decisions may require more than one form.

Appeal and correction mechanisms complete accountability. People should know how to contest an outcome, present additional information, and obtain human review. Institutions need procedures for correcting records, reversing decisions, and examining whether similar cases were affected.

Governance dimension	Key control questions	Possible metrics	Evidence
Fairness	Which groups may be disadvantaged? Which errors matter most? What threshold is acceptable?	Error rates by group, selection rates, calibration, complaint patterns	Fairness assessment, metric rationale, remediation record
Explainability	What must each audience understand? Are limitations stated?	User comprehension, explanation consistency, review success	Model card, explanation samples, usability test
Accountability	Who owns the system and accepts residual risk? Who can stop it?	Review completion, unresolved findings, response time	Responsibility matrix, approvals, audit trail

Governance dimension	Key control questions	Possible metrics	Evidence
Human oversight	Can reviewers recognise and override unsafe outputs?	Override rate, review time, escalation rate, outcome quality	Training record, interface test, override log
Participation	Were affected groups involved before and after deployment?	Representation, issues raised, changes adopted, satisfaction	Consultation report, design change record, complaint register

Participation and Public Trust

The source literature places citizen participation and stakeholder engagement at the edge of mainstream governance discussions. This is a significant weakness. Public-sector AI affects people with different experiences, resources, and exposure to harm. Technical and legal experts may not anticipate how a system changes everyday interaction with a service.

Participation should begin during problem definition. Communities can question whether automation is appropriate, identify harms, and propose alternatives. Domain professionals can explain operational realities that datasets do not capture. Frontline staff can reveal where a model may create new burdens or workarounds. Consultation should continue during testing and after deployment.

Participation must be designed carefully. A single public meeting or survey does not create democratic legitimacy. Institutions should explain what can change, provide accessible information, include under-represented groups, and report how feedback influenced decisions. Where recommendations are not adopted, reasons should be given.

Trust should be treated as an outcome of trustworthy practice rather than a communication objective. Public confidence grows when institutions are honest about limitations, respond to complaints, publish meaningful evidence, and correct mistakes. Efforts to persuade people to trust an unaccountable system are unlikely to succeed.

Assurance, Limitations, and Future Work

The principles-controls-evidence model offers a practical assurance structure. Fairness principles lead to subgroup testing, metric selection, and remediation controls. Transparency principles lead to documentation and audience-specific explanation. Accountability principles lead to named owners, review gates, appeals, and incident procedures. Evidence shows whether these controls were used and whether they changed outcomes.

The literature remains limited by uneven empirical validation. Many frameworks describe desirable practices but do not report long-term effects on discrimination, trust, appeals, or

institutional behaviour. Metrics may be tested on static datasets without studying how staff and citizens interact with the system. More field research is needed.

Future studies should compare different oversight designs, measure the quality of human review, and examine whether explanations improve challenge and correction. Research should also develop methods for fairness monitoring when demographic data are incomplete or legally restricted. Participatory governance needs stronger evaluation, including which methods genuinely influence design and which merely create an appearance of consultation.

Conclusion

Fair, explainable, and accountable AI requires more than a technically accurate model. Bias can enter through objectives, data, labels, thresholds, workflows, and institutional history. Explanation must be designed for real audiences. Accountability needs named owners, evidence, challenge routes, and correction. Human oversight must provide genuine authority rather than ceremonial approval. The reviewed governance landscape gives strong rhetorical attention to ethics and transparency but less consistent attention to bias controls, participation, and operational accountability. A lifecycle approach can correct this imbalance. It combines early stakeholder involvement, context-specific fairness testing, layered explanation, meaningful human review, appeals, monitoring, and auditable evidence. Public institutions earn trust by demonstrating responsible practice, acknowledging uncertainty, and repairing harm. When fairness and accountability are built into the structure of decision-making, AI can support public services without weakening rights, dignity, or democratic legitimacy.

References

[1] Lwakatare, L. E., Crnkovic, I., & Bosch, J. (2020, September). DevOps for AI–Challenges in Development of AI-enabled Applications. In *2020 international conference on software, telecommunications and computer networks (SoftCOM)* (pp. 1-6). IEEE.

[2] Mallempati, A., & Jaladi, D. S. (2021). A Scalable Master Data Management Architecture for Enterprise Data Integration and Governance in Full-Stack Application Environments. *International Journal of Emerging Research in Engineering and Technology*, *2*(1), 101-110.

[3] Kuntamukkala, N. K., & Thalary, S. (2021). Self-Optimizing Angular Applications: A Novel Framework for AI-Driven Performance Adaptation in Production Environments. *International Journal of AI, BigData, Computational and Management Studies*, *2*(2), 107-117.

[4] Aluri, Y. S. (2021). Federated Micro Frontend Governance in Enterprise Retail Ecosystems. *International Journal of Artificial Intelligence, Data Science, and Machine Learning*, *2*(2), 114-125.

[5] Subashree, M. (2020). Agile Software Development in AI-Driven Applications: Challenges and Strategies. *International Journal of Emerging Trends in Computer Science and Information Technology*, *1*(3), 10-16.

[6] Pillai, S. (2016). Continuous Integration/Continuous Deployment (CI/CD) in DevOps: Principles, Practices, andChallenges. *International Journal of Artificial Intelligence and Machine Learning*, *6*(3).

[7] Yuvaraj, N., & Kumar, M. S. (2021). From Governed Data to Customer Health Signals: Integrating Telemetry with Enterprise Data Quality Controls. *International Journal of Emerging Trends in Computer Science and Information Technology*, *2*(4), 115-125.

[8] Erik, S., & Emma, L. (2018). The Future of Software Development: AI-Driven Testing and Continuous Integration for Enhanced Reliability. *International Journal of Trend in Scientific Research and Development*, *2*(4), 3082-3096.

[9] Kumar, M. S., & Yuvaraj, N. (2020). Building a Privacy-Aware Customer Data Foundation: A Governance-First Approach to Digital Service Systems. *International Journal of Emerging Research in Engineering and Technology*, *1*(4), 55-68.