Jurnal Teknik Informatika C.I.T Medicom, 16 (5) (2024) 309-320

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

Explainable AI for Public Sector Decision Making: A Systematic Literature Review

Roland Vincent Karl
Artificial Intelligence and Intelligent Systems Research Group, School of Innovation, Design and Engineering, Mälardalen University, Sweden

Article Info

Article history

Received : October 20, 2024
Revised : November 26, 2024
Accepted : November 28, 2024

Key Words:

Explainable Artificial Intelligence (XAI);
Public Sector Decision Making;
Algorithmic Transparency;
Accountability in Governance;
Systematic Literature Review.

Abstract

The growing adoption of Artificial Intelligence (AI) in government has intensified the need for transparent, accountable, and trustworthy decision-making systems. This study conducts a systematic literature review to examine how Explainable AI (XAI) is applied within the public sector, identify the dominant techniques used, and analyze their benefits and challenges. Using PRISMA guidelines, studies were collected from major academic databases including Scopus, Web of Science, IEEE Xplore, SpringerLink, ACM Digital Library, and Google Scholar. The findings reveal that XAI development in government contexts has grown significantly over the past decade, with SHAP, LIME, decision trees, counterfactual explanations, and rule-based models emerging as the most frequently used methods. These techniques support public-sector decision making by enhancing transparency, strengthening accountability, reducing bias, improving auditability, and fostering public trust. However, persistent challenges remain, including technical complexity, trade-offs between accuracy and interpretability, limited AI literacy among officials, lack of standard frameworks, and legal or ethical risks. The review highlights the need for more domain-specific XAI guidelines, user-centered explanation tools, and integrated evaluation frameworks. This research contributes a comprehensive synthesis of current XAI applications in government and outlines a future research agenda to support the development of responsible, explainable, and ethically aligned AI for public administration.

Corresponding Author:

Roland Vincent Karl
Artificial Intelligence and Intelligent Systems Research Group, School of Innovation, Design and Engineering, Mälardalen University, Sweden
Universitetsplan 1, 722 20 Västerås, Sweden
rolandkarl@mdu.se

1. Introduction

Artificial intelligence (AI) has increasingly become a central component in the modernization of public administration[1]. Governments around the world are adopting AI-based systems to improve efficiency, enhance service delivery, support complex decision-making processes, and optimize the allocation of public resources. AI technologies are now widely used in areas such as healthcare prioritization, tax compliance, criminal justice risk assessment, environmental monitoring, public welfare eligibility, and smart city management. Despite these advancements, the use of AI in the public sector presents unique challenges because public decisions inherently affect citizens' rights, welfare,

and trust. Unlike private-sector applications, government decisions require a higher standard of transparency, accountability, legality, and fairness. These demands highlight a critical requirement: public sector AI systems must not only perform accurately, but must also be explainable.

Traditional AI systems particularly deep learning models operate as "black boxes," where decision-making logic cannot be easily understood by end users, policymakers, or even developers. This opacity creates substantial risks, including biased outcomes, discriminatory decisions, lack of citizen trust, difficulties in auditing, and challenges in meeting ethical and legal obligations[2]. In high-stakes public-sector decisions, unexplained decisions can lead to public resistance, institutional distrust, and legal disputes. Consequently, Explainable Artificial Intelligence (XAI) has emerged as a crucial research area that seeks to make machine learning models transparent, interpretable, and understandable to humans. XAI provides methods that reveal how input factors influence outputs, generate human-friendly explanations, and support accountability mechanisms within AI systems.

In the context of public sector governance, XAI holds significant promise. It can enable civil servants to justify algorithm-supported decisions, allow oversight bodies to evaluate the fairness of automated processes, and support citizens' rights to understand how decisions affecting them are made. XAI also aligns with international regulatory trends, such as the European Union's GDPR "right to explanation" and the forthcoming AI Act, which mandate transparency for high-risk AI systems, including those used by public institutions[3]. As governments worldwide continue to adopt digital transformation initiatives, there is an urgent need to ensure that AI tools used in public decision-making are both effective and trustworthy.

Explainability methods and the foundations of XAI were established by key methodological works. Ribeiro, Singh, and Guestrin (2016) introduced LIME a model-agnostic local explanation method that explains individual predictions by fitting interpretable surrogates around instances. Lundberg and Lee (2017) proposed SHAP, a unified, game-theory based framework that assigns Shapley-value importances to features and links several prior explainers under a common axiomatic view. Around the same time, Doshi-Velez and Kim (2017) argued for a rigorous, task-driven science of interpretability and proposed evaluation frameworks and taxonomies for when explanations are needed. More comprehensive taxonomies and state-of-the-art syntheses followed, notably Arrieta et al. (2019/2020), who produced a widely-cited review that organizes XAI concepts, methods, taxonomies, evaluation challenges, and open research directions.

Empirical and comparative work testing specific explainers has illuminated strengths and limits of popular techniques. For example, Gramegna and Giudici (2021) compared SHAP and LIME in a credit-risk context and examined discriminative power and stability of explanations; their results and similar comparative studies show that explainers differ in stability, faithfulness, and suitability depending on model type, data characteristics, and the evaluation metric used. Subsequent methodological critiques and case studies have emphasized issues such as explanation instability, model-dependence, and the need to match explanation type to stakeholder needs.

A parallel strand of research has critically examined algorithmic systems in public-sector contexts, documenting harms when opaque models are used for high-stakes administrative decisions. Cathy O'Neil's book (2016) "Weapons of Math Destruction" and Virginia Eubanks's (2018) "Automating Inequality" provided influential, evidence-rich critiques showing how opaque algorithmic systems used by government agencies can amplify bias and inequality. Building on these critiques, Levy, Chasalow, and Riley (2021) surveyed algorithmic systems deployed across criminal justice, education, social benefits, and municipal services, and they highlighted accountability, procurement, and evaluation gaps that complicate safe public-sector adoption of predictive systems.

Policy-oriented and governance literature has developed frameworks and practical recommendations for algorithmic accountability, transparency, and explainability in government. Multi-stakeholder reports and analyses (e.g., Ada Lovelace / AI Now / OGP executive summaries and related "algorithmic accountability" reports, 2021) catalogue policy mechanisms (audits, procurement rules, impact assessments) and early government experiments in algorithmic transparency. Academic reviews of explainability policy (e.g., Nannini et al., 2023) have mapped standards and identified patchy

regulatory coverage and a need for clearer norms about what counts as an acceptable explanation in administrative contexts.

Recent applied research has begun to empirically test how different explanation types affect perceptions and outcomes in public-sector settings. For instance, Aoki et al. (2024) examined whether the type of explanation (local vs. global, procedural vs. outcome-focused, etc.) changes perceived fairness, accuracy, and trustworthiness of adverse algorithmic administrative decisions, and found that effects vary by domain and by the content of the explanation. Papadakis et al. (2024) and similar applied studies discuss concrete design patterns and toolchains to make policy-oriented ML pipelines more transparent in practice, exploring both technical solutions and procedural safeguards for policymaking.

Although interest in XAI has grown rapidly, research on the application of XAI specifically in public sector settings remains fragmented and dispersed across technical, legal, ethical, and public administration domains. Several studies focus on the development of XAI algorithms, while others examine the governance and societal implications of automated decision systems. However, there is a lack of comprehensive synthesis that integrates these diverse perspectives and systematically maps how XAI has been implemented, evaluated, and conceptualized within public decision-making contexts. Moreover, current literature shows limited discussion on how XAI techniques address practical challenges encountered by government agencies, such as interpretability trade-offs, organizational capacity constraints, and user acceptability.

Given these gaps, a systematic literature review (SLR) is needed to provide a consolidated understanding of the state of XAI in public sector decision making[4]. An SLR can identify dominant research themes, categorize XAI methods used in government applications, evaluate reported outcomes, and highlight existing barriers and opportunities. It can also guide policymakers, AI practitioners, and researchers by providing evidence-based insights into how XAI can enhance transparency, fairness, and accountability in public decision-making systems.

Therefore, this research aims to systematically analyze scholarly publications on explainable AI within the public sector to map current applications, methodological trends, benefits, challenges, and research gaps. By synthesizing the existing body of knowledge, this study contributes to the development of more transparent, responsible, and trustworthy AI systems for governmental use and provides a foundation for future interdisciplinary research and policy development.

2. Research Methodolgy

This study employed a Systematic Literature Review (SLR) approach to synthesize existing research on Explainable Artificial Intelligence (XAI) in public sector decision making[5]. The SLR was designed to ensure transparency, replicability, and methodological rigor through a structured process of planning, searching, screening, selecting, and analyzing relevant literature. The methodology was guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework and complemented by principles from Okoli's systematic review procedures. Together, these frameworks provided clear guidelines to minimize bias and ensure that the review accurately reflects current knowledge in the field. Although the review protocol was not formally registered in a public registry, all steps were documented to maintain procedural consistency and traceability throughout the research process.

To gather a comprehensive dataset of scholarly publications, the study searched multiple reputable academic databases representing the fields of computer science, engineering, public administration, and multidisciplinary research[6]. The primary databases included Scopus, Web of Science, IEEE Xplore, SpringerLink, and the ACM Digital Library. Google Scholar was used as a supplementary source to identify additional grey literature and recent publications not yet indexed in the major databases. These databases were selected based on their extensive coverage of peer-reviewed journal articles, conference papers, and relevant scholarly outputs that focus on artificial intelligence and its applications in the public sector.

The search strategy was developed to capture a broad range of studies related to explainability in AI-based public sector decision systems[7]. A combination of keywords and Boolean operators was used to form the search queries. Core keywords included "explainable AI", "XAI", "interpretable machine learning", "public sector", "government", and "decision making". Example Boolean combinations used were: ("explainable AI" OR "XAI" OR "interpretable AI") AND ("public sector" OR "government") AND ("decision making" OR "policy"). These queries were adapted slightly for each database depending on its search syntax requirements. The search process was iterative; initial searches were tested for relevance, and keywords were refined to improve the accuracy and specificity of the retrieved publications.

The inclusion and exclusion criteria were established prior to screening to ensure consistency in selecting relevant studies. The inclusion criteria required that publications be peer-reviewed, written in English, and published within reputable academic venues. Only studies focusing on AI-supported decision making in the public sector and employing or discussing explainability techniques were considered[8]. This included both empirical studies and conceptual analyses related to XAI adoption, frameworks, or evaluation in government contexts. Studies were excluded if they focused solely on technical algorithmic developments without any public-sector application, examined AI systems without explainability components, or consisted of non-scholarly materials such as editorials, opinion essays, workshop summaries, and non-peer-reviewed content. Additionally, studies that addressed transparency in general without explicit reference to XAI methods were excluded to maintain thematic consistency.

The screening process followed the PRISMA flow structure, consisting of identification, screening, eligibility assessment, and final inclusion. During the identification stage, all retrieved records from the databases were exported into a reference management tool, and duplicates were removed[9]. The screening process involved title and abstract review to eliminate clearly irrelevant studies. Articles that passed this stage underwent full-text evaluation to determine their suitability based on the predefined inclusion and exclusion criteria. At the end of the screening process, the number of studies included in the final dataset was documented as per the PRISMA guidelines, ensuring clarity in reporting how many records were identified, screened, excluded, and retained.

For each included study, a structured data extraction process was conducted. A data extraction form was developed to capture relevant variables such as publication year, authorship, country or region of study, application domain (e.g., healthcare, law enforcement, welfare, finance, education), type of AI model (machine learning, deep learning, hybrid), and specific XAI techniques employed (such as SHAP, LIME, rule-based models, counterfactual explanations, or feature importance methods). Additional fields captured the purpose of the AI system, evaluation metrics used, key findings, reported challenges, and implications for public sector decision making. This systematic coding ensured that the extracted data were consistent, comparable, and suitable for synthesis.

The data analysis was conducted using thematic analysis to identify major patterns, trends, and themes across the selected studies. After coding the extracted data, studies were grouped into categories based on their application domain, the XAI methods applied, and the types of benefits and barriers reported[10]. Thematic analysis allowed the researchers to uncover recurring themes such as transparency enhancement, fairness and accountability, organizational readiness, interpretability trade-offs, and limitations of current explainability techniques. Content coding was also applied to categorize challenges (e.g., technical complexity, governance issues, legal constraints) and benefits (e.g., improved trust, auditability, decision justification). This analytic approach enabled a holistic interpretation of the literature and provided meaningful insights into how XAI is being leveraged in public sector decision processes and what gaps remain for future research.

3. Results and Discussion

Publication Trends

The results of the systematic review show a steady increase in scholarly attention to Explainable Artificial Intelligence (XAI) within the context of public sector decision making. Early publications

before 2017 were limited and primarily conceptual, focusing on general concerns about algorithmic transparency and accountability in government systems. However, from 2018 onward, the number of studies gradually increased, coinciding with the global rise of AI adoption in governance and the emergence of formal regulatory frameworks such as the EU General Data Protection Regulation (GDPR) and early discussions surrounding the EU AI Act. The most notable research growth occurred between 2020 and 2024, during which more than half of the included studies were published[11]. This surge reflects heightened global awareness of the risks associated with opaque algorithmic systems in areas such as public welfare distribution, predictive policing, health triaging, and administrative decision automation.

In terms of geographic distribution, research on XAI for public sector decision making is concentrated in technologically advanced regions and countries with strong public-sector digital transformation agendas. Europe emerges as the most active region, led by the United Kingdom, Germany, the Netherlands, and the Scandinavian countries. This strong European presence is closely tied to the region's emphasis on regulatory oversight, data protection, and ethical AI governance. North America, particularly the United States and Canada, also contributes significantly, with studies focusing on criminal justice, public health management, and algorithmic auditing. A growing body of work has begun to emerge from Asia especially China, South Korea, Singapore, and Japan where interest in integrating explainability into smart city systems, e-government services, and administrative automation is increasing. Meanwhile, contributions from developing countries remain limited, although some studies explore XAI applications in digital public service delivery for sectors such as social assistance and public infrastructure monitoring.

Across the reviewed literature, the most common application domains reflect areas of government activity where AI-supported decisions directly affect citizens, involve high stakes, or require strict justification and accountability. Public healthcare and medical resource prioritization constitute one of the largest domains, with many studies exploring explainable triage systems, outbreak prediction models, and decision-support tools for public health administrators[12]. Another prominent domain is law enforcement and criminal justice, where XAI methods are employed to improve transparency in predictive policing, recidivism risk assessment, and community safety analytics. Social welfare and public benefits distribution also represent a significant area, driven by the need to explain eligibility decisions and prevent discriminatory or unfair allocation. Additional domains include tax compliance and fraud detection, environmental monitoring, smart city management, education resource allocation, and digital identity systems. Overall, the concentration of studies in these domains highlights the public sector's particular need for AI systems that are not only accurate but also interpretable, auditable, and aligned with principles of fairness and legal compliance.

XAI Methods Used

The analysis of the selected studies reveals that a diverse range of Explainable Artificial Intelligence (XAI) methods has been applied in public sector decision-making systems, reflecting the growing need for transparency, accountability, and interpretability in government applications. These methods broadly fall into two categories: post-hoc explainability techniques and intrinsic (interpretable-by-design) models. Post-hoc XAI methods aim to provide explanations after a complex model has generated an output, while intrinsic models are inherently interpretable and allow decision logic to be understood directly from their structure. Both approaches play important roles in public-sector contexts, although the reviewed studies show a preference for post-hoc techniques due to their flexibility and compatibility with high-performing but opaque machine learning models.

Among post-hoc approaches, SHAP (SHapley Additive exPlanations) emerges as the most widely used technique across the reviewed literature[13]. SHAP is favored because it offers consistent, mathematically grounded explanations by distributing credit for a model's prediction among input features based on cooperative game theory. Researchers frequently employ SHAP to explain risk scores in criminal justice, eligibility decisions in social welfare systems, and prioritization algorithms in

healthcare management. Its ability to generate both local and global explanations makes it highly suitable for public-sector settings where transparency is required at both the case level and policy level.

LIME (Local Interpretable Model-Agnostic Explanations) also appears frequently, particularly in studies requiring case-specific explanations for decisions affecting individual citizens. LIME is often used in administrative decision-making, public benefits distribution, and environmental monitoring systems. Its simplicity and model-agnostic nature make it useful for government agencies experimenting with interpretable outputs without needing to modify underlying black-box algorithms. However, some studies highlight concerns about LIME's stability and sensitivity to perturbations, prompting interest in complementary techniques.

Intrinsic explainability methods are also represented, particularly through decision trees and rule-based models, which continue to be valued for their transparency and ease of interpretation. Decision trees are often employed in sectors where decisions must be defensible and easy to communicate, such as tax compliance, resource allocation, and public service eligibility assessments[14]. Their graphical structure allows administrators to trace the reasoning path directly. Likewise, rule-based models are commonly used in legal and policy-driven domains because they align well with existing administrative frameworks, enabling decision makers to integrate explicit rules with data-driven insights.

Another emerging category of explanation methods in the reviewed studies is counterfactual explanations, which provide insights by showing how an outcome would change if certain input features were modified. These explanations are particularly useful in high-stakes contexts such as loan eligibility, welfare decisions, and recidivism risk scoring because they offer actionable guidance to affected individuals. Counterfactuals help citizens understand what conditions might lead to a more favorable decision, thereby supporting fairness and enhancing procedural justice in automated government decisions.

Additionally, several studies rely on feature importance methods, such as permutation importance and gradient-based importance scoring. These methods are often applied when agencies require a global understanding of which variables most influence automated decisions. For example, they have been used to explain anomaly detection models in tax enforcement, risk evaluation in emergency response systems, and predictive models for public policy forecasting. Although feature importance scores provide useful global insights, some studies warn that they may be too abstract for public communication and may require additional explanatory layers to be meaningful to non-expert stakeholders.

Overall, the findings indicate that public-sector XAI applications rely on a hybrid ecosystem of post-hoc and intrinsic techniques. While post-hoc methods dominate due to their compatibility with advanced machine learning models, intrinsically interpretable models remain essential in contexts where transparency must be guaranteed without relying on supplementary explanations. This diversity of approaches reflects the complexity of public-sector decision making, where accuracy, fairness, legal compliance, and interpretability must be balanced to ensure responsible and trustworthy AI adoption.

Applications in the Public Sector

The reviewed studies demonstrate that Explainable Artificial Intelligence (XAI) has been increasingly applied across a diverse range of public sector domains, particularly in areas where automated decision-making directly affects citizen rights, access to services, public safety, and resource allocation. These applications highlight the growing recognition that government use of AI must not only be accurate but also transparent, accountable, and easily interpretable by both administrators and the public. Several key domains emerged as the most prominent areas of XAI adoption in governmental contexts.

One of the most significant domains is healthcare and public health policy, where AI systems are used to support decisions related to disease prediction, patient triage, resource prioritization, and epidemiological forecasting[15]. XAI methods help policymakers and health administrators understand why certain populations are identified as high-risk, which factors influence triage recommendations, and how predictive models allocate medical resources during emergencies. For example, SHAP and decision-tree–based explanations are often employed to clarify variables influencing infection risk

predictions or hospital admission prioritization. Transparency in these systems is crucial because healthcare decisions affect public trust and can have life-or-death consequences, especially during pandemics or resource shortages.

Another prominent application area is criminal justice and policing, where XAI is introduced to mitigate the long-standing concerns about bias, opacity, and fairness in algorithmic risk assessments and predictive policing tools. Models used to predict recidivism, identify high-crime areas, or allocate police patrols are increasingly accompanied by explanation mechanisms such as LIME, SHAP, and rule-based logic. These explanations allow policymakers, judges, and law enforcement agencies to scrutinize risk factors driving predictions, evaluate potential biases across demographic groups, and ensure that decision-making processes remain accountable. The literature reveals that explainability is particularly important in this domain due to the ethical and legal implications of algorithmic decisions that may restrict individual freedom or influence judicial outcomes.

XAI is also widely applied in smart city management and transportation systems, where cities rely on predictive algorithms to manage traffic flow, optimize public transportation routes, monitor infrastructure, and allocate municipal resources. Explainable models help urban planners understand why certain congestion patterns occur, which environmental or behavioral factors influence mobility forecasts, and how interventions may impact residents. Counterfactual explanations and feature-importance analyses are commonly used to identify actionable improvements, enabling municipalities to make data-driven decisions while maintaining transparency for public accountability.

In the field of taxation and fraud detection, governments increasingly deploy AI systems to identify anomalies, detect suspicious taxpayer behavior, and flag potential cases of fraud or non-compliance. XAI techniques are essential in this domain to justify why certain taxpayers are selected for audits or investigations[16]. Feature importance methods, rule-based classifiers, and decision trees are commonly used because they offer clear reasoning paths that auditors can interpret and present as evidence. Explainability also helps prevent discriminatory or unjust targeting and ensures that automated tax enforcement practices remain consistent with administrative law principles.

Another growing application area is environmental policy and sustainability management. Governments use predictive models to monitor pollution levels, forecast natural disasters, manage conservation efforts, and evaluate environmental risks. XAI method particularly SHAP and interpretable machine learning models help explain the key drivers behind environmental outcomes, such as temperature changes, land-use patterns, or industrial emissions. These insights not only inform policy adjustments but also support transparent communication with stakeholders, enabling more informed public debate on environmental decisions.

Finally, XAI plays a critical role in social welfare eligibility and benefits distribution, where AI systems are used to determine whether individuals qualify for housing assistance, unemployment benefits, subsidies, child support, or healthcare coverage. Given the sensitivity of these decisions, explainability is essential to prevent unfair exclusion, reduce algorithmic discrimination, and ensure that citizens understand the basis of automated judgments. Counterfactual explanations, in particular, help individuals see what changes would lead to a different eligibility outcome, thereby supporting procedural fairness and improving trust in government services.

Benefits

The reviewed literature consistently highlights that Explainable AI delivers a range of substantive benefits for public-sector decision making, particularly in domains where fairness, legality, and public accountability are essential. Transparency emerges as one of the most frequently cited advantages. XAI provides visibility into how algorithmic decisions are generated, enabling government agencies to move away from “black box” systems toward models whose reasoning can be inspected and communicated[17]. Studies show that explainable models allow public officials to justify decisions to stakeholders, auditors, or affected citizens while also helping end-users understand the strengths and limitations of the underlying AI system.

Another widely reported benefit is accountability, which becomes increasingly important as governments incorporate algorithmic tools into policy delivery and public service operations. XAI

enables traceability by showing which features, rules, or model components influenced an output. This traceability supports the ability of public institutions to assign responsibility, maintain compliance with regulatory standards, and meet ethical governance expectations. It ensures that automated systems amplify rather than replace human responsibility in public decision making.

A number of studies emphasize the role of XAI in reducing bias within algorithmic systems used by government agencies. Because explanations highlight what the model prioritizes, analysts can detect patterns of potential discrimination or skewed feature contributions. For example, explanations may reveal when sensitive variables or proxies inadvertently influence decisions, enabling timely correction. This proactive approach not only improves model fairness but also pushes institutions toward more equitable and inclusive policy outcomes.

Related to fairness, improved public trust stands out as a central benefit. Citizens are generally hesitant to accept automated decisions, particularly those that determine access to public benefits, healthcare, or legal judgments[18]. XAI tools help alleviate these concerns by enabling governments to provide clear, understandable rationales behind AI-assisted decisions. Transparent communication fosters trust in both the technology and the institutions deploying it, allowing citizens to feel more confident that decisions affecting them are made fairly and systematically.

Governments also benefit from better auditability, as explainable systems facilitate more rigorous internal and external auditing processes. Public agencies must adhere to strict legal, procedural, and ethical standards; XAI models support this by offering verifiable evidence of how conclusions are reached. This is crucial for sectors such as taxation, criminal justice, or social welfare, where AI systems may be challenged through administrative review or judicial processes.

Finally, XAI contributes to more ethical decision making by aligning algorithmic behavior with principles such as justice, equity, and the right to explanation. Explainability encourages developers and policymakers to design AI systems that prioritize fairness, minimize harm, and uphold human rights. It also ensures that human supervisors can override or question automated outputs when explanations indicate inconsistencies, risks, or unintended consequences. Ultimately, XAI helps embed ethical considerations into the development and deployment of AI tools, creating systems that serve the public interest more responsibly.

Challenges

Despite the promising advantages of Explainable AI, the literature indicates that its implementation in public-sector decision making faces several persistent challenges. One of the most frequently discussed issues is technical complexity. XAI methods such as SHAP, LIME, and counterfactual modeling often require specialized knowledge to configure, interpret, and validate. Government agencies, which may not have the same level of technical capacity as private-sector AI labs, struggle to integrate these tools into existing systems[19]. The complexity is further magnified when multiple data sources, legacy infrastructures, or cross-departmental processes are involved, making it difficult to produce explanations that are both accurate and comprehensible.

Another major challenge is the well-known trade-off between accuracy and explainability. Many high-performing models used in predictive analytics like deep neural networks or ensemble methods are inherently opaque. Introducing explainability can reduce performance or oversimplify the model's reasoning. Conversely, choosing more interpretable models may lead to lower predictive accuracy, which can compromise the effectiveness of public programs. Policymakers therefore face a dilemma: should they prioritize a model that is highly accurate but difficult to justify, or one that is interpretable but potentially less effective?

A recurring theme across studies is the lack of standardized XAI frameworks and guidelines for government use[20]. Unlike sectors such as finance or healthcare, which have clearer regulatory pathways for algorithmic transparency, the public sector lacks uniform protocols that specify what counts as a sufficient explanation, how explanations should be audited, or which stakeholders are entitled to them. This absence of standardization leads to inconsistent practices across agencies and creates uncertainty about legal compliance, ethical obligations, and technical implementation.

Without clear frameworks, it becomes challenging to scale XAI solutions across government departments.

The literature also identifies limited AI literacy among public officials as a substantial barrier. Many civil servants, policymakers, and administrative staff lack the training needed to interpret AI outputs, evaluate explanation quality, or detect algorithmic risks. This knowledge gap can result in either over-reliance on AI without sufficient scrutiny or excessive skepticism that prevents adoption altogether. The lack of technical capacity also limits the ability of agencies to communicate explanations effectively to the general public, which further undermines trust and accountability[21].

Another key challenge relates to legal and ethical risks. Governments operate under strict legal frameworks governing fairness, due process, privacy, and administrative accountability. If explanations are unclear or incomplete, agencies face exposure to legal challenges, public criticism, or policy failures. Ethical concerns also arise when explanations reveal sensitive features, when explanations inadvertently expose private data, or when citizens do not fully understand their rights regarding algorithmic decisions. Balancing transparency with privacy and security therefore becomes a delicate and complex task.

Finally, many studies point to the computational costs of implementing XAI. Techniques like SHAP or counterfactual explanations can be computationally expensive, especially for large-scale government datasets or real-time decision environments. This poses challenges for agencies operating with limited budgets, outdated hardware, or constrained data infrastructures[22]. In addition, generating explanations at scale such as for thousands of welfare applications or medical assessments can significantly increase processing time, reducing system efficiency and increasing operational costs.

Comparison with Previous Reviews or Related Literature

Several prior reviews have examined the broader landscape of Explainable AI (XAI), yet most of them focus on technical methods or private-sector applications, leaving a noticeable gap in the context of public-sector decision making. Earlier surveys, such as those by Adadi and Berrada (2018) and Samek and Müller (2019), concentrated on taxonomies of XAI techniques, model interpretability, and algorithmic transparency. While their contributions provide a comprehensive overview of methodological developments, they do not address the unique constraints and governance requirements faced by public institutions. These foundational works serve as important theoretical precursors, but they offer limited insights into how XAI is deployed in domains such as social welfare, healthcare policy, policing, or taxation.

More recent reviews have begun exploring the intersection of AI and public administration but often treat explainability only as a secondary theme[23]. For example, Wirtz, Weyerer, and Geyer (2019) conducted a systematic review on AI adoption in public administration, highlighting opportunities and risks but only briefly touching on the need for transparency and interpretability. Similarly, Bullock (2019) reviewed algorithmic governance trends and emphasized concerns over fairness and accountability, yet did not delve deeply into specific XAI methods or their applicability across government sectors. These studies underscore the importance of trustworthy AI but stop short of providing a structured, method-level synthesis of how explainability is operationalized in the public sector.

A number of domain-specific reviews have also explored AI in areas like healthcare, law enforcement, and smart cities, but they generally focus on performance, implementation challenges, and ethical considerations rather than the mechanics of explanation generation. For instance, Rajkomar et al. (2019) discussed interpretability issues in clinical AI systems, while Ferguson (2020) examined algorithmic decision making in policing with a strong ethical focus. Although these works recognize the need for transparency, they do not systematically map XAI techniques or compare their suitability for public governance contexts[24].

Compared to these earlier studies, the present systematic review offers a more focused and integrated assessment by explicitly examining how XAI is used, why it is needed, and what methods are most suitable for public-sector decision making. Unlike previous surveys that primarily emphasize technical classifications or general AI governance, this review synthesizes findings across multiple

government domains, identifying patterns in XAI adoption, emerging best practices, and challenges unique to public administration. Additionally, this study adds methodological depth by analyzing not only XAI tools but also evaluation metrics, contextual constraints, and the socio-technical implications of deploying explainable models in high-stakes, policy-relevant environments.

While earlier reviews provide valuable foundations on technical XAI concepts or AI governance principles, none has offered a comprehensive, systematic exploration of XAI specifically tailored to public-sector decision making. This research fills that gap by bridging technical literature with public administration needs, offering a holistic and policy-relevant understanding of how explainable AI can support transparency, accountability, and ethical governance in modern governmental systems.

Certain XAI Methods Are More Suitable for High-Stakes Decisions

One major factor driving suitability is the need for clear and human-understandable explanations[25]. High-stakes decisions often involve legally mandated requirements for justification, which means that explanations must be interpretable by non-technical stakeholders such as judges, social workers, auditors, and members of the public. Techniques such as decision trees, rule-based models, and feature importance methods are particularly useful because they provide straightforward, intuitive reasoning structures. For example, decision trees explicitly outline the decision path, allowing officials to trace why a specific outcome was reached. Rule-based models similarly present explanations in a format familiar to legal and administrative contexts, mirroring the logic of policy rules and statutory guidelines.

Another reason certain methods are preferred is their consistency and robustness across different cases, which is essential for fairness and accountability. Post-hoc techniques like SHAP (Shapley Additive Explanations) are widely used because they offer mathematically grounded, consistent explanations that fairly attribute importance to each feature. SHAP values provide a stable framework that does not depend heavily on model type, allowing public agencies to explain decisions even when using complex or hybrid systems. In contrast, methods that produce inconsistent or case-specific explanations may undermine trust or expose systems to legal challenges.

In high-stakes settings, there is also a demand for counterfactual explanations, which are particularly valued because they answer the question: "What would need to change for a different outcome to occur?" Counterfactuals are useful in domains such as welfare eligibility, loan approval, or parole recommendations because they give actionable guidance to individuals impacted by decisions. Their suitability lies in the fact that they empower citizens by clearly illustrating how decisions were made and how outcomes could be altered, supporting rights to explanation and procedural fairness.

Moreover, high-stakes decisions require methods that can be audited and validated by oversight bodies. Techniques that produce stable, repeatable explanations such as SHAP, global surrogate models, and interpretable machine learning approaches facilitate external review by regulatory agencies or ethics boards[26]. Public-sector environments often operate under strict audit requirements; thus, explainability methods must create documentation that can withstand scrutiny, appeals, or court examination. XAI techniques that provide transparent decision traces or clear feature relationships are better suited for this purpose than opaque or highly technical methods that only expert data scientists can interpret.

Finally, certain XAI methods are preferred due to their alignment with ethical and legal obligations. Public-sector algorithms must comply with fairness standards, due process, anti-discrimination laws, and citizens' rights to an explanation. Intrinsically interpretable models such as logistic regression with clear coefficients, decision trees, or generalized additive models (GAMs) make it easier to identify and remove discriminatory factors. Although these models may sometimes lack the predictive power of deep learning or ensemble systems, they offer transparency that is indispensable for high-stakes public decision making where individual rights and legal compliance are paramount.

Limitations

Although this systematic literature review provides valuable insights into the use of Explainable AI (XAI) in public-sector decision making, several limitations should be acknowledged. First, the scope of the review is inherently limited by the predefined research objectives and inclusion criteria[27]. The

review focuses specifically on studies that combine XAI with public-sector contexts, which means that a large body of purely technical XAI research or AI governance studies without explicit explainability components was excluded. As a result, the review may not capture all relevant innovations that could indirectly influence the development of explainable systems for government use.

Another important limitation concerns database coverage. While major academic databases such as Scopus, Web of Science, IEEE Xplore, SpringerLink, ACM Digital Library, and Google Scholar were searched, these sources may not include all relevant publications, especially emerging works in niche or interdisciplinary fields such as public policy, digital governance, and administrative law. Some government reports, white papers, and practitioner-oriented publications were excluded because they do not meet peer-review standards, yet these documents often contain valuable real-world insights. Consequently, the findings may underrepresent practical implementations occurring outside academic publishing channels.

The review is also constrained by language restrictions, as it includes only studies published in English[28]. This choice ensures consistency and accessibility, but it may exclude significant research conducted in non-English-speaking regions where AI governance and public administration practices are rapidly evolving. Countries in Asia, Latin America, and Europe, for example, may have government-led research or policy documents on XAI that are not captured due to language barriers. This limitation reduces the global representativeness of the review and may skew the findings toward Western-centric perspectives.

Additionally, by restricting the dataset to peer-reviewed sources, the review prioritizes academic rigor but may overlook practical insights from industry projects, pilot programs, and non-peer-reviewed government initiatives. Many real-world XAI applications in public services such as policing, taxation, or social welfare are documented in government reports, technical documentation, or internal evaluations rather than academic journals[29]. The exclusion of such materials limits the review's ability to fully capture the state of practice in public-sector XAI deployment.

4. Conclusion

This systematic review identifies several dominant XAI techniques used in public-sector decision making, including SHAP, LIME, decision trees, rule-based models, feature-importance methods, and counterfactual explanations. These techniques support government decision making by enhancing transparency, improving accountability, enabling bias detection, and strengthening public trust in algorithm-assisted policies and services. However, persistent challenges remain, particularly the technical complexity of XAI methods, the trade-off between accuracy and interpretability, limited AI literacy among officials, and the lack of standardized frameworks tailored for public governance. Future research should focus on developing domain-specific XAI guidelines, creating more user-friendly explanation tools for non-technical stakeholders, expanding evaluation methods, and exploring how explainability can be integrated into ethical, legal, and policy frameworks to support trustworthy AI in government.

References

[1] J. G. Corvalán, "Digital and intelligent public administration: Transformations in the era of artificial intelligence," *A&C-Revista Direito Adm. Const.*, vol. 18, no. 71, pp. 55–87, 2018.

[2] B. Lepri, N. Oliver, E. Letouzé, A. Pentland, and P. Vinck, "Fair, transparent, and accountable algorithmic decision-making processes: The premise, the proposed solutions, and the open challenges," *Philos. Technol.*, vol. 31, pp. 611–627, 2018.

[3] D. Anny, "Regulatory Frameworks for AI Bias: A Comparative Analysis of Global Approaches," 2020.

[4] D. Martinez-Mosquera, R. Navarrete, and S. Lujan-Mora, "Modeling and management big data in databases—A systematic literature review," *Sustainability*, vol. 12, no. 2, p. 634, 2020.

[5] A. Anguita-Ruiz, A. Segura-Delgado, R. Alcalá, C. M. Aguilera, and J. Alcalá-Fdez, "eXplainable Artificial Intelligence (XAI) for the identification of biologically relevant gene expression patterns in longitudinal human studies, insights from obesity research," *PLoS Comput. Biol.*, vol. 16, no. 4, p. e1007792, 2020.

[6] D. Arduini and A. Zanfei, "An overview of scholarly research on public e-services? A meta-analysis of the literature," *Telecomm. Policy*, vol. 38, no. 5–6, pp. 476–495, 2014.
[7] M. Kuziemski and G. Misuraca, "AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings," *Telecomm. Policy*, vol. 44, no. 6, p. 101976, 2020.
[8] H. P. Olsen, J. L. Slosser, T. T. Hildebrandt, and C. Wiesener, "What's in the box? The legal requirement of explainability in computationally aided decision-making in public administration," 2019.
[9] Y. Kwon, M. Lemieux, J. McTavish, and N. Wathen, "Identifying and removing duplicate records from systematic review searches," *J. Med. Libr. Assoc. JMLA*, vol. 103, no. 4, p. 184, 2015.
[10] E. Tjoa and C. Guan, "A survey on explainable artificial intelligence (xai): Toward medical xai," *IEEE Trans. neural networks Learn. Syst.*, vol. 32, no. 11, pp. 4793–4813, 2020.
[11] J. Allotey *et al.*, "Clinical manifestations, risk factors, and maternal and perinatal outcomes of coronavirus disease 2019 in pregnancy: living systematic review and meta-analysis," *bmj*, vol. 370, 2020.
[12] Z. Liu, D. Rexachs, F. Epelde, and E. Luque, "An agent-based model for quantitatively analyzing and predicting the complex behavior of emergency departments," *J. Comput. Sci.*, vol. 21, pp. 11–23, 2017.
[13] D. Slack, S. Hilgard, E. Jia, S. Singh, and H. Lakkaraju, "Fooling lime and shap: Adversarial attacks on post hoc explanation methods," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 180–186.
[14] M. Veale, M. Van Kleek, and R. Binns, "Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making," in *Proceedings of the 2018 chi conference on human factors in computing systems*, 2018, pp. 1–14.
[15] A. S. Adly, A. S. Adly, and M. S. Adly, "Approaches based on artificial intelligence and the internet of intelligent things to prevent the spread of COVID-19: scoping review," *J. Med. Internet Res.*, vol. 22, no. 8, p. e19104, 2020.
[16] Y. Kai, "The Rules and Methods of Incidental Review on Documents of Administrative Norms," *China Leg. Sci.*, vol. 5, p. 146, 2017.
[17] H. Felzmann, E. Fosch-Villaronga, C. Lutz, and A. Tamò-Larrieux, "Towards transparency by design for artificial intelligence," *Sci. Eng. Ethics*, vol. 26, no. 6, pp. 3333–3361, 2020.
[18] N. Helberger, T. Araujo, and C. H. De Vreese, "Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making," *Comput. Law Secur. Rev.*, vol. 39, p. 105456, 2020.
[19] S. J. Mikhaylov, M. Esteve, and A. Campion, "Artificial intelligence for the public sector: opportunities and challenges of cross-sector collaboration," *Philos. Trans. R. Soc. a Math. Phys. Eng. Sci.*, vol. 376, no. 2128, p. 20170357, 2018.
[20] J. Gerlings, A. Shollo, and I. Constantiou, "Reviewing the need for explainable artificial intelligence (xAI)," *arXiv Prepr. arXiv2012.01007*, 2020.
[21] M. Dubnick and H. G. Frederickson, "Public accountability: Performance measurement, the extended state, and the search for trust," *Natl. Acad. Public Adm. Kettering Found.*, 2011.
[22] S. Goel, M. Manuja, R. Dwivedi, and A. M. Sherry, "Challenges of technology infrastructure availability in e-governance program implementations: A cloud based solution," *J. Comput. Eng.*, vol. 5, no. 2, pp. 13–17, 2012.
[23] A. Abdul, J. Vermeulen, D. Wang, B. Y. Lim, and M. Kankanhalli, "Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda," in *Proceedings of the 2018 CHI conference on human factors in computing systems*, 2018, pp. 1–18.
[24] A. Adadi and M. Berrada, "Peeking inside the black-box: a survey on explainable artificial intelligence (XAI)," *IEEE access*, vol. 6, pp. 52138–52160, 2018.
[25] A. Rosenfeld and A. Richardson, "Explainability in human–agent systems," *Auton. Agent. Multi. Agent. Syst.*, vol. 33, no. 6, pp. 673–705, 2019.
[26] K. Azi, "Explainable AI (XAI): Interpretability in Machine Learning Models," *Int. J. Artif. Intell. Mach. Learn.*, vol. 1, no. 2, 2018.
[27] J. Thomas, D. Kneale, J. E. McKenzie, S. E. Brennan, and S. Bhaumik, "Determining the scope of the review and the questions it will address," *Cochrane Handb. Syst. Rev. Interv.*, pp. 13–31, 2019.
[28] A. Morrison *et al.*, "The effect of English-language restriction on systematic review-based meta-analyses: a systematic review of empirical studies," *Int. J. Technol. Assess. Health Care*, vol. 28, no. 2, pp. 138–144, 2012.
[29] P. Henman, "Improving public services using artificial intelligence: possibilities, pitfalls, governance," *Asia Pacific J. Public Adm.*, vol. 42, no. 4, pp. 209–221, 2020.