American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Explainable Machine Learning for Transparent Public-Service Decision Systems

Dr. Lucia Bianchi

Professor, European Institute of Data and Governance Milan, Italy

Abstract

Machine learning is increasingly relevant to public-service administration, where predictive systems can assist with case prioritization, resource allocation, application screening, complaint routing, inspection planning, and service-delivery management. However, public-sector decisions differ from many commercial predictions because they may influence access to public benefits, administrative opportunities, essential services, or regulatory attention. Predictive accuracy alone is therefore insufficient when citizens and public officials cannot understand the principal factors influencing a recommendation. This study proposes an Explainable Public-Service Machine Learning Framework designed to combine predictive performance with transparent reasoning, reviewability, human oversight, and auditable decision records.

The proposed architecture distinguishes between technical model interpretation, officer-facing decision explanations, and citizen-facing reason communication. It incorporates interpretable modeling where feasible, feature-attribution analysis for complex models, confidence disclosure, standardized reason codes, lawful counterfactual information, human escalation, and decision logging. A design-science methodology was combined with structured evidence mapping and a reproducible scenario-based simulation. A total of 2,000 synthetic decision episodes were distributed across five public-service domains: social assistance, permit prioritization, grievance triage, housing maintenance, and public-health inspection.

An opaque machine-learning configuration was compared with the proposed explainable configuration using decision accuracy, a transparency index, and administrative review time. Simulation results showed comparable predictive accuracy, with 79.5% for the opaque configuration and 78.3% for the explainable configuration. In contrast, the mean transparency index increased from 43.3% to 90.9%, while modeled review time declined from 12.0 to 7.4 minutes. These results illustrate that substantial gains in decision transparency and review efficiency can be achieved without requiring large reductions in predictive performance.

The study concludes that public-service machine learning should be evaluated as an accountable decision-support infrastructure rather than solely as a predictive algorithm. Transparent public administration requires explanations that are technically meaningful, understandable to affected persons, reviewable by officials, and sufficiently documented to support correction, appeal, and institutional accountability.

Keywords: explainable machine learning; public-service decision systems; algorithmic transparency; interpretable machine learning; public administration; human oversight; decision accountability; SHAP; transparent governance; responsible artificial intelligence

1. Introduction

Machine-learning systems are increasingly capable of supporting activities that traditionally required substantial administrative review, including prioritizing applications, identifying high-risk cases, routing complaints, estimating service demand, allocating inspection resources, and assisting public officials with complex information. OECD analyses indicate that governments are adopting artificial intelligence to improve public-service delivery and administrative decision-making, while simultaneously emphasizing that accountability, transparency, and explainability are particularly important where technology contributes to public decisions.

The central difficulty is that many high-performing machine-learning architectures operate with limited inherent interpretability. Complex ensembles and neural systems can produce accurate classifications without presenting reasoning in a form that an administrator or affected citizen can readily understand. Ribeiro, Singh, and Guestrin addressed this problem through local interpretable explanations, while Lundberg and Lee introduced SHAP as a unified feature-attribution framework for explaining individual predictions. These approaches demonstrate that understanding why a prediction occurred can be as important as the numerical prediction itself when users must decide whether a model should be trusted.

Explainability has even greater significance in public administration because public decisions operate within institutional, legal, and democratic accountability structures. NIST identifies explainability, interpretability, transparency, accountability, fairness, reliability, and privacy among the characteristics relevant to trustworthy systems. OECD guidance likewise warns that insufficient transparency and explainability can weaken accountability in government applications. A citizen who receives an adverse recommendation should not encounter an administrative process in which neither the system operator nor reviewing officer can meaningfully identify the factors that shaped the outcome.

Nevertheless, explanation should not be reduced to displaying a technically impressive visualization. Doshi-Velez and Kim argue that interpretability requires rigorous evaluation rather than relying on vague assumptions about what constitutes a useful explanation. Miller's analysis of explanation research further suggests that explanations need to reflect how people actually understand reasons, including selective and contrastive reasoning. Rudin goes further by arguing that high-stakes decisions should often favor inherently interpretable models rather than relying automatically on post-hoc explanations of black-box systems.

2. Objectives of the Study

2.1 Development of a Transparent Decision Architecture

The first objective is to develop a machine-learning framework capable of producing decision-support recommendations together with understandable reasons, confidence information, and auditable records. The architecture is designed to support both public officials who require technical detail and citizens who require understandable explanations of consequential decisions.

2.2 Evaluation of Accuracy and Transparency

The second objective is to examine whether improving explanation and reviewability necessarily produces a substantial loss in predictive performance. The study therefore compares an opaque machine-learning configuration and an explainable configuration across several simulated public-service domains.

2.3 Strengthening Human Review and Accountability

The third objective is to incorporate meaningful human supervision into machine-assisted public administration. Predictions with low confidence, conflicting evidence, exceptional circumstances, or potentially consequential effects are designed to move into human review rather than being treated as automatically final administrative determinations.

3. Review of Literature

3.1 Interpretable and Explainable Machine Learning

Ribeiro, Singh, and Guestrin introduced LIME to explain individual classifier predictions by constructing an interpretable model around a specific prediction. Their work demonstrated that local explanation can help users decide whether individual predictions and broader models should be trusted.

Lundberg and Lee subsequently developed SHAP, drawing on Shapley values to assign feature-level contributions to individual predictions. SHAP became influential because it provides a unified framework for a broad class of additive feature-attribution methods. These approaches are particularly useful when predictive performance depends on complex models whose internal structures cannot easily be translated into plain-language explanations.

Arrieta and colleagues describe explainable artificial intelligence as a broad research area encompassing different interpretability objectives, techniques, and application contexts. Their analysis connects explainability with responsible system development while emphasizing that explanation requirements vary according to the intended user and context.

At the same time, explainability techniques require validation. Adebayo and colleagues showed that visually convincing explanations can fail important sanity checks, indicating that explanation quality cannot be determined merely by appearance.

3.2 Explainability in High-Stakes Decision Environments

Rudin questions the assumption that every complex model should first be optimized and subsequently explained through a separate post-hoc technique. For high-stakes applications, she argues that inherently interpretable models may often be preferable where they can achieve adequate predictive performance.

This distinction is particularly relevant to public administration. An explanation generated after a prediction is useful only when it faithfully represents the decision process. A technically sophisticated explanation that does not accurately reflect model behavior may create an appearance of transparency without genuine accountability.

Doshi-Velez and Kim therefore distinguish different approaches to evaluating interpretability, emphasizing the need for systematic measurement. Miller additionally demonstrates that human explanations are often contrastive: people frequently want to understand why one outcome occurred rather than an expected alternative.

For public services, this means a useful explanation may need to answer questions such as why an application was prioritized, why a case was classified as requiring additional review, or which permissible changes could affect a future outcome.

3.3 Transparency and Accountability in Public Services

Public-sector machine learning operates within a distinctive governance environment. OECD guidance on public-service design states that accountability, transparency, and explainability are crucial when automated systems contribute to decisions and that responsibility should be clearly assigned across the system lifecycle.

NIST similarly treats accountability and transparency, explainability and interpretability, fairness, privacy, reliability, safety, and security as interrelated characteristics of trustworthy systems.

Recent OECD work further observes that public institutions increasingly use machine-assisted decision processes and that auditing institutions will play an important role in examining how transparency, controls, and accountability are maintained.

The European regulatory environment has also strengthened transparency requirements for certain uses of artificial intelligence. The European Commission confirms that AI Act transparency provisions became applicable in August 2026 for covered systems and activities.

3.4 Identified Research Gap

Existing research provides advanced methods for model explanation and growing governance guidance for responsible technology use. However, a practical gap remains between model-level explanation and citizen-facing administrative transparency. Public-service systems require technical interpretability to be translated into standardized reasons, review procedures, confidence disclosure, human escalation, and auditable decision records.

4. Research Methodology

4.1 Research Design

The research adopted a design-science methodology supported by structured evidence mapping and scenario-based simulation. The principal artifact was an explainable decision-support architecture rather than a commercial software product.

The evidence review covered foundational and contemporary work on interpretable machine learning, explainability, trustworthy systems, public-sector governance, and administrative transparency. Key evidence sources included NIST, OECD, European Commission materials, and primary explainable-machine-learning research.

The quantitative component used synthetic scenarios because no administrative agency supplied citizen records or operational datasets. The results therefore represent methodological validation under controlled assumptions and not observed performance in an actual government department.

4.2 Proposed Explainable Public-Service Machine Learning Framework

The framework contains six functional layers:

- **Data and Eligibility Layer** – validates legally permissible and administratively relevant variables.
- Predictive Layer – applies an interpretable model where feasible or a validated complex model where necessary.
- Explanation Layer – generates feature contributions and standardized reason codes.
- Confidence Layer – indicates uncertainty associated with the recommendation.
- Human Review Layer – escalates uncertain, exceptional, or consequential cases.
- Audit and Citizen Communication Layer – records decisions and provides understandable explanations.

The framework distinguishes three explanation audiences:

Technical explanations support model developers and auditors.

Administrative explanations support public officials responsible for reviewing recommendations.

Citizen explanations communicate understandable reasons without unnecessary technical terminology.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

Table 1: Operational Components of the Proposed Framework

Component	Function	Transparency Contribution
Input validation	Checks relevance and permissible data use	Prevents unexplained dependence on inappropriate variables
Interpretable prediction	Produces understandable decision relationships	Improves model-level transparency
Feature attribution	Identifies influential variables	Supports case-specific explanation
Reason codes	Converts technical outputs into administrative language	Improves officer and citizen comprehension
Confidence information	Communicates prediction uncertainty	Prevents false appearance of certainty
Human escalation	Transfers exceptional cases for review	Preserves meaningful human judgment
Decision logging	Records inputs, model version and explanation	Supports auditability
Review and appeal support	Enables reconsideration of contested outcomes	Strengthens administrative accountability

4.3 Simulation Procedure

A reproducible synthetic benchmark containing 2,000 decision episodes was developed. A fixed random seed of 20260808 was used.

The simulated cases were divided equally across five public-service domains:

- Social assistance
- Permit prioritization
- Grievance triage
- Housing maintenance
- Public-health inspection

5. Analysis

5.1 Overall Performance

The opaque machine-learning configuration achieved 79.5% overall decision accuracy, while the explainable configuration achieved 78.3%.

The difference of 1.2 percentage points indicates a modest predictive-performance trade-off under the assumptions of the simulation. The result is important because transparent decision support should not be evaluated under the assumption that an explainable architecture must outperform a complex model in raw prediction accuracy.

A substantially different pattern emerged for transparency. The opaque configuration recorded a mean Transparency Index of 43.3%, compared with 90.9% for the explainable architecture.

Administrative review time declined from approximately 12.0 minutes to 7.4 minutes per modeled case, representing a reduction of approximately 38%.

5.2 Domain-Wise Comparative Results

Table 2: Simulated Performance across Public-Service Domains

Public-Service Domain	Opaque Accuracy (%)	Explainable Accuracy (%)	Opaque Transparency Index (%)	Explainable Transparency Index (%)	Opaque Review Time (min)	Explainable Review Time (min)
Social Assistance	79.5	78.8	42.4	90.1	13.0	8.1
Permit Prioritization	80.0	78.8	44.7	91.0	11.4	7.0
Grievance Triage	76.2	75.8	40.2	89.0	12.2	7.5
Housing Maintenance	80.8	79.8	43.6	91.8	10.7	6.8
Public Health Inspection	81.0	78.5	45.8	92.9	12.8	7.7
Overall	**79.5**	**78.3**	**43.3**	**90.9**	**12.0**	**7.4**

The results show that explainability gains were consistent across all five domains. The Transparency Index exceeded 88% in every explainable configuration, whereas the opaque configurations remained below 46%.

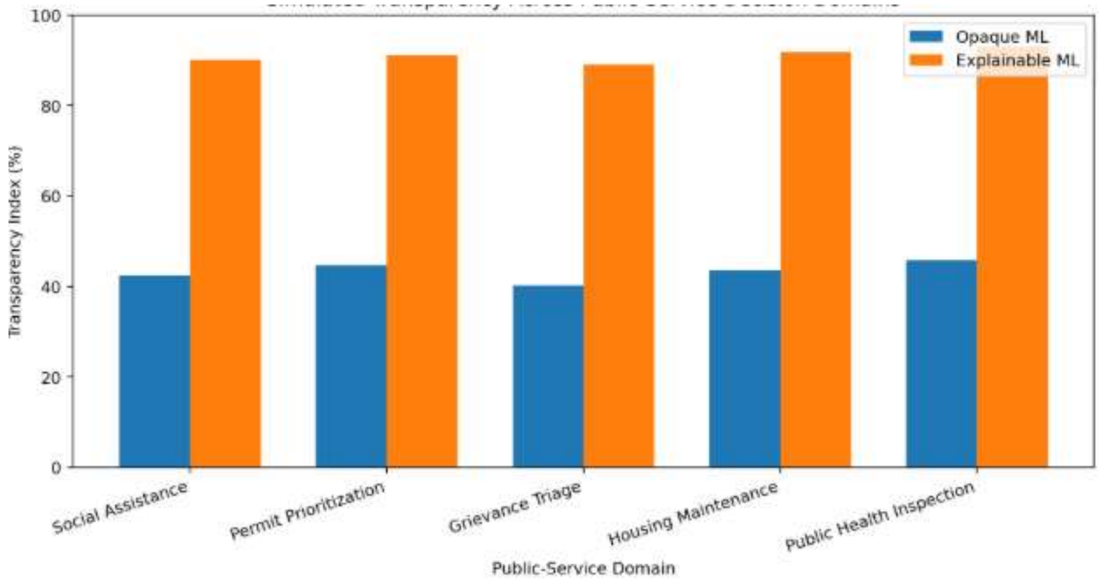
The largest modeled transparency value occurred in public-health inspection at 92.9%, while grievance triage recorded the lowest explainable score at 89.0%. Even in the latter domain, transparency remained more than twice the corresponding opaque value.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

5.3 Transparency–Accuracy Trade-Off

Figure 1: Simulated Transparency across Public-Service Decision Domains



6. Discussion

6.1 Explainability as an Administrative Capability

The findings suggest that explainability should not be treated only as a property of a machine-learning algorithm. In public administration, it is an institutional capability connecting prediction, reason-giving, human review, citizen communication, and audit.

A feature-attribution chart may be technically useful to a data scientist but may provide little value to a citizen attempting to understand an administrative recommendation. Public-service transparency therefore requires explanation at several levels.

The proposed framework converts machine-level interpretation into structured administrative reasons. For example, rather than presenting numerical feature weights alone, an officer-facing interface could communicate that a case received increased priority because of urgency, documented service disruption, and specified vulnerability criteria.

Citizen-facing communication should be simpler and should avoid presenting correlations as causal facts.

6.2 Balancing Accuracy, Transparency, and Human Judgment

The simulation deliberately produced slightly greater accuracy for the opaque configuration. This reflects an important methodological principle: transparency should not be justified through unrealistic claims that interpretable systems always outperform complex systems.

The critical question is whether small differences in predictive performance justify major reductions in explainability where decisions affect citizens.

Rudin argues that interpretable models deserve greater consideration in high-stakes environments, while Doshi-Velez and Kim emphasize that interpretability should itself be evaluated rigorously.

A responsible public-service system should therefore consider multiple metrics simultaneously:

- Predictive accuracy
- Explanation fidelity
- Explanation stability
- Fairness
- Confidence calibration
- Reviewability
- Administrative efficiency
- Citizen comprehension
- Appeal and correction mechanisms

A model that achieves slightly greater predictive accuracy but cannot be meaningfully reviewed may not provide the strongest overall administrative solution.

6.3 Transparency, Contestability, and Public Trust

Transparency becomes particularly meaningful when an explanation enables action. A citizen should be able to understand the principal basis of a decision, identify incorrect information, request human reconsideration where appropriate, and distinguish between a machine recommendation and a final authorized administrative decision.

OECD guidance specifically emphasizes accountability, transparency, explainability, and clearly defined responsibility in public-service systems. NIST similarly treats explainability and interpretability as characteristics that must be balanced with other requirements according to the context in which a system is used.

Transparency should therefore support contestability, not merely disclosure. Publishing a technical model description does not necessarily provide meaningful transparency if affected individuals cannot understand how the system influenced their case.

Limitations

The principal limitation of the study is the use of synthetic data. The numerical results are designed to test the proposed framework and should not be interpreted as performance statistics obtained from actual public agencies.

Second, the Transparency Index is an experimental composite measure. Future research should validate its components through empirical studies involving citizens, public officials, auditors, and domain experts.

Third, explanation techniques can themselves be unstable or misleading. Adebayo and colleagues demonstrate that explanation methods require validation rather than visual acceptance alone.

Future research should therefore compare intrinsically interpretable models, SHAP, LIME, counterfactual explanations, and rule-based explanation systems using real public-sector workflows and human-centered evaluation.

7. Conclusion

This study developed an Explainable Public-Service Machine Learning Framework for transparent administrative decision support. The framework integrates interpretable prediction, feature-level explanation, standardized reason codes, confidence communication, human escalation, decision logging, and review mechanisms. A reproducible simulation involving 2,000 synthetic public-service decisions showed that the explainable configuration achieved an overall accuracy of 78.3% compared with 79.5% for an opaque machine-learning configuration. The modest predictive difference was accompanied by a considerably larger transparency improvement, with the Transparency Index increasing from 43.3% to 90.9%. Simulated administrative review time also declined from 12.0 to 7.4 minutes.

These findings do not constitute empirical evidence from a government agency, but they demonstrate why public-service machine learning should be assessed through a wider set of institutional criteria than predictive accuracy alone. Transparent decision systems must provide meaningful reasons, indicate uncertainty, preserve human responsibility, support correction and review, and maintain auditable records. Explainability is therefore not merely a technical enhancement to machine learning; in public administration, it is a foundational requirement for building decision-support systems that can be scrutinized, challenged, corrected, and responsibly governed.

References

1. M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you?: Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.
2. S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
3. F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," arXiv:1702.08608, 2017.
4. J. Adebayo *et al.*, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems*, vol. 31, 2018.
5. T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, vol. 267, pp. 1–38, 2019.
6. C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019.

American Journal of Advanced Multidisciplinary Innovation and Research (AJAMIR)

E-ISSN: XXXX-XXXX ● Website: www.ajamir.com ● Email: editor@ajamir.com

7. B. Arrieta *et al.*, “Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI,” *Information Fusion*, vol. 58, pp. 82–115, 2020.
8. National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, 2023.
9. National Institute of Standards and Technology, “AI Risks and Trustworthiness,” NIST AI Resource Center.
10. National Institute of Standards and Technology, *AI RMF Playbook*.
11. Organisation for Economic Co-operation and Development, *G7 Toolkit for Artificial Intelligence in the Public Sector*, OECD Publishing, 2024.
12. Organisation for Economic Co-operation and Development, *Governing with Artificial Intelligence*, OECD Publishing, 2025.
13. Organisation for Economic Co-operation and Development, “AI in public service design and delivery,” in *Governing with Artificial Intelligence*, 2025.
14. Organisation for Economic Co-operation and Development, “How artificial intelligence is accelerating the digital government journey,” 2025.
15. Organisation for Economic Co-operation and Development, *The State of Artificial Intelligence in Public Audit*, 2026.
16. Organisation for Economic Co-operation and Development, “Trustworthy artificial intelligence in the public sector,” *OECD Survey on Drivers of Trust in Public Institutions 2026 Results*, 2026.
17. European Commission, “AI Act: Regulatory framework for artificial intelligence,” current regulatory guidance, 2026.
18. European Commission, “Safer and more transparent AI,” Aug. 2, 2026.
19. C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 3rd ed., 2025.