Received: 12 November 2024 | Accepted: 03 April 2025 | Published: 14 June 2025

ARK: 24285/RCC.v9i17.254

ARTICLE

REVISTA COLETA CIENTÍFICA

Governing Artificial Intelligence in the Public Sector: Accountability, Transparency, and Risk-Based Oversight

Governança da inteligência artificial no setor público: responsabilização, transparência e supervisão baseada em riscos

Amy West West
https://orcid.org/0009-0002-5111-4481
University of Michigan: Ann Arbor, Michigan, US
E-mail: umich@edu.com

Publication information

ARK: 24285/RCC.v9i17.254

ISSN: 2763-6496

Keywords:
Artificial intelligence.
Public administration.
Accountability.
Transparency.
Risk management.

Palavras-chave:
Inteligência artificial.
Administração pública.
Responsabilização.
Transparência.
Gestão de riscos.

Abstract
Public agencies use artificial intelligence to expand analytical capacity, personalize services, and automate work, redistributing decisions, risks, and responsibilities across sociotechnical arrangements that can be difficult to inspect. This conceptual integrative review connects public-administration scholarship with international governance frameworks. The synthesis identifies six requirements: a clear public purpose and legal basis; proportionate risk classification; lifecycle data governance and documentation; human oversight with actual authority; audience-specific transparency; and accessible mechanisms for contestation, audit, and remedy. High-level principles have limited operational value unless converted into named roles, decision gates, and reviewable evidence. The article therefore proposes a minimum governance architecture organized around ex ante, concurrent, and ex post controls for internally developed systems and procured services. It emphasizes traceability, responsible procurement, continuous monitoring, and the capacity to suspend a system when harms or performance failures emerge. Explainability is one component of accountability, not a substitute for legality, due process, or institutional responsibility. Legitimate public-sector AI depends on the organization's ability to justify, review, contest, and discontinue automated practices.

Resumo

A incorporação de inteligência artificial por órgãos públicos pode ampliar capacidade analítica, personalizar serviços e reduzir tarefas repetitivas, mas também desloca decisões, riscos e responsabilidades para arranjos sociotécnicos difíceis de auditar. Este artigo desenvolve uma revisão conceitual integrativa sobre governança de IA no setor público, articulando literatura de administração pública com referenciais normativos internacionais. A síntese identifica seis requisitos interdependentes: finalidade pública e base jurídica claras; classificação proporcional de risco; governança de dados e documentação do ciclo de vida; supervisão humana com autoridade real; transparência orientada aos públicos afetados; e mecanismos de contestação, auditoria e reparação. Argumenta-se que listas genéricas de princípios têm baixo valor operacional quando não são traduzidas em papéis, evidências e pontos de decisão. Como contribuição, propõe-se uma arquitetura mínima de governança organizada em controles ex ante, concomitantes e ex post, aplicável tanto a sistemas desenvolvidos internamente quanto a soluções adquiridas de fornecedores. O modelo enfatiza rastreabilidade, contratação responsável e monitoramento contínuo, sem tratar explicabilidade como substituto de legalidade ou devido processo. Conclui-se que a legitimidade da IA pública depende menos da sofisticação técnica isolada e mais da capacidade institucional de justificar, revisar e interromper seu uso.

1. Introduction

Artificial intelligence (AI) is increasingly embedded in public-sector activities such as document classification, fraud detection, resource allocation, inspection prioritization, and citizen-facing assistance. The attraction is understandable: governments face rising service expectations, fragmented information, fiscal constraints, and backlogs that appear amenable to computational support. Yet public authority changes the normative stakes. An erroneous product recommendation can inconvenience a consumer; an erroneous public decision may delay a benefit, intensify surveillance, or impair access to rights. Consequently, the relevant question is not simply whether an AI system works, but whether its use is lawful, justified, reviewable, and compatible with democratic values.

Research on AI in government shows a field rich in exploratory applications but still short of explanatory and longitudinal evidence about institutional effects (Wirtz et al., 2019; Zuiderwijk et al., 2021). This gap matters because the same model can produce different outcomes depending on procurement terms, data quality, worker discretion, administrative procedures, and avenues for appeal. A governance approach must therefore examine the complete decision system: policy goals, organizational routines, datasets, models, interfaces, human actors, vendors, and affected communities.

This article asks: Which institutional controls are necessary to make public-sector AI accountable, transparent, and proportionate to risk? Its contribution is a practical synthesis that translates established principles into a minimum operating architecture. The proposal is not a universal compliance checklist. It is a set of decision rights, evidence requirements, and review mechanisms that public organizations can adapt to sectoral law and local administrative capacity.

2. Methodological approach

The manuscript is a conceptual integrative review. It combines peer-reviewed syntheses on AI and public governance with authoritative frameworks from the OECD, UNESCO, NIST, and the European Union. Sources were selected for their direct relevance to public accountability, lifecycle risk management, transparency, and administrative decision-making. The analysis compared the problem definitions, governance functions, and evidence artifacts proposed across these sources, then organized recurring elements by when oversight occurs: before deployment, during operation, and after decisions or harms.

The review does not claim systematic exhaustiveness or generate a pooled effect estimate. Its purpose is theory building and institutional design. Before submission,

future human authors should reproduce and document searches in multidisciplinary databases, define inclusion criteria, check for relevant national rules, and validate whether the proposed controls fit the jurisdiction and public service under study. This limitation is deliberate: it avoids presenting an AI-assisted initial synthesis as a completed systematic review.

3. From ethical principles to an operating system

International instruments converge around human rights, fairness, transparency, robustness, safety, and accountability. The OECD AI Principles connect trustworthy AI to human-centred values and systematic risk management; UNESCO frames ethical impact in relation to dignity, inclusion, and environmental and social consequences; and the NIST AI Risk Management Framework organizes practice around the functions govern, map, measure, and manage (NIST, 2023; OECD, 2024; UNESCO, 2021). The European Union's AI Act adds legally binding, risk-based obligations for systems within its scope (European Union, 2024). Although these instruments differ in legal status and geography, they share a lifecycle perspective.

The implementation problem is that values do not assign work. An agency may endorse fairness without deciding who approves a training dataset, who measures disparate error rates, what threshold triggers remediation, or who can halt deployment. Accountability requires a chain linking a public purpose to a responsible office, a documented design choice, an observable outcome, and a remedy. Busuioc (2021) accordingly emphasizes that accountability cannot be delegated to an algorithm: public organizations must remain answerable for the authority exercised through technical systems.

A useful governance architecture therefore resembles an operating system for decisions. It specifies roles, permissions, mandatory evidence, escalation paths, and review intervals. The architecture must include political and administrative leadership, legal and procurement functions, data and security specialists, frontline professionals, internal audit, and representatives of affected groups. Technical teams can explain model behavior, but only duly authorized public actors can justify why that behavior is acceptable in a particular administrative context.

4. A minimum risk-based governance architecture

Risk classification is the first control because it determines the intensity of every other control. Relevant dimensions include the significance of the decision, the vulnerability of affected persons, scale, reversibility, data sensitivity, automation level, and the possibility of cumulative or discriminatory harm. A low-impact writing assistant used by staff should not be governed identically to a system that ranks families for inspection. Proportionality, however, should not become a loophole: classifications and exemptions must be reasoned, documented, and periodically revisited as use changes.

Before deployment, agencies should establish purpose, legal authority, necessity, and a credible non-AI alternative. They should assess data provenance and fitness, define expected benefits and unacceptable harms, consult affected stakeholders, and set measurable acceptance criteria. For high-impact uses, an algorithmic or human-rights impact assessment should be reviewed by a body independent from the project team. Documentation must also identify residual risks and the official empowered to accept them.

During operation, oversight requires more than a nominal human in the loop. Human reviewers need time, competence, contextual information, and authority to

disagree with the system. Monitoring should track technical performance, distributional effects, overrides, complaints, drift, incidents, and changes in the surrounding policy environment. Logs must support investigation without creating disproportionate surveillance. The NIST lifecycle approach is particularly useful because it treats governance as cross-cutting rather than as a one-time approval (NIST, 2023).

After decisions and incidents, affected persons need intelligible notice, an accessible route to contest the outcome, and review by someone capable of providing an effective remedy. Kroll et al. (2017) show why accountability cannot rely only on revealing source code: procedural safeguards, testing, and institutional powers are often more meaningful. Ex post controls should include incident reporting, root-cause analysis, remediation deadlines, independent audit, and, when warranted, suspension or retirement.

Table 1. Minimum evidence architecture for public-sector AI

Governance dimension	Minimum control	Reviewable evidence	Trigger for escalation
Purpose and legality	Document public purpose, authority, necessity, and alternatives	Decision memorandum; legal analysis; service map	Purpose expansion or change in law
Data and model	Assess provenance, fitness, performance, bias, security, and accessibility	Datasheets; test plan; validation report; threat model	Drift, data-source change, or subgroup disparity
Human oversight	Define reviewer competence, authority, workload, and override path	Operating procedure; training record; override logs	Automation bias or ineffective review
Transparency	Provide audience-specific notice and publish system-level information	Public register; notices; model/service documentation	Material update or misleading disclosure
Contestability and remedy	Offer accessible review, correction, and redress	Appeal protocol; response times; outcomes	Complaint pattern or unresolved harm
Lifecycle assurance	Monitor, audit, report incidents, and retire safely	Monitoring dashboard; audit report; incident and exit plans	Threshold breach, severe incident, or loss of vendor support

Source: initial synthesis for subsequent human validation.

5. Transparency, procurement, and institutional capacity

Transparency has several audiences and purposes. Leaders need risk and performance information; auditors need access to evidence and logs; frontline workers need limitations and escalation rules; and affected people need to know when AI materially influences a process, what information was relevant, and how to seek review. A single technical model card cannot satisfy all these needs. Proactive public registers can improve discoverability, but their fields should connect to deeper documentation rather than encourage performative disclosure.

Procurement is a decisive governance point because proprietary systems can restrict inspection, data portability, and remedy. Contracts should allocate responsibilities for data protection, security, accessibility, performance testing, incident notice, subcontractors, model updates, audit access, and orderly exit. Agencies should avoid clauses that prevent meaningful explanation or bind essential public functions to a vendor without migration options. Data governance is equally central: unreliable, historically biased, or poorly documented data can undermine even a technically sound model (Janssen et al., 2020).

Capacity constraints are not solved by adopting more guidance. Smaller agencies may need shared assessment services, model clauses, sectoral assurance labs, and pooled technical expertise. Training must include not only data science but also administrative

law, records management, procurement, public participation, and service design. The goal is institutional competence to ask the right questions and act on the answers, including the decision not to automate.

6. Implementation roadmap and research agenda

Implementation can begin with an inventory of AI-enabled systems and a moratorium on undisclosed high-impact use. Agencies can then classify risk, assign accountable owners, standardize impact assessments, establish review gates, and publish a public-facing register. Mature programs should connect governance evidence to budgeting, procurement, cybersecurity, privacy, records, and internal audit rather than create a parallel compliance bureaucracy.

Research should move beyond catalogs of principles toward comparative studies of how controls work in practice. Priority questions include whether impact assessments alter deployment decisions; how different appeal mechanisms affect error correction; how procurement terms shape transparency; which metrics reveal distributional harms; and how frontline discretion changes after automation. Longitudinal and cross-jurisdictional designs are needed to identify when governance reduces harm without freezing beneficial experimentation.

7. Final considerations

Public-sector AI is legitimate only when public institutions remain capable of explaining and revising the exercise of authority. Risk-based governance should not mean less accountability for lower-risk systems; it means proportionate depth of evidence, scrutiny, and monitoring. The minimum architecture proposed here aligns purpose and legality, data and technical assurance, meaningful human control, differentiated transparency, contestability, audit, and remedy across the lifecycle.

The practical test is simple but demanding: for every consequential system, an agency should be able to state why it is used, who is responsible, what evidence supports it, who may be harmed, how that harm is detected, how a person can challenge an outcome, and who can stop the system. Where those answers are absent, deployment is not merely immature; it is institutionally unjustified.

References

Busuioc, M. (2021). Accountable artificial intelligence: Holding algorithms to account. Public Administration Review, 81(5), 825–836. https://doi.org/10.1111/puar.13293

European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence. Official Journal of the European Union. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

Janssen, M., Brous, P., Estevez, E., Barbosa, L. S., & Janowski, T. (2020). Data governance: Organizing data for trustworthy artificial intelligence. Government Information Quarterly, 37(3), 101493. https://doi.org/10.1016/j.giq.2020.101493

Kroll, J. A., Huey, J., Barocas, S., Felten, E. W., Reidenberg, J. R., Robinson, D. G., & Yu, H. (2017). Accountable algorithms. University of Pennsylvania Law Review, 165(3), 633–705. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3/

National Institute of Standards and Technology. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0) (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

OECD. (2024). Governing with artificial intelligence: Are governments ready? OECD Publishing. https://www.oecd.org/content/dam/oecd/en/publications/reports/2024/06/governing-with-artificial-intelligence_f0e316f5/26324bc2-en.pdf

OECD. (2024). Recommendation of the Council on Artificial Intelligence (OECD/LEGAL/0449, amended 2024). https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449

UNESCO. (2021). Recommendation on the ethics of artificial intelligence. https://unesdoc.unesco.org/ark:/48223/pf0000381137

Wirtz, B. W., Weyerer, J. C., & Geyer, C. (2019). Artificial intelligence and the public sector—Applications and challenges. International Journal of Public Administration, 42(7), 596–615. https://doi.org/10.1080/01900692.2018.1498103

Zuiderwijk, A., Chen, Y.-C., & Salem, F. (2021). Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. Government Information Quarterly, 38(3), 101577. https://doi.org/10.1016/j.giq.2021.101577