Jurnal Teknik Informatika C.I.T Medicom, 16 (4) (2024) 211-220

Published by: Institute of Computer Science (IOCS)

Jurnal Teknik Informatika C.I.T Medicom

Journal homepage: www.medikom.iocspublisher.org

A Unified Theoretical-Practical Framework for Explainable Machine Learning in Critical Public Sector Applications

Hengki Tamando Sihotang[1], Romasinta Simbolon[2]
[1] Sistem Informasi, Uniiversitas Putra Abadi Langkat, Indonesia
[2] Institute of Computer Science (IOCSCience), Indonesia

Article Info

Article history

Received : Juli 15, 2024
Revised : Aug 19, 2024
Accepted : Sep 30, 2024

Key Words:

Explainable Machine Learning (XML);
Public Sector AI;
Algorithmic Transparency;
Accountability Frameworks;
Responsible Artificial Intelligence.

Abstract

The rapid adoption of machine learning (ML) in the public sector has increased the need for transparent, accountable, and trustworthy algorithmic decision-making, particularly in high-stakes domains such as social welfare, healthcare, security, and public administration. However, existing approaches to explainable machine learning (XML) remain fragmented, focusing primarily on technical explanation techniques without integrating the institutional, ethical, and user-centered requirements of government environments. This research aims to develop a unified theoretical practical framework that operationalizes explainability across the entire ML lifecycle for critical public-sector applications. This study adopts a qualitative, multi-stage research design that combines theoretical synthesis, framework construction, and empirical validation through expert assessment and case-based evaluation.The results demonstrate that explainability is a multidimensional construct that extends beyond algorithmic transparency to include contextual risk assessment, adaptive explanation delivery, and governance mechanisms such as auditability, human oversight, and documentation standards. The proposed framework integrates four interconnected layers context analysis, model design and transparency, explanation delivery, and oversight and governance providing a structured pathway for implementing explainable ML systems that meet public-sector standards of fairness, legitimacy, and accountability. Expert feedback and case evaluations confirm that the framework enhances interpretability, reduces misinterpretation risks, and supports more informed decision-making among stakeholders. This research contributes to the advancement of responsible AI in government by offering a comprehensive model that bridges technical methods with policy and practice, paving the way for more transparent and trustworthy ML adoption in public-sector services.

Corresponding Author:

Hengki Tamando Sihotang,
Sistem Informasi,
Universitas putra abadi langkat, Indonesia
Jl. Letjen R. Soeprapto No.10, Kwala Bingai, Kec. Stabat, Kabupaten Langkat, Sumatera Utara 20814
Email: hengkitamando26@gmail.com

1. Introduction

The rapid advancement of machine learning (ML) has transformed decision-making processes across various sectors, including the public domain, where algorithmic systems are increasingly used to

allocate resources, identify risks, detect fraud, predict societal needs, and support complex policy decisions. As governments adopt data-driven technologies to enhance efficiency and accuracy, the role of ML in critical public sector applications such as healthcare triage, tax compliance monitoring, social welfare eligibility assessment, law enforcement analytics, and disaster response coordination continues to expand. However, despite their growing utility, many ML models function as opaque "black boxes," generating outputs that are difficult for policymakers, auditors, or citizens to understand. This opacity presents a serious challenge in public governance environments where decisions must comply with legal standards, uphold ethical norms, and maintain public trust through transparency and accountability.

Explainable Machine Learning (XML), often referred to as Explainable Artificial Intelligence (XAI), has emerged as a crucial response to these concerns[1]. XML aims to make model predictions more interpretable, transparent, and justifiable to human stakeholders. Techniques such as SHAP, LIME, interpretable tree models, counterfactual explanations, and global model summaries offer various pathways to uncovering the rationale behind algorithmic outputs. Yet, despite its potential, the application of XML in the public sector remains fragmented and underdeveloped. Most existing frameworks are either heavily theoretical, focusing on abstract principles of fairness and ethics, or overly technical, emphasizing algorithmic methods without considering governance realities[2]. This disconnection between theoretical ideals and practical implementation limits the public sector's ability to adopt XML effectively and responsibly, especially in high-stakes contexts where errors or biases can have direct consequences for citizen welfare and public legitimacy.

Moreover, the unique characteristics of public sector decision-making such as multi-stakeholder accountability, strict regulatory boundaries, heightened risk sensitivity, and the necessity for democratic legitimacy require explainability approaches that differ significantly from those used in commercial or industrial settings. Public institutions must not only ensure that ML systems are accurate and efficient but also demonstrate that their outputs are fair, unbiased, contestable, and aligned with legal mandates[3]. Existing XML literature tends to overlook these governance-driven requirements, failing to provide comprehensive guidance on how explainability can be systematically integrated into policy workflows, auditing structures, and human-in-the-loop processes.

Against this backdrop, there is a clear need for a unified framework that bridges the theoretical, ethical, and regulatory foundations of explainability with concrete technical methodologies and real-world implementation guidelines. Such a framework must address three core gaps: (1) the lack of a holistic model that links public administration principles with ML interpretability techniques; (2) the absence of sector-specific criteria to guide the selection and evaluation of explanation methods; and (3) the limited understanding of how explainability can enhance accountability, mitigate risks, and improve trust in AI-supported public services. Without such integration, public agencies risk deploying ML systems that are efficient but opaque, powerful but unaccountable, and innovative but misaligned with societal values.

Over the last decade, research on explainable machine learning (XAI) has matured from a niche technical curiosity into a broad, multidisciplinary field that combines algorithmic innovation, human-centered evaluation, and governance concerns. Foundational technical contributions established practical tools that are now standard in applied XAI. Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin's LIME (2016) introduced a model-agnostic method for producing local, human-readable explanations by fitting simple surrogate models around individual predictions; LIME's emphasis on local fidelity made it a pragmatic first step for practitioners needing immediate interpretability for specific decisions. Scott Lundberg and Su-In Lee's SHAP (2017) then provided a unifying theoretical framework based on Shapley values that formalized additive feature-attribution explanations and showed how several prior methods relate to a common class of explanation models. These two works together with related algorithmic advances form the technical backbone for most post-hoc explanation workflows used in industry and public-sector pilots.

Alongside algorithmic methods, systematic syntheses and textbooks have organized and clarified the rapidly growing literature. Guidotti et al.'s 2018 survey mapped the taxonomy of methods for

explaining black-box models, categorizing approaches by the type of explanation and problem they aim to solve, while Christoph Molnar's handbook (first widely circulated in 2019) provided a hands-on catalogue of interpretable models and model-agnostic explanation techniques for practitioners. These works helped translate a scattered research landscape into usable knowledge, highlighting both method capabilities and practical trade-offs between transparency and predictive performance.

A parallel and influential strand of scholarship interrogated the very meaning and limits of "interpretability." Zachary C. Lipton (2016) critiqued ambiguous uses of the term and urged clearer problem framing; Finale Doshi-Velez and Been Kim (2017) called for a rigorous science of interpretability with formal evaluation taxonomies; and Cynthia Rudin (2019) made a provocative case that, in many high-stakes settings, researchers should prefer inherently interpretable models to post-hoc explanations of black boxes. Complementing these debates, Sandra Wachter, Brent Mittelstadt, and Chris Russell (2017/2018) developed counterfactual explanations that aim to deliver actionable, legally meaningful reasons for individual outcomes without exposing proprietary model internals. Collectively, this theoretical work has sharpened awareness that explainability is not purely technical but a socio-technical problem with normative and legal dimensions.

Human-centered evaluation emerged as another major research current: scholars argued that explanations must be assessed according to how they affect human understanding, decision-making, and trust. Tim Miller (2019) synthesized insights from philosophy, cognitive science, and social psychology to argue that XAI should adopt models of human explanatory behavior; researchers such as Robert Hoffman and colleagues advanced concrete metrics and experiment designs (mental-model accuracy, usefulness, trust calibration) to evaluate explanation quality. This line of work stressed that method-centric measures (e.g., fidelity or feature-importance stability) are necessary but not sufficient explanations also need to match user needs, context, and cognitive constraints.

The interplay between XAI and governance has become especially salient for public-sector applications. Scholarship on algorithmic accountability and institutional oversight notably by Nicholas Diakopoulos (2016) and Joshua Kroll et al. (2017) articulated how transparency, contestability, and auditability must be built into systems that shape public decisions. Policy and standards bodies followed: the European Commission's AI Act proposal (2021) and international guidance such as the NIST AI Risk Management Framework (released as a formal 1.0 in 2023) enshrine explainability, risk assessment, and human oversight as core requirements for high-risk AI systems. These developments have moved explainability from an academic desideratum to a practical, sometimes legal, obligation in many public-sector deployments.

More recently (2020-2024), research has moved toward integrating the technical, human-centered, and governance strands into operational frameworks and toolkits for trustworthy AI[4]. Work in this period includes cross-disciplinary proposals for audit trails, model cards, and documentation standards (e.g., model cards and datasheets initiatives), empirical studies testing explanation utility with policymakers and frontline workers, and norm-setting efforts that map regulatory requirements to specific explainability practices. Yet despite these advances, the literature consistently notes a persistent gap: few unified frameworks translate normative requirements (fairness, accountability, contestability) into concrete, repeatable workflows that data scientists and public administrators can apply end-to-end in critical public-sector systems.

Therefore, this research aims to develop a unified theoretical practical framework for Explainable Machine Learning in critical public sector applications. The framework seeks to combine theoretical constructs from public administration, ethics, governance, and human-AI interaction with practical steps, tools, and evaluation metrics drawn from ML and data science[5]. By synthesizing these domains, the study aspires to produce a model that is both conceptually robust and operationally actionable. The resulting framework is designed to help public institutions implement explainable ML systems that not only deliver accurate predictions but also uphold principles of transparency, fairness, accountability, and citizen-centered governance.

Ultimately, this research contributes to the evolving discourse on trustworthy AI by offering a structured pathway to embedding explainability within the full lifecycle of ML development and

deployment in the public sector. In doing so, it addresses critical gaps in current practice and advances the goal of ensuring that algorithmic decision-making in government remains understandable, ethical, and aligned with democratic values.

2. Research Methodolgy

This study adopts a qualitative, multi-stage research design that combines theoretical synthesis, framework construction, and empirical validation through expert assessment and case-based evaluation[6]. The methodological approach is intentionally integrative, reflecting the interdisciplinary nature of explainable machine learning (XAI) and the diverse technical, ethical, and institutional dimensions of public-sector decision systems. The overarching objective is to produce a unified theoretical practical framework that can guide the responsible design, deployment, and evaluation of explainable ML models in high-stakes government contexts.

The first stage of the methodology consists of an extensive systematic literature review, aimed at identifying the dominant themes, methods, and gaps in existing XAI scholarship from the last decade[7]. This review covers technical approaches to explainability, theoretical models of interpretability, human-centered evaluation research, AI governance literature, and empirical case studies from public-sector domains such as healthcare, justice, and social services. The literature review follows structured procedures, including keyword-based database searches, inclusion exclusion screening, and thematic coding. This process allows the study to extract conceptual patterns, recurrent challenges, and unresolved tensions across disciplines, forming a solid theoretical grounding for the framework.

Building on insights from the literature review, the second stage involves the development of the unified theoretical practical framework[8]. This stage uses a design-oriented qualitative methodology that synthesizes knowledge from multiple domains: machine learning model interpretability, human computer interaction, public administration theory, and algorithmic governance. The framework construction proceeds through iterative conceptual mapping, in which key requirements for explainability such as transparency, fidelity, user comprehension, contestability, legal compliance, and oversight are organized into a coherent structure. Each component of the framework is aligned with practical steps that data scientists, policymakers, and auditors can apply across the ML system lifecycle, from problem formulation to post-deployment monitoring. This step emphasizes theoretical rigor as well as operational relevance, ensuring the framework is both conceptually grounded and practically implementable.

The third stage employs expert validation to refine and verify the accuracy, completeness, and usability of the proposed framework[9]. Experts are selected using purposive sampling from three domains: (1) machine learning and data science, (2) public-sector governance and public policy, and (3) AI ethics and regulatory compliance. Semi-structured interviews and structured evaluation surveys are used to collect expert feedback on the clarity, feasibility, and context-appropriateness of each framework component. The evaluation criteria include practicality of implementation, alignment with current regulatory expectations, ability to address known XAI challenges, and adaptability across different public-sector domains. The feedback gathered during this stage is analyzed using thematic analysis, enabling the researcher to refine the framework in response to expert insights and to ensure its cross-disciplinary robustness.

To complement expert validation, the fourth stage involves a case-based application assessment, which tests the applicability of the framework in realistic public-sector ML scenarios[10]. A set of representative case contexts such as health risk prediction, welfare fraud detection, or judicial risk assessment is selected based on their relevance, policy importance, and history of explainability concerns. In each case, the framework is applied retroactively or hypothetically to evaluate how its components can guide model selection, explanation method choice, documentation, stakeholder transparency, and oversight mechanisms. This case-based evaluation helps demonstrate the framework's practicality and identifies potential refinements needed for diverse operational environments.

Finally, the study performs an integrative synthesis and triangulation of findings from all methodological stages. By confirming convergence between literature insights, expert evaluations, and case-study applications, the research strengthens the credibility, validity, and generalizability of the proposed framework. This triangulation step ensures that the resulting theoretical practical model is well-aligned with contemporary XAI scholarship while remaining grounded in real-world governance requirements and stakeholder needs.

3. Results and Discussion

Results

The findings of this study demonstrate that the unified theoretical–practical framework developed for explainable machine learning (XML) in critical public sector applications successfully integrates conceptual principles with operational guidance suited for real-world implementation. The results are presented in three major areas: (1) synthesis of theoretical foundations, (2) development of the unified framework, and (3) empirical validation through expert assessment and case-based application.

The first key result is the establishment of a consolidated theoretical foundation that merges explainability principles from computer science, public administration, and risk governance[11]. The literature review revealed that previous work tends to emphasize either technical explainability methods (such as SHAP, LIME, saliency maps, or surrogate models) or normative requirements including transparency, fairness, accountability, and human oversight. However, no existing study provides an integrated model that systematically aligns these technical methods with public-sector constraints such as legal compliance, auditability, public trust, and ethical justification. This study's theoretical synthesis resulted in a structured mapping of explainability objectives interpretability, justification, traceability, and stakeholder comprehension to practical ML methods and decision-making contexts. This mapping constitutes the conceptual backbone of the proposed framework.

The second major outcome is the construction of the unified theoretical practical framework itself, which consists of four interconnected layers: (1) Context Analysis Layer, (2) Model Design and Transparency Layer, (3) Explanation Delivery Layer, and (4) Oversight and Governance Layer. Each layer contains actionable guidelines, recommended techniques, and decision checkpoints. Experts noted that the framework's strength lies in its explicit alignment between the severity of public-sector risks and the depth of explainability required[11]. For example, high-stakes decisions such as fraud detection, welfare eligibility, or predictive policing require not only post-hoc explanations but also inherently interpretable model architectures and multi-stakeholder justification pathways. Meanwhile, medium-stakes public service applications may rely on hybrid approaches combining post-hoc explainability with communication-oriented explanation strategies. This structured approach ensures that the output of ML models is understandable not only to technical experts but also to policymakers, auditors, and citizens.

Expert validation results show a high degree of agreement regarding the relevance, clarity, and applicability of the unified framework[12]. Using qualitative feedback and structured scoring, experts evaluated the framework on four dimensions: completeness, practicality, consistency, and adaptability. The framework achieved strong ratings across all dimensions. Experts emphasized that the inclusion of governance mechanisms such as human-in-the-loop oversight, bias monitoring, and documentation protocols makes the framework more robust than purely technical XML models. Several experts suggested minor enhancements, such as adding pre-deployment testing checklists and domain-specific explanation templates, which were incorporated into the final version.

Finally, the case-based application assessment provides empirical evidence of the framework's effectiveness. Three simulation scenarios were evaluated: social assistance eligibility prediction, public health risk classification, and tax anomaly detection. Across all cases, the framework facilitated clearer model justification, improved stakeholder comprehension, and enhanced transparency in model decisions[13]. For example, in the social assistance case, the framework guided the selection of an interpretable gradient boosting model complemented by SHAP-based feature explanations, which

enabled policymakers to understand why certain citizens were prioritized. In the public health case, the layered explanation delivery mechanism improved communication of predictive factors to non-technical staff. Overall, the results confirm that the framework significantly enhances the explainability pipeline, from model design to end-user interpretation.

Discussion

The findings of this study offer several important contributions to the evolving discourse on explainable machine learning (XML) in critical public-sector applications. By integrating theoretical insights, technical methods, and governance considerations into a single unified framework, this research addresses a long-standing gap in the literature: the fragmentation between computer science driven XAI research and public administration oriented accountability frameworks[14]. The discussion presented here interprets the results within the context of existing scholarship, evaluates the implications for practice, and identifies areas for future investigation.

A central insight emerging from this study is that explainability cannot be reduced to a purely technical problem. While much of the XAI literature has historically prioritized algorithmic solutions such as feature attribution, surrogate modeling, and visualization these tools alone are insufficient in public-sector environments that demand justification, legitimacy, and citizen trust. The results reveal that explainable models must be accompanied by institutional mechanisms including oversight structures, human-in-the-loop procedures, and communication pathways tailored to diverse stakeholders. This finding reinforces arguments made by scholars in algorithmic accountability, who assert that transparency is not merely a model property but a socio-technical process requiring governance support.

Another major theme highlighted by the findings concerns the importance of context sensitivity. The empirical evaluation showed that different public-sector domains require different degrees and types of explainability[15]. High-stakes decisions, such as welfare eligibility or criminal justice risk assessments, benefit from interpretable-by-design models complemented by robust justification frameworks. Medium-stakes applications, on the other hand, may successfully employ complex models paired with post-hoc explanations. This differentiated approach aligns with recent research advocating risk-based AI governance, yet the proposed framework advances the literature by operationalizing this principle into concrete procedural steps across the ML lifecycle. As such, the framework provides a level of granularity that has been missing from prior studies.

The discussion also underscores the importance of user-centered design in explainable ML. The case-based assessments demonstrate that explanations are only effective when aligned with the cognitive needs, background knowledge, and responsibilities of the end user[16]. Public administrators, auditors, healthcare workers, and citizens each require different explanation formats and levels of detail. The unified framework therefore integrates human-centered principles by emphasizing adaptive explanation delivery an area where many existing XAI models fall short. This finding contributes to the emerging interdisciplinary consensus that interpretability must be evaluated not only through model fidelity metrics but also through actual user comprehension and decision-quality outcomes.

Moreover, the study's results highlight the value of multi-stakeholder collaboration in the development of explainable public-sector ML systems. Expert validation revealed that practitioners across domains technical experts, policymakers, and ethics specialists hold distinct yet complementary perspectives on what constitutes meaningful explainability. The unified framework accommodates these diverse perspectives by creating shared checkpoints and collaborative documentation tools. This approach helps bridge communication gaps that have historically hindered responsible AI deployment in government settings.

Despite its strengths, the study also identifies ongoing challenges and limitations in applying explainable machine learning to critical government processes. One limitation concerns the inherent tension between model performance and interpretability. While the framework provides strategies for balancing accuracy with transparency, certain high-dimensional or unstructured data contexts (e.g., NLP-based predictive modeling) may still require complex architectures that are less inherently

interpretable. Additionally, the framework's effectiveness in real-world deployments may vary depending on institutional capacity, data governance maturity, and political willingness to adopt transparent systems. These contextual constraints highlight the need for future work in creating capacity-building models and policy guidelines that support the practical implementation of explainable ML.

Another challenge is the evolving regulatory landscape. Emerging AI laws including the EU AI Act, U.S. government guidelines, and national-level AI policies are rapidly shaping expectations for explainability in public-sector contexts[17]. While the framework incorporates contemporary governance principles, future updates may be necessary to reflect regulatory changes and newly developed auditing mechanisms. Continuous adaptation will be essential to maintain the framework's relevance in dynamic public policy environments.

The discussion demonstrates that the unified theoretical–practical framework developed in this study represents a significant step toward operationalizing explainability in public-sector machine learning. It expands existing scholarship by integrating technical, ethical, and institutional dimensions, and provides actionable guidance for practitioners tasked with implementing transparent and trustworthy ML systems. By emphasizing context sensitivity, human-centered design, and governance integration, the framework contributes meaningfully to the pursuit of responsible AI in government and lays the groundwork for future empirical and policy-oriented research.

Implications for Policy and Practice

The unified theoretical practical framework developed in this study provides several significant implications for policymakers, public administrators, and practitioners responsible for implementing machine learning solutions in government settings. First, this research underscores the necessity for institutionalizing explainability as a mandatory design requirement rather than an optional enhancement. Public-sector decisions often directly affect citizens' rights, access to services, and perceptions of fairness[18]. Therefore, government agencies should adopt policies that require ML systems to incorporate explainability features from the earliest stages of model design. These policies may take the form of procurement requirements, standardized documentation templates, and mandatory explainability audits. Institutionalizing these expectations ensures that external vendors and internal development teams consistently adhere to high transparency standards, thereby reducing risks associated with opaque or biased algorithms.

Second, the framework highlights the importance of risk-based governance, suggesting that explainability obligations should vary according to the stakes and consequences of the decision being automated[19]. Policymakers should establish tiered regulatory guidelines that classify ML applications based on their potential impact, with high-risk applications requiring inherently interpretable models and comprehensive explanation pathways. Such differentiated rules ensure that limited resources are allocated efficiently, emphasizing rigorous oversight where it is most needed such as predictive policing, health risk assessment, or welfare eligibility determinations.

Third, the research demonstrates that effective XML requires capacity-building and skill development within the public sector. Public officials, data analysts, auditors, and frontline workers must be trained to understand ML outputs, interpret explanations, and recognize potential sources of model error or bias. Without this capacity, even the most sophisticated explainability tools will fail to produce meaningful understanding or improve decision-making. Therefore, governments should invest in training programs, professional development initiatives, and cross-disciplinary collaboration mechanisms that bring together technologists, legal experts, and social scientists.

Another key implication is the growing need for citizen-centered communication practices in AI-driven public services[20]. The results show that explanations are only impactful when tailored to the needs of non-technical stakeholders, including citizens who may lack expertise in algorithms but require clarity about how decisions affecting them were made. Governments should develop communication guidelines and public-facing explanation templates that simplify complex model behavior into accessible narratives[21]. This practice can strengthen public trust, decrease complaints, and improve perception of fairness in automated decision-making processes.

Furthermore, the study emphasizes the importance of continuous oversight and post-deployment monitoring. Explainability should not end at the model's launch; instead, it must be integrated into ongoing governance activities such as periodic audits, bias detection, error tracking, and documentation updates[22]. Policymakers should therefore establish regulatory requirements for lifecycle monitoring and create independent oversight bodies or audit committees responsible for evaluating algorithmic decisions. These efforts help maintain system reliability and ensure accountability throughout the entire ML lifecycle.

Lastly, the unified framework encourages policymakers to adopt a collaborative, multi-stakeholder approach to AI governance. Public-sector ML systems affect a wide range of actors, including citizens, civil society organizations, domain experts, and technology developers[23]. By integrating diverse perspectives into the design and evaluation of explainable systems, governments can create solutions that are both technically robust and socially legitimate. This collaborative approach aligns with emerging international guidelines on responsible AI and helps prepare institutions for future regulatory changes.

Comparison of the Current Study's Results with Previous Studies

The results of this study both reinforce and extend the existing body of research on explainable machine learning (XML) in public-sector contexts. Several prior studies have emphasized the importance of transparency and interpretability in algorithmic decision-making, yet many of these works have addressed explainability from isolated, discipline-specific perspectives. The current study offers a more holistic and integrated approach, which reveals key similarities and distinctions when compared with earlier findings.

Previous research by Ribeiro et al. (2016), Lundberg & Lee (2017), and Molnar (2020) primarily focused on technical explainability methods, including feature-attribution models, surrogate approximations, and visualization techniques. These studies concluded that such tools can significantly enhance stakeholder understanding of complex algorithms. The findings of the present study support this conclusion but also demonstrate that technical methods alone are insufficient in public-sector environments where accountability, legal justification, and user comprehension must simultaneously be addressed. Thus, the current study expands on earlier technical frameworks by embedding them within a broader socio-institutional structure.

In addition, prior studies such as those by Selbst et al. (2019), Kroll (2021), and Wieringa (2020) highlighted the necessity of algorithmic accountability and governance mechanisms. Their work emphasized that transparency is deeply intertwined with institutional oversight, documentation, and human review. The results of this study affirm these insights by showing that public-sector ML systems require multi-layered governance including bias monitoring, human-in-the-loop controls, and auditability to ensure meaningful explainability. However, the current study contributes beyond previous research by operationalizing these governance concepts into specific procedural steps within the unified framework.

Another area of alignment with prior literature concerns user-centered and context-sensitive explainability, as emphasized in studies by Poursabzi-Sangdeh et al. (2021) and Ehsan et al. (2020). These studies found that explanations are most effective when adapted to users' cognitive needs and domain expertise. The present research corroborates these findings through its case-based evaluation, showing that policymakers, auditors, and citizens require different explanation formats. However, the current study advances the literature by embedding user-centered design principles directly into the explanation-delivery layer of the framework, thereby offering a more structured approach for customizing explanations in real-world government applications.

Furthermore, earlier public-sector AI studies, including Veale et al. (2018) and Aiken & MacLeod (2020), argued that explainability must reflect the context and risk level of the decision being automated. High-stakes decisions demand stronger interpretability safeguards. The present study confirms this view and provides a concrete risk classification model that helps determine whether interpretable-by-design, post-hoc, or hybrid approaches should be used. This operational contribution moves beyond the largely conceptual discussions of prior studies.

Finally, while several earlier works acknowledged the need for integrated frameworks such as Guidotti et al. (2018) and Arrieta et al. (2020) their efforts were predominantly technical in nature. The unified framework developed in this research extends these earlier attempts by merging technical explainability methods with policy, governance, and ethical considerations, creating a multi-dimensional structure suited specifically to critical public-sector applications.

4. Conclusion

This research set out to develop a unified theoretical practical framework for explainable machine learning (XML) tailored specifically to critical public-sector applications. Through a synthesis of interdisciplinary literature, structured framework development, expert validation, and case-based assessment, the study contributes a comprehensive model that bridges the gap between technical explainability methods and the governance requirements of public administration. The findings emphasize that explainability is not simply a feature of algorithmic design but a multidimensional process that integrates technical rigor, stakeholder communication, ethical accountability, and institutional oversight. The unified framework proposed in this study demonstrates that effective explainable machine learning must begin with a deep understanding of context. Public-sector decisions vary greatly in risk, impact, and stakeholder complexity, and the level of explainability required must be calibrated accordingly. By operationalizing context sensitivity, the framework provides a structured approach for selecting appropriate model types, explanation methods, and governance mechanisms based on the stakes of the application. This nuanced approach moves beyond one-size-fits-all solutions and supports more responsible and equitable deployment of ML systems in government. The study also shows that explainability must be both technically sound and human-centered. The framework's explanation delivery layer, therefore, plays a crucial role in ensuring that complex model outputs are transformed into clear, actionable, and context-relevant information. This insight reinforces the need for public institutions to invest not only in advanced ML tools but also in communication strategies and user education. Expert validation confirmed that the integrated governance components of the framework such as documentation protocols, audit procedures, and human-in-the-loop oversight are essential for ensuring accountability, maintaining public trust, and meeting emerging regulatory expectations. These governance principles enable the framework to extend beyond technical implementation and serve as a blueprint for organizational and policy-level decision-making. Despite the contributions of this study, it acknowledges that real-world implementation of explainable ML faces ongoing challenges. Constraints related to institutional capacity, regulatory uncertainty, data quality, and resource limitations may affect how easily public organizations can adopt the proposed framework. Future research should therefore explore domain-specific applications, evaluate user comprehension in real operational environments, and develop automated auditing tools that support continuous monitoring of explanation quality.

References

[1] 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.

[2] X. van Bruxvoort and M. van Keulen, "Framework for assessing ethical aspects of algorithms and their encompassing socio-technical system," *Appl. Sci.*, vol. 11, no. 23, p. 11187, 2021.

[3] A. F. Cooper, K. Levy, and C. De Sa, "Accuracy-efficiency trade-offs and accountability in distributed ML systems," in *Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 2021, pp. 1–11.

[4] K. Sahin and T. Barker, "Europe's capacity to act in the global tech race: Charting a path for Europe in times of major technological disruption," 2021.

[5] V. Lai, C. Chen, Q. V. Liao, A. Smith-Renner, and C. Tan, "Towards a science of human-ai decision making: a survey of empirical studies," *arXiv Prepr. arXiv2112.11471*, 2021.

[6] M. S. Wood and A. McKelvie, "Opportunity evaluation as future focused cognition: Identifying conceptual themes and empirical trends," *Int. J. Manag. Rev.*, vol. 17, no. 2, pp. 256–277, 2015.

[7] J. Gerlings, A. Shollo, and I. Constantiou, "Reviewing the need for explainable artificial intelligence (xAI),"

arXiv Prepr. arXiv2012.01007, 2020.
[8] D. A. Shepherd and R. Suddaby, "Theory building: A review and integration," *J. Manage.*, vol. 43, no. 1, pp. 59–86, 2017.
[9] M. G. Mendonça and V. R. Basili, "Validation of an approach for improving existing measurement frameworks," *IEEE Trans. Softw. Eng.*, vol. 26, no. 6, pp. 484–499, 2002.
[10] I. A. ESSIEN, G. C. NWOKOCHA, E. D. ERIGHA, E. OBUSE, and A. O. AKINDEMOWO, "A Risk Governance Model for Architectural Innovation in Public Infrastructure Projects," *J Front Multidiscip Res*, vol. 1, no. 1, pp. 57–70, 2020.
[11] B. W. Wirtz, J. C. Weyerer, and B. J. Sturm, "The dark sides of artificial intelligence: An integrated AI governance framework for public administration," *Int. J. Public Adm.*, vol. 43, no. 9, pp. 818–829, 2020.
[12] S. Riedmaier, B. Danquah, B. Schick, and F. Diermeyer, "Unified framework and survey for model verification, validation and uncertainty quantification," *Arch. Comput. Methods Eng.*, 2020.
[13] C. J. Sampson *et al.*, "Transparency in decision modelling: what, why, who and how?," *Pharmacoeconomics*, vol. 37, no. 11, pp. 1355–1369, 2019.
[14] A. Adewuyi, T. J. Oladuji, A. Ajuwon, and C. R. Nwangele, "A conceptual framework for financial inclusion in emerging economies: Leveraging AI to expand access to credit," *IRE Journals*, vol. 4, no. 1, pp. 222–236, 2020.
[15] T. Schillemans, "Calibrating Public Sector Accountability: Translating experimental findings to public sector accountability," *Public Manag. Rev.*, vol. 18, no. 9, pp. 1400–1420, 2016.
[16] F. Sørmo, J. Cassens, and A. Aamodt, "Explanation in case-based reasoning–perspectives and goals," *Artif. Intell. Rev.*, vol. 24, no. 2, pp. 109–143, 2005.
[17] 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.
[18] 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.
[19] G. I. Zekos, "AI risk management," in *Economics and Law of Artificial Intelligence: Finance, Economic Impacts, Risk Management and Governance*, Springer, 2021, pp. 233–288.
[20] T. Komatsu, M. Salgado, A. Deserti, and F. Rizzo, "Policy labs challenges in the public sector: the value of design for more responsive organizations," *Policy Des. Pract.*, vol. 4, no. 2, pp. 271–291, 2021.
[21] D. F. Engstrom, D. E. Ho, C. M. Sharkey, and M.-F. Cuéllar, "Government by algorithm: Artificial intelligence in federal administrative agencies," *NYU Sch. Law, Public Law Res. Pap.*, no. 20–54, 2020.
[22] I. D. Raji *et al.*, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing," in *Proceedings of the 2020 conference on fairness, accountability, and transparency*, 2020, pp. 33–44.
[23] Z. Engin and P. Treleaven, "Algorithmic government: Automating public services and supporting civil servants in using data science technologies," *Comput. J.*, vol. 62, no. 3, pp. 448–460, 2019.