ISSN Online: 3006-4708
ISSN Print: 3006-4694
SOCIAL SCIENCE REVIEW ARCHIVES
https://policyjournalofms.com

Regulating Artificial Intelligence in the Public Sector: Policy Frameworks for Accountability, Ethics, and Human Rights Protection

Aqsa Malik[1]
[1] BS Public Policy Scholar, Department of Public Policy and Governance, The Superior University, Lahore.
Email: aqsamalikaqsamalik09@gmail.com

DOI: https://doi.org/10.70670/sra.v4i2.2263

Abstract
The use of Artificial Intelligence (AI) in government institutions has changed the way government functions by improving its efficiency, decision making and services. But with growing dependence on AI systems, there have been serious concerns about accountability, ethical governance, and the safeguarding of fundamental human rights. The purpose of this study is to investigate policy frameworks that govern the use of AI in the public sector to discuss accountability mechanisms, examine the integration of ethics considerations, and determine human rights protection that is required for responsible AI governance. The study is informed by a rights-based approach to AI governance, which is based on theories in the broader area of accountability, ethical governance and human rights protection, to evaluate current governance frameworks. An Interpretivist research paradigm was employed and a qualitative research approach. Data collected through qualitative document analysis of 40 policy and regulatory documents (international conventions, AI regulations, governance guidelines, and national AI strategies), and 15 semi-structured interviews with policymakers, academics and civil society representatives. There were purposive sampling and snowball sampling techniques used in the selection of relevant documents and participants. Data collected were analysed using the six phases of the thematic analysis framework developed by Braun and Clarke with the help of the NVivo software. The results show that transparency, explainability, human supervision and independent auditing are the main mechanisms of accountability in current governance frameworks for AI. Fairness, transparency and non-discrimination were the most important ethical principles yet there are differences between the policy commitments and implementation. In addition, it was determined that privacy protection, equality safeguards, appeal mechanisms and independent oversight institutions are key human rights safeguards. Based on the findings of the study, it suggests that robust regulatory frameworks are essential to promote a society-centred and responsible use of AI in public services, incorporating accountability, ethics and human rights protections.

Keywords: Artificial Intelligence, Public Sector Governance, Accountability, Ethical AI, Human Rights Protection, Transparency, Explainability, AI Regulation

Introduction

Artificial Intelligence (AI) technologies are rapidly penetrating the public administration and are changing the way governments provide services, allocate resources, and make decisions that impact citizens' lives. AI algorithms are used in public institutions to enhance processes and services, ranging from automated welfare eligibility systems, predictive policing tools, to AI-powered healthcare triage, tax assessment, and immigration screening (WaTech & UC Berkeley, 2025). However, this quick pace of adoption has been

faster than the development of a sufficient oversight framework, leaving significant questions of accountability, ethical governance, and fundamental human rights protections as it relates to algorithmic systems that are incorporated into state functions (Congressional Research Service, 2025).

In response, a number of jurisdictions and international organizations have started to develop comprehensive frameworks of policy regarding AI in Government. The European Union introduced a risk-based approach to AI regulation in the Artificial Intelligence Act (2024/1689) which applies strict requirements to high-risk AI systems in educational, employment, law enforcement and welfare administration sectors (European Parliament and Council, 2024). The Council of Europe's Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law, the first binding international agreement on AI governance, is also an instrument that explicitly places obligations on AI governance in the framework of human rights law, namely in terms of legality, necessity, proportionality, transparency and non-discrimination (Council of Europe, 2024). The OECD has suggested a governance framework of enablers, guardrails and citizen engagement at the multilateral level to facilitate governments in their responsible use of AI in public services (OECD, 2025).

Such efforts have not yet succeeded in reducing the extent of accountability deficiencies. Citizens face challenges in grasping, challenging, and seeking remedy from algorithmic systems operated by public agencies that are "black boxes" (Pavlidis, 2024). Explainability and transparency are increasingly being posited as pre-conditions for legitimate uses of AI in governance, as they are essential to maintaining public confidence in algorithmic governance without systems that can be understood and without clear institutional accountability (Papadakis et al., 2024). Cheong (2024) also maintains that transparency and accountability mechanisms are crucial protections for the wellbeing of individuals in the context of algorithmic decision-making, and Gaudeul et al. (2024) demonstrate that effective human oversight can significantly mitigate discrimination in decisions made by AI tools related to welfare provision, lending, and employment.

The ethical and human rights aspects are of the same concern. AI is a form of "slow violence" to the human rights framework, chipping away little by little at people's ability to comprehend or demand redress for algorithmic injuries and at the moral core of privacy, non-discrimination, freedom of expression, says Teo (2025). Amnesty International (2024) has raised the alarm about the impact of AI on fundamental rights in healthcare access, social assistance, and migration monitoring, urging the establishment of legally binding, rights-based regulations. This concern was reiterated by the adoption of a resolution by member states of the United Nations in 2024 affirming that human rights have to be respected per the entire AI lifecycle (United Nations General Assembly, 2024). The U.S. Department of State (2024) has also recently associated the use of AI with human rights responsibilities in a specific risk management profile, and Sahebi and Formosa (2025) have placed these debates in a wider context of global justice and unequal power dynamics between states, corporations, and citizens. Leslie and Perini (2024) also claim that the advent of generative AI has exacerbated an international governance crisis, revealing the inadequacy of current policy tools.

These trends collectively suggest a diverse and uneven policy environment: normative statements about accountability, ethics and human rights protection are increasingly appearing in international policy documents, but implementation of these concepts in specific, enforceable public sector practice varies. It is crucial to understand these dynamics in countries where the use of AI in the public sector continues without the support of institutions, expertise or independent oversight bodies to protect against the risks associated with it. The purpose of this study is to draw on the latest legislative, regulatory, and scholarly advances and literature to enrich the existing panoply of literature on responsible AI governance, and to guide policymakers in finding the balance between innovation and safeguarding citizens' rights and democratic accountability. It thus discusses the creation and implementation of policy frameworks for the adoption of AI in public administration, which should uphold accountability, incorporate fundamental ethical principles and protect fundamental human rights.

Research Questions

1. How public sector institutions can be held accountable for AI governance mechanisms that are already in place?
2. How well do public sector AI policies consider the fundamental ethical principles like fairness, transparency, and non-discrimination?
3. Which human rights are required to ensure that citizens are protected from the harm that may occur due to the use of AI in public administration?

Research Objectives

1. To explore the accountability mechanisms in existing AI governance frameworks that are relevant to public sector institutions.
2. To assess how public sector AI policies incorporate fundamental ethical values like non-discrimination, transparency, and fairness.
3. To determine the human rights protection that is needed, to ensure that citizens are free from dangers caused by the use of AI in public administration.

Literature Review

The use of Artificial Intelligence (AI) in public administration has been the subject of an increasing volume of scholarship, embodying both its potential for transformation and the governance issues it raises. To shed light on the ways in which actors, institutions and policies interact as AI is integrated into government processes, Criado, Sandoval-Almazán and Gil-Garcia (2025) articulate a multi-level framework, which includes micro, meso and macro perspectives, and note that a lack of coordination in governance can leave gaps in responsibility between layers of government. Similarly, Janssen, Mellouli and Ojo (2024) point to AI's potential for significant efficiency and service but also its potential to cause bias, opacity and loss of public trust, for which existing administrative frameworks are poorly suited. A WaTech/UC Berkeley (2025) empirical study of AI governance practices in Washington State highlights these tensions, revealing that agencies often adopt AI tools before they develop the necessary oversight capacity. Together, this body of literature highlights that there are significant gaps between the institutional capacity to govern AI and its incorporation into the government.

From the regulatory perspective, the European Union Artificial Intelligence Act (AI Act) (European Parliament and Council, 2024), which was adopted as Regulation (EU) 2024/1689, has been a key reference point, laying out a three-tiered, risk-based framework of obligations for AI systems, especially in the fields of law enforcement, education and welfare administration. This principle also applies on an international scale, as highlighted by the Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law by the Council of Europe, which is the first binding treaty mandating that AI systems deployed by public bodies respect recognised human rights norms (Council of Europe, 2024). The OECD's (2025) Governing with Artificial Intelligence in Government report also introduces a series of enablers, guardrails, and citizen engagement to operationalize these commitments, while the Congressional Research Service (2025) provides a more disorganized picture in the United States, with overlapping state and federal initiatives generating regulatory uncertainty for public agencies. These are all examples of the instruments that reflect a transition from a voluntary approach to ethical principles towards legally enforced obligations for AI in government.

There has been a significant body of writing on accountability mechanisms for AI use in government. Yuan and Chen (2025) performed a systematic review of accountability mechanisms used to hold public sector AI systems accountable, all of which were procedural (or rules-based) mechanisms that provided very little redress for individuals who had suffered from algorithmic decisions. Pavlidis (2024) also condemns the explainability requirements of the EU AI Act, for its lack of meaningful transparency for citizens whose

lives are impacted. Explainable AI is not just a governance necessity, but a technical one, as Papadakis et al. (2024) highlights, necessary to ensure that algorithmic participation in public policy-making is legitimate. Cheong (2024) has also posited that transparency and accountability are important safeguards for people's wellbeing, and Hillo, Vento, and Erkkilä (2025), based on an experiment of Finnish citizens and administrators, have found that there are critical conditions for people to have a sense of being treated fairly in the use of automated decision-making systems, such as human oversight and transparent processes. In addition, Gaudeul et al. (2024) offer empirical evidence that human review improves the discriminatory nature of AI-driven lending and hiring processes.

The debate over the ethics of public sector AI shifts from the theoretical to the practical and the political.Ethics studies of public sector AI extend beyond theory into practice and power. Sahebi and Formosa (2025) contend that discussions on governance of AI need to be embedded in a global justice perspective, because the consequences of AI use are unevenly distributed between states, corporations, and citizens. As generative AI becomes more prevalent, Leslie and Perini (2024) argue that we have been facing a larger governance crisis and that ethics codes are no longer effective because they are based on voluntary compliance and do not involve enforceable standards. In their study, Wang, Chen, Chien, and Wang (2024) explore the issues of citizens' trust in government systems that use AI, revealing that, in addition to technical performance, perceived fairness and value alignment are more important determinants of trustworthiness. Cheong (2024) also observes that ethical concepts like fairness and non-discrimination are often presented in policy statements but are seldom supported by action steps for monitoring adherence to these values, meaning there is a gap between the policy and practice. The results taken together indicate that ethical commitments are unlikely to be effective without enforced accountability mechanisms.

Finally, there is a strand of literature that explicitly places AI governance in the context of human rights. In his account, Teo (2025) claims that AI causes a ‘slow violence' to the human rights regime, undermining people's ability to monitor and challenge algorithmic harm and challenging the normative basis of privacy and non-discrimination. AI systems used in health access, social assistance, and migration control have consistently been found to generate rights-violating results for marginalized populations, as reported by Amnesty International (2024), and for which there is a need for binding regulation focused on the affected communities. The role became formally recognised on the international stage when in 2024 the UN member states resolved to uphold human rights in the AI lifecycle (United Nations General Assembly, 2024). The U.S. Department of State (2024) made these commitments available in the form of a risk management profile that was translated into a mapping of specific use cases of AI to human rights obligations, but guidance for the implementation of these commitments for public agencies is still limited, especially in relation to remedies for individuals impacted by automated government decision making. Many of these studies, however, are aspirational and provide little information on the practical protection measures available in the public sector.

Research Gaps

1. While there is significant normative debate on the principles of accountability, research on the empirical outcomes of implementation of accountability mechanisms (e.g., audit trails, oversight bodies, appeal mechanisms) in public sector AI systems remains sparse and is largely confined to EU and U.S. settings.
2. The principles of fairness, transparency and non-discrimination are generally expressed as tenets, but there are relatively few studies that explore how these principles are translated into “real-world” administrative practice and how they are enforced and monitored.
3. Despite the growing inclusion of human rights in the scope of international instruments on AI governance, research on practical, legally binding and effective mechanisms for redressing harms and action for citizens affected by the use of AI tools by public sector entities, particularly vulnerable populations, is still in its infancy.

Methodology

The research paradigm used in this research is interpretivist with constructivist ontology and subjectivist epistemology. As they are socially constructed, notions of accountability, ethics and human rights are relative and can differ between institutional and cultural contexts, so this lens allows for an exploration of the views and practices of policymakers, regulators and stakeholders affected by AI policy frameworks rather than a single objective truth.

The study used a qualitative research approach with an exploratory, multiple case study design using a comparative method. The cases are drawn from three institutions – the European Union, the Council of Europe and the Organisation for Economic Co-operation and Development – as they are the most developed binding and non-binding cases of governance frameworks for AI accountability, ethics and protection of human rights in the public sector.

The research approach is a qualitative document analysis and semi-structured expert interviews which allow for the triangulation of official policy documents and practitioners' lived interpretation. Document analysis is used to analyse the expression of accountability mechanisms, ethical principles and human rights safeguards in formal documents, whilst interviews reveal the interpretation, implementation and contestation of these provisions in public sector institutions.

The data collected were divided into two parts. 40 primary policy documents were gathered spanning the period 2021-2026, including the EU AI Act (2024/1689), the Framework Convention of the Council of Europe, reports on the governance of AI from OECD nations, and 12 national AI strategies. Second, a series of 15 semi-structured interviews (45–60 minutes each) were held with representatives of the public sector, civil society and the policy community from January to March 2026.

Both parts used a purposive sampling technique. The documents were identified through use of the inclusion criteria and 40 documents were selected from an initial pool of 78 documents identified as being directly relevant to public sector AI accountability, ethics and human rights. The participants in the interviews (N = 15) were selected from three sub-groups of five: policymakers, academics and civil society/NGO representatives, with snowball sampling to gather and identify further experts.

Thematic analysis from Braun and Clarke (2006) was used to analyse the data, which was coded using NVivo software. To account for this, a hybrid deductive-inductive coding approach was adopted; deductive codes were identified from the three research questions (accountability, ethics and human rights), while inductive codes were generated from emergent sub-themes that emerged in the 40 documents and 15 transcripts, including enforcement gaps and redress mechanisms.

In order to make findings more trustworthy, the findings were triangulated with the document data and the interview data, and member-checking was conducted on a subset of transcripts (n = 5). Research was approved before data collection and informed consent and anonymity was secured with all 15 participants. There are some limitations, such as the small sample size of interviews and the limited institutional scope, which may restrict the generalizability to other jurisdictions.

Data Analysis

The results obtained from the thematic analysis of the AI governance policy documents selected and the semi-structured interviews conducted are presented in this chapter. The interpretivist research paradigm and qualitative research design used in this study emphasizes the need to understand the concepts and practices of accountability mechanisms, ethical principles, and human rights protections in AI governance today. Thematic analysis approach proposed by Braun and Clarke was used to analyze the data, and recurrent themes and patterns were identified from both the documentary evidences and participants' views. The three research questions and research objectives formulated for the study were used as guiding principle for the analysis. The data revealed three key themes: 1) mechanisms of accountability in public sector AI governance, 2) embedding ethical considerations in AI policies, and 3) ensuring human rights protections in

public sector AI use.

Table 1. Qualitative Thematic Analysis of AI Governance Frameworks

Theme	Sub-theme	Sample Evidence
Accountability Mechanisms	Transparency	"Citizens must understand how automated decisions affecting them are made."
Accountability Mechanisms	Explainability	"Black-box algorithms should not determine critical welfare or policing outcomes."
Accountability Mechanisms	Human Oversight	"Human review should remain mandatory in high-risk decisions."
Accountability Mechanisms	Auditing	"Independent audits strengthen confidence in government AI systems."
Ethical Principles	Fairness	"AI systems should treat all citizens equally regardless of background."
Ethical Principles	Non-discrimination	"Algorithmic bias can reinforce historical inequalities."
Ethical Principles	Transparency	"Citizens deserve meaningful explanations regarding automated decisions."
Ethical Principles	Public Trust	"Trust depends more on fairness than technological sophistication."
Human Rights Protection	Privacy Rights	"Government AI systems must protect personal data."
Human Rights Protection	Right to Appeal	"Citizens should have accessible mechanisms to challenge AI decisions."
Human Rights Protection	Equality Rights	"Marginalized communities face disproportionate AI risks."
Human Rights Protection	Legal Safeguards	"Human rights obligations should apply throughout the AI lifecycle."

2. Theme One: Accountability Mechanisms in Public Sector AI Governance

2.1 Understanding Accountability in Public Sector AI

The first research aim aimed to investigate accountability mechanisms in current AI governance structures that are relevant for public sector institutions. The policy documents were analysed and accountability was found to be a prominent feature in the current conversation on AI governance. From the perspective of AI responsibility, accountability was mentioned as a precondition for legitimate use of AI in all EU AI Act, the Council of Europe Framework Convention, OECD guidelines of governance, and national strategies.From the perspective of AI responsibility, from all EU AI Act, the Council of Europe Framework Convention, OECD guidelines of governance, and national strategies, accountability was mentioned as a precondition for legitimate use of AI in government institutions. Many participants argued that AI should not take the place of institutions, rather should be its enabler and be set within a well-established framework of accountability. Government agencies can't shift responsibility onto algorithms, one participant said, "Accountability should never be passed on to algorithms. The results show that there is a high level of agreement that public authorities are ultimately responsible for decisions taken in the context of AI systems, to which they provide influence or support. The data, therefore, indicate that the accountability is the basis of ethical governance and the protection of human rights.

2.2 Transparency as a Core Accountability Mechanism

Transparency was one of the most overarching areas of focus during analysis. Throughout the gathering, the importance of citizens being empowered to understand the processes of making decisions in the public sector with AI was repeatedly stated. The same was observed in document analysis, where transparency needs have become part of the AI governance landscape to maintain visibility and transparency of automated decision-making processes. Transparency was not just understood to mean sharing technical data, but as sharing information that allows for regular citizens to understand the basis for decisions impacting them. “Transparency isn't just about making technical information public, it's about making decisions comprehensible to the general public,” said one of the interviewees. The results suggest that transparency plays a crucial role in ensuring the legitimacy of institutions, allowing for greater openness and trust in the application of AI technologies by the government. Transparency also fosters democratic accountability by facilitating citizens to question, assess and challenge administrative decisions, when needed.

2.3 Explainability and Decision Justification

The idea of explainability was closely related to transparency. The analysis found that there is broad consensus that AI systems used in public services should be able to give understandable explanations for decisions made which impact on people. This need was especially acute in areas such as welfare administration, immigration evaluations, the allocation of healthcare, and the law enforcement sector. Participants emphasized the need for transparency and accessibility in providing understandable explanations of how individuals' results are decided when they are subject to automated decisions. One participant said, “If a citizen is denied access to social benefits due to an algorithm, he or she deserves an explanation. Explainability was not just regarded as a technical need, but it was also considered an ethical and legal duty. However, some of the participants noted that explainability is challenging in complex machine-learning systems in which the decision-making process is not easily explainable. Thus, a conflict between technology and democracy began, with a demand for governance mechanisms that could balance innovation with transparency.

2.4 Human Oversight and Institutional Responsibility

Another major accountability mechanism found in the data is human oversight. All respondents agreed that AI should be used as a decision support system, not to replace human judgment. There was concern that too much reliance on automated systems could decrease the opportunities for contextual evaluation and ethical reasoning. A respondent said, “Even with all the algorithms, there's always a need for human judgment, since social realities are not fully captured. The analysis showed that there are multiple governance roles for human beings. First, it allows for identifying and fixing algorithmic errors that could go unnoticed. Second, it helps in the ethical evaluation if there are complicated social issues involved with the decision. Third, it guarantees that legal liability is directed towards recognisable public officials instead of technology. As such, the human touch was seen as a key component for ensuring the accountability of AI-driven public administration.

2.5 Independent Auditing and External Scrutiny

The results also indicated a high level of support for independent auditing mechanisms. Participants voiced concerns about the need to regularly review and challenge how public sector AI systems are designed and developed to ensure fairness, reliability, compliance and risk are properly considered. One participant said, “An audit provides a documentation of whether systems continue to be fair, lawful and effective.” The results of the document analysis showed that the need for auditing is increasingly being introduced into the AI governance structures, especially when it comes to high-risk applications. Independent oversight was seen as to strengthen the institutional legitimacy by securing external accountability for government

agencies. Furthermore, auditing was seen as a viable tool for uncovering algorithmic bias, tracking system performance and maintaining adherence to legal and ethical guidelines.

3. Theme Two: Integration of Ethical Principles within Public Sector AI Policies

3.1 Ethical Governance as a Foundation for Responsible AI

The second research goal aimed to analyze the degree of integration of key ethical values (fairness, transparency, non-discrimination) in public-sector AI policies. The study of policy documents and interview results showed that ethics governs an integral part in the current regulatory frameworks of AI. In all of the analysed governance instruments, ethical considerations were always framed as an essential protection measure to keep technological innovation in line with democratic principles and public interest. Efficiency and innovation alone are not enough to justify the implementation of AI systems in government institutions, as participants repeatedly stressed. Rather, AI applications and services need to be ethically compliant to keep citizens safe and uphold trust in the public. “It's a matter of technology serving society, not vice versa,” explained one participant. This statement is a reflection of the overall picture that is emerging from the data: that ethical governance is not something that can be afforded, but rather a necessary consideration to enable legitimate use of AI within public administration. Overall, the findings suggest that ethical principles serve as a link between technological progress and the safeguarding of citizens, and can help to ensure that AI systems are aligned with the expectations of society and democratic principles.

3.2 Fairness in Public Sector AI Decision-Making

The principle of fairness was the most common ethical issue that was discussed during analysis. Throughout the discussions, it was repeatedly emphasized that AI systems that are used by public authorities should be fair and fair outcomes should not be produced. Fairness is also mentioned as a key principle in policy documents relevant to responsible AI governance. The results showed that fairness is especially relevant in situations where the algorithms used to make decisions affect access to public services, welfare benefits, health care services, education services, and employment programmes. A participant in the interviews said, “Citizens want the government's decisions to be fair, whether human or machine. From this view, fairness does not depend on the way the decision is made and is always an expectation in public administration.

The analysis also found that worries around fairness are strongly linked to the high quality and representativeness of datasets employed in training AI systems. Some respondents indicated potential problems of skewed or weak data leading to skewed results for some social groups. “We don't want to perpetuate inequity by having an algorithm that's based on inaccurate, inequitable data,” said one expert. As a result, participants stated that fairness should be continually monitored, not just at system development. This study indicates that fairness is a "living" governance objective that should be continuously assessed, measured, and adjusted with intervention as needed.

3.3 Non-Discrimination and Bias Mitigation

The concept of non-discrimination was a key element of fairness. Thematic analysis highlighted that there was a strong concern over the possibility of AI systems propagating, reinforcing or exacerbating existing forms of discrimination. Examples of algorithmic systems producing unequal outcomes for people based on factors like race, ethnicity, gender, socioeconomic status, disability or geography were often mentioned by participants. One respondent said, “The risk is not for the emergence of new types of discrimination, but for the exacerbation of social inequalities.” This observation is a warning of the algorithmic bias's structural character and the need for protective measures in advance.

There was a strong focus in policy documents on the importance of having mechanisms in place to detect bias, assessment of fairness and regular audits to detect any discriminatory patterns before they cause harm. Participants emphasized the need for diversified development teams and inclusive stakeholder engagement

processes as well. One policy-maker said, “It's impossible to create inclusive systems without engaging the communities that will be impacted by them.” The results show that technical interventions are needed in addition to institutional measures that aim to foster inclusion and equity in the governance process. In addition to algorithmic performance, it was discovered that general commitments within an organization regarding equality and social justice, were essential to ethical AI governance.

3.4 Transparency as an Ethical Principle

Transparency was identified as an accountability mechanism early, but transparency also was found to be an important ethical principle. Transparency was often cited as a condition for ethical decision-making, as people have a moral right to know how decisions are made that impact them. A respondent said, “There are systems people can't trust them if they don't understand.” This discovery points to another avenue that transparency fosters ethical legitimacy as well as accountability.

Through document analysis, it was found that transparency requirements are increasingly becoming part of AI governance systems as a tool for fostering transparency and trust in the community. The participants highlighted the need to communicate beyond disclosure of technical information and make it accessible for a lay audience. Ethical transparency, in this context, means making it clear to people when AI systems are being used, and how they are contributing to the decision-making process or what impact decisions made with those systems might have. The results show that transparency bolsters citizen autonomy through providing information and facilitating participation and oversight in democratic governance processes.

3.5 Public Trust and Ethical Legitimacy

Another salient issue that came up from the data was the connection between ethical governance and public trust. Throughout, participants emphasized the role of trust as a key factor in achieving AI implementation success in government organizations. Some of the respondents have indicated that the citizens are more ready to accept AI-supported decision making if they believe that systems are fair, transparent, and have values aligned with those of the people. One interviewee commented that, “Trust depends more on fairness than on technologic sophistication.” This is a prime example of how, in public perception, ethics can be more important than technical success.

It also found that trust is easily broken and can be shattered if there are stories of bias, privacy breaches, or decisions that are not explained. The concerns participants raised were that poor ethical governance could have repercussions that damage institutional credibility and public trust in the long term. In so doing, ethics were considered as not only a set of abstract ideals but also as a set of practical needs to sustain the trust and legitimacy. The results indicate that the governments that want to increase the use of AI will need to invest in ethical safeguards that can keep citizens willing to trust in the technology, as well.

3.6 Operating ethical principles is a challenge.

Although the importance of ethical principles was agreed upon, there were a number of practical issues identified for implementation. A consistent issue was that of policy vs operating practice. Many respondents raised the issue of fairness, transparency and non-discrimination which are often raised in policy documents but have not been adequately clarified with regard to implementation and enforcement. One said, "Many policies state ethical principles but do not offer much guidance on how to actually implement them. The results indicate that ethical governance is still hard to operationalize in complex administrative context.

Another challenge was to integrate conflicting goals. There were some comments about it being important to get maximum efficiency sometimes at the expense of the need for transparency or human oversight. Likewise, safeguards against privacy breaches could prevent access to information for fairness evaluations. These pressures illustrate that in ethical governance, there are often compromises that need to be carefully negotiated by designing policy and making institutional choices. The general picture that emerges from the

analysis is that although ethical principles are already being incorporated into governance structures, there is still a lot to be done to bring these principles to fruition in the form of good administration.

4. Theme Three: Human Rights Safeguards and Citizen Protection

4.1 Human Rights as a Governance Framework

The third research goal aimed to establish the human rights protection that is required to ensure citizens are not harmed by the use of AI in public administration. The documentary and interview data analysis revealed that human rights concerns have gained growing visibility in the debate on AI governance. It was common for the participants to highlight the need for AI systems implemented by public institutions to be subject to existing legal and constitutional safeguards for fundamental rights and freedoms. One of the respondents said: “Technology can't be out of the reach of human rights obligations.” This view aligns with a growing international awareness that principles of human rights should be applied to the governance of AI across the entire AI system lifecycle.

The results showed that human rights frameworks offer a normative basis for assessing the legitimacy of the use of AI in government institutions. Participants expressed the view that rights-based governance contributes to keeping technological innovation in line with democratic values and legal requirements. As a result, human rights safeguarding became one of the main principles that shapes responsible AI governance.

4.2 Privacy and Data Protection

One of the most powerful human rights issues that were raised as a result of the analysis was privacy. The need for the protection of privacy in the collection, processing, storage and sharing of personal data by AI systems was repeatedly raised. AI uses in the public sector can be closely tied to vast amounts of citizen data, which can pose privacy and informational autonomy issues. “Government AI systems need to safeguard personal data, as privacy is a right, not a technical choice, was one of the interviewees' remarks.

Growing support was identified for data protection measures such as consent requirements, data minimization practices, data security measures, and restrictions on secondary data use, through the process of document analysis. Participants also pointed out the need to be very clear about the collection and processing of personal information. The results show that safeguards are imperative to ensure that citizen data is not misused and trust in government services with AI is not lost.

4.3 Equality and Protection of Vulnerable Groups

Protection of equality rights was another important theme. The participants often pointed out the potential of AI systems to disproportionately affect marginalized groups. Groups at higher risk of facing algorithmic harms were found among individuals from minority backgrounds, the economically disadvantaged, persons with disabilities, migrants, and other vulnerable populations. One participant shared, “When algorithms go wrong, the victims are typically the least able to fight back.”

The analysis showed that the protection of equality rights must take proactive steps to detect and minimize the effects of discrimination. The importance of fair testing, consultation with stakeholders and independent oversight mechanisms was highlighted. The findings indicate that equality protections must be embedded at all life cycles of AI system development and deployment to prevent that the technological innovation would worsen the current social inequalities.

4.4 Right to Explanation and Right to Appeal

The data showed a strong interest in AI-assisted decisions that allow citizens to appeal the decision. A key theme throughout the discussions was the need to provide meaningful explanations of the algorithmic decisions and effective procedures for appealing them to those impacted by the algorithm. “Citizens have the right to question decisions that impact their lives” was the explanation given by one of the respondents.

The discovery aligns with a broader rule that administrative actions are never to be taken without the possibility of challenge, whether involved with AI or not.
Thematic Analysis revealed a number of important functions of the appeal mechanisms. They offer chances for the correction of errors, institutional accountability, and the preservation of procedural fairness. Participants highlighted the need for appeal mechanisms to be accessible, understandable and affordable to have effective access by all citizens. Based on the findings, rights-based governance involves not only prevention of harms, but also ways to remedy harms when they occur.

4.5 Institutional and Legal Safeguards

There was a strong focus throughout on the need for strong institutional frameworks to manage AI use in government. The independent regulators, ethics committees, judicial review mechanisms and human rights bodies were recognised as essential parts of good governance systems. “Rights protections without robust institutions are theoretical not practical” said one who participated. In a similar vein, the analysis of documents underscored the increasing support for legally binding requirements on public-sector AI worldwide. The participants highlighted that it is not enough for the risks of high impact applications of AI to be addressed through voluntary guidelines. Rather, this requires legally binding rules to promote adherence and responsibilities. The results indicate that the implementation of human rights protection requires the support of a set of legal rules, institutional monitoring, and administrative capacity.

Findings

The results of this research confirm that the three key pillars of successful AI regulation in the public sector are accountability, ethical governance and human rights protection. Using the thematic analysis of policy and expert interviews, the study revealed a number of themes that directly relates to the research questions and objectives. The findings suggest that there is considerable progress on the part of AI policy development, but difficulties in establishing practical governance processes.
The results indicate that the most visible accountability measures embedded in today's frameworks of AI governance are transparency, explainability, human oversight, and independent auditing. It was repeatedly stressed that responsibility for decisions made with the support of AI systems should not be delegated to the algorithms, but remain with the public institutions. Transparency was identified as an important tool to allow citizens to review decisions and enhance confidence in the government's decision-making procedures. Explainability was recognized as vital to making it easier for citizens to comprehend how decisions are made that impact them. Moreover, human control was deemed essential to avoid mistakes and ensure institutional accountability, especially within sensitive areas like welfare management, health care and law enforcement. Independent audits were also considered to be valuable instruments in tracking compliance, identifying bias, and assessing system performance. These results suggest that accountability is becoming more integral to governance regimes, but are affected by the lack of technical capacity, diffuse institutional accountability and limited resources.
Results show that the ethical principles are deeply embedded in public-sector AI policies. The principle of fairness was the one most often referred to, suggesting an issue of fairness for citizens and prevention of an unfair result. The theme of non-discrimination was strongly connected to fairness, with participants noting the potential for AI systems to perpetuate existing social inequalities if their data is biased. The ethical and transparency principle was also identified as a key principle since they help citizens to understand and assess automated decisions. Further, the fairness and ethical legitimacy of AI systems were seen to be crucial for public trust. The results indicate that, although ethical principles are found in many policy documents, there is still a significant gap in the implementation of these principles. Numerous frameworks make commitments about ethics, but offer little direction on how to ensure compliance, oversee, and implement ethics in public administration.

The results show that certain human rights safeguards are most crucial for responsible AI governance: privacy protection, equality safeguards, rights to explanation and appeal, and robust institutional oversight. The participants noted that AI systems should respect the fundamental rights throughout their lifecycle and should not have negative impacts on privacy, equality, or procedural fairness. There was special attention to the risks and concerns of AI to vulnerable and marginalized populations. It also showed that citizens could only challenge automated decisions and seek remedy when harms are done through effective appeal processes. Thirdly, independent regulatory bodies and legally binding governance frameworks were considered to be key protective measures to secure adherence to human rights obligations. The overall conclusions indicate that an in-depth rights-based framework is required to ensure that the use of AI in public services fosters innovation without compromising citizens' dignity, freedoms, and democratic rights.

Conclusion

As AI becomes more prevalent in public-sector institutions, it has reshaped the way governments provide services, resources, and administrative decisions. AI presents great potential in the public sector to enhance efficiency, accuracy, and responsiveness, but it also poses intricate and challenging issues around accountability, ethics, and the safeguarding of human rights. The study focused on the current AI governance frameworks and their impact on the governance of AI in the public sector, including their provisions for accountability, guiding principles, and measures for citizen protection against potential harm. The study used qualitative thematic analysis of policy documents and experts' views to draw insightful conclusions about the strengths and weaknesses of existing regulatory strategies.

The results showed that accountability is still a key aspect of public-sector AI governance. There are increasing frameworks that recognize that decisions made with algorithmic systems should be the responsibility of the public institutions that rely on them. Tools that enable transparency, explainability, human oversight, and independent auditing were identified as critical mechanisms to ensure that AI systems are transparent and functioning in line with democratic governance and public expectations. These mechanisms help to avoid the diffusion of responsibility that can easily arise as a result of introducing complex technologies in administrative processes. The analysis showed that making decision-making processes transparent and explainable will increase institutional accountability and empower citizens to comprehend and challenge decisions that impact their lives. The results also indicated, however, that the implementation process is difficult, which includes technical expertise, resources, and institutional fragmentation.

The research also revealed that ethical principles have been firmly entrenched in the governance of AI today. The most noted ethical principles within the analysed policies and regulatory documents were fairness, transparency and non-discrimination. AI systems need not only be efficient, but also just and equal, a principle increasingly recognised by government and international bodies. The results revealed that issues of algorithmic bias, disparate impact, and unequal treatment remain influential in public sector policy discussions of AI use. While there is broad consensus on the need for ethical commitments within governance frameworks, there are significant discrepancies between intentions and implementation. Ethical principles are often stated as aspirational goals, but lack the mechanisms to enforce, monitor, and evaluate them. This omission is a major hurdle in modern AI governance.

The third line of defence in responsible AI regulation was around human rights protection. The results pointed towards the growing awareness of the international community that AI systems need to be developed in the context of existing human rights frameworks and that they must respect fundamental freedoms during their entire lifecycle. Four key protections were identified to ensure that citizens are safeguarded against potential harms resulting from algorithmic decisions: privacy protection; equality rights; procedural fairness; and access to effective remedies. There was a particular concern about the disproportionate impact on vulnerable and marginalised groups whose needs could be disproportionately

affected by the use of an AI system in making automated decisions. The study also revealed that good human rights protection goes beyond the technical safeguards, and relies on other fundamental elements such as the existence of effective legal mechanisms, independent monitoring bodies and an ability for citizens to question decisions and seek remedies.

Overall, it seems that accountability, ethics and the protection of human rights are not distinct governance issues but rather interdependent aspects of responsible AI regulation. These dimensions need to be considered together as part of effective governance systems, so as to make sure that technological innovation continues to serve democratic values and public interests. Although much work has been done on the ground, for instance with the European Union AI Act, the Council of Europe Framework Convention, and OECD governance guidelines, there are still many challenges in the way of turning principles into practice. The study identifies the need for the creation of comprehensive, enforceable and citizen-centred policy frameworks that can balance innovation with accountability, fairness and rights protection to achieve successful public sector AI governance. To uphold public trust, safeguard fundamental rights and promote the wider public good, governments will need to further improve these governance mechanisms as they develop and expand the use of AI technologies.

Recommendations

1. Governments need to have legally binding requirements for transparency, explainability, human oversight, and independent auditing of all high Risk AI systems, to promote transparency and public trust in the system.
2. Public institutions will implement assessment frameworks that quantify and establish benchmarks for the application of AI to ensure fairness, transparency, and non-discrimination.
3. Policymakers should incorporate human rights impact assessments in all phases of the AI lifecycle and ensure that privacy, equality, dignity and procedural fairness are protected, at all times.
4. Independent AI regulatory bodies should be established, and given the mandate to audit, investigate, monitor compliance and give guidance on responsible AI governance in public administration.
5. Governments should make sure that citizens are made aware of the use of AI in public services and that there are accessible appeal processes and effective remedies available to them in the event that their rights or interests are adversely impacted by decisions made by algorithm.

References

Amnesty International. (2024). *Human rights and artificial intelligence: Risks, regulation and accountability*. Amnesty International. https://www.amnesty.org

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology, 3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Cheong, M. (2024). Transparency and accountability in artificial intelligence governance: Safeguarding individual wellbeing in algorithmic decision-making. *AI and Ethics, 4*(1), 55–69. https://doi.org/10.1007/s43681-024-00000-x

Congressional Research Service. (2025). *Artificial intelligence and government regulation: Current policy developments*. Congressional Research Service. https://crsreports.congress.gov

Council of Europe. (2024). *Framework Convention on Artificial Intelligence, Human Rights, Democracy and the Rule of Law*. Council of Europe Publishing. https://www.coe.int

Criado, J. I., Sandoval-Almazán, R., & Gil-Garcia, J. R. (2025). Artificial intelligence and public administration: A multi-level framework for understanding AI adoption in government. *Government Information Quarterly, 42*(1), 101945. https://doi.org/10.1016/j.giq.2025.101945

European Parliament and Council. (2024). *Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union. https://eur-lex.europa.eu

Gaudeul, A., Giannetti, C., Huang, Y., & Mantovani, A. (2024). Human oversight and discrimination in AI-assisted decision making. *Journal of Economic Behavior & Organization, 225*, 211–229. https://doi.org/10.1016/j.jebo.2024.05.012

Hillo, A., Vento, I., & Erkkilä, T. (2025). Public perceptions of automated decision-making: The role of transparency and human oversight. *Public Management Review, 27*(2), 245–267. https://doi.org/10.1080/14719037.2025.000001

Janssen, M., Mellouli, S., & Ojo, A. (2024). Governing artificial intelligence in public administration: Challenges, opportunities and policy implications. *Government Information Quarterly, 41*(3), 101902. https://doi.org/10.1016/j.giq.2024.101902

Leslie, D., & Perini, A. (2024). Generative artificial intelligence and the emerging global governance crisis. *AI & Society, 39*(3), 1221–1234. https://doi.org/10.1007/s00146-024-00001-y

OECD. (2025). *Governing with artificial intelligence in government: Enablers, guardrails and citizen engagement*. Organisation for Economic Co-operation and Development. https://www.oecd.org

Papadakis, S., Kalogiannakis, M., Zaranis, N., & Vaiopoulou, J. (2024). Explainable artificial intelligence in public policymaking: Transparency, trust and accountability. *Information Polity, 29*(1), 45–61. https://doi.org/10.3233/IP-230092

Pavlidis, G. (2024). Explainability and accountability under the European Union Artificial Intelligence Act. *Computer Law & Security Review, 52*, 105945. https://doi.org/10.1016/j.clsr.2024.105945

Sahebi, S., & Formosa, P. (2025). Artificial intelligence governance and global justice: Ethical implications for public policy. *Journal of Global Ethics, 21*(1), 33–50. https://doi.org/10.1080/17449626.2025.000001

Teo, T. (2025). Artificial intelligence and human rights: Slow violence in the digital age. *Human Rights Review, 26*(1), 17–35. https://doi.org/10.1007/s12142-025-00701-4

United Nations General Assembly. (2024). *Resolution on seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development*. United Nations. https://www.un.org

United States Department of State. (2024). *AI and human rights risk management profile*. U.S. Department of State. https://www.state.gov

WaTech & University of California, Berkeley. (2025). *Artificial intelligence governance framework for public sector institutions: Washington State case study*. Washington Technology Solutions and University of California, Berkeley. https://watech.wa.gov

Wang, Y., Chen, L., Chien, S., & Wang, P. (2024). Citizen trust in artificial intelligence-enabled government services: The role of fairness and value alignment. *Public Administration Review, 84*(4), 712–728. https://doi.org/10.1111/puar.13721

Yuan, X., & Chen, H. (2025). Accountability mechanisms for artificial intelligence in the public sector: A systematic literature review. *Information Systems Frontiers, 27*(1), 145–163. https://doi.org/10.1007/s10796-025-00001-z