

Securing AI/ML Pipelines in the State Government Sector: Threat Models, Controls, and Governance

Swapan Arora

Independent Researcher, USA

Received: 31.01.2026

Revised: 02.02.2026

Abstract

Government authorities and public sector organizations increasingly deploy artificial intelligence and machine learning systems in critical services including benefits eligibility verification, fraud detection, visitor flow optimization, and health monitoring. However, existing cybersecurity frameworks inadequately address vulnerabilities inherent in machine learning pipelines. Traditional security controls protect infrastructure and data storage against conventional attacks but fail to counter AI-specific threats such as data poisoning, model extraction, and adversarial manipulation, which require specialized protective mechanisms. A significant gap exists between the rapid adoption of AI technologies and the maturity of security governance frameworks in public sector environments. This article presents a comprehensive threat model addressing vulnerabilities across the entire machine learning lifecycle. Technical controls including data provenance verification, cryptographic model signing, secure training environments, and continuous inference monitoring form the foundation of a secure MLOps architecture. Alignment with the NIST AI Risk Management Framework provides structured approaches for risk identification, assessment, and mitigation. Clear governance accountability structures are established through system owners, independent risk reviewers, and oversight committees. Fairness considerations integrate alongside security controls to address public sector equity obligations. Audit trail requirements satisfy transparency mandates and support retrospective review of algorithmic decisions. This framework enables innovation in artificial intelligence while maintaining compliance requirements essential for State government applications.

Keywords: Artificial Intelligence Security, Machine Learning Pipelines, State Government Systems, Threat Modeling, Risk Management Framework, Secure MLOps

1. Introduction

State governments have embraced artificial intelligence deployments to modernize public services and enhance operational effectiveness. Applications span benefits eligibility screening, fraud detection in public assistance programs, predictive analytics for traffic management systems, and surveillance algorithms for disease pattern monitoring in health departments. These deployments represent a fundamental transformation in how government agencies serve citizens. However, this algorithmic transformation raises significant ethical and security considerations that existing frameworks fail to address adequately [1].

Algorithmic systems operating within government contexts create distinct ethical challenges. Research identifies six core concerns surrounding algorithmic decision-making: inconclusive evidence, inscrutable evidence, misguided evidence, unfair outcomes, transformative effects, and traceability [1]. Public sector AI systems directly affect citizen welfare through consequential decisions. Benefits determinations grant

10.48047/jocaaa.2026.35.01.96

or deny essential services. Risk assessment tools influence law enforcement resource allocation. Health surveillance systems guide public health interventions. Each application carries significant implications for individual rights and social equity as documented in algorithmic ethics literature [1].

Machine learning models exhibit reduced transparency compared to traditional rule-based systems, creating accountability difficulties. Affected residents often lack the means to challenge algorithmic determinations. Traditional administrative procedures involve human decision-makers who can provide explanations for their judgments. Machine learning systems operate differently. Complex feature interactions resist straightforward explanation. Model behavior may shift as systems encounter new data distributions. These characteristics create tensions with established principles of administrative law and due process. Algorithmic opacity fundamentally undermines traditional accountability mechanisms [1].

Security considerations add another layer of complexity to government AI deployments. Comprehensive frameworks for understanding security and privacy threats in machine learning reveal an attack surface spanning the entire system lifecycle [2]. Training data can be manipulated to introduce systematic biases. Model parameters can be extracted through carefully crafted queries. Adversarial inputs can trigger misclassification at inference time. Privacy attacks can reveal sensitive information about training data subjects. The breadth of vulnerabilities facing deployed systems necessitates specialized defensive approaches [2].

Traditional cybersecurity controls prove insufficient for AI-specific threats. Firewalls and access controls protect infrastructure but not model integrity. Encryption protects data in transit and at rest but does not prevent poisoning attacks. Intrusion detection systems monitor network traffic but miss adversarial inputs designed to manipulate model behavior. Machine learning systems require specialized defensive approaches beyond conventional cybersecurity measures [2]. Government agencies require security frameworks specifically designed for machine learning systems.

The disparity between AI adoption rates and security maturity creates substantial risk exposure. Agencies deploy sophisticated models without corresponding security governance. Attack techniques documented in research literature become increasingly accessible [2]. Nation-state adversaries and criminal organizations recognize the value of manipulating government AI systems. The consequences of successful attacks extend beyond data breaches to corrupted decision-making affecting millions of citizens. This article addresses the security governance gap through a comprehensive approach integrating technical controls with governance mechanisms aligned to established risk management standards.

2. Related Work and Framework

Prior literature on machine learning security primarily addresses commercial and enterprise contexts. Government-specific security frameworks remain underdeveloped despite the distinct requirements of public sector environments. Existing information security standards such as NIST SP 800-53 provide fundamental controls for information systems. However, the unique vulnerabilities inherent in AI technology demand specialized defensive approaches. The NIST AI Risk Management Framework provides structured guidance for AI risk management, yet practical implementation guidance for State government ML pipelines remains scarce.

This article bridges the gap between theoretical AI security principles and actionable government implementation. A comprehensive threat model identifies vulnerabilities spanning training, deployment,

10.48047/jocaaa.2026.35.01.96

and inference phases. Technical controls map directly to NIST AI RMF functions including Govern, Map, Measure, and Manage. The proposed secure MLOps architecture integrates data provenance verification, cryptographic model signing, and continuous monitoring. Governance structures establish clear accountability through designated owners and independent reviewers. Fairness considerations integrate alongside security controls to address public sector equity obligations. The framework positions security as embedded throughout the machine learning lifecycle rather than appended after deployment. Alignment with established risk management standards enables regulatory compliance while maintaining operational effectiveness. This contribution advances practical security governance for State government AI applications.

3. AI/ML Threat Landscape in Government Operations

3.1 Training Phase Vulnerabilities

Machine learning models inherit the characteristics of their training data. This fundamental property creates opportunities for adversarial manipulation during the training phase. Data poisoning attacks exploit this dependency by injecting malicious samples into training datasets. Corrupted data influences model parameters in ways that serve attacker objectives. Government systems that aggregate data from multiple sources face elevated exposure to such attacks. Systematic analysis documents the evolution and current state of data poisoning threats [3].

Comprehensive taxonomies of data poisoning threats and defenses categorize attacks based on attacker objectives [3]. Availability attacks aim to degrade overall model performance, rendering systems unreliable or unusable. Integrity attacks pursue more subtle objectives, seeking to manipulate model behavior for specific inputs while preserving normal performance otherwise. Integrity attacks prove particularly dangerous because standard validation procedures may fail to detect them. Research demonstrates that integrity attacks present greater challenges for defensive systems [3].

Backdoor attacks represent a sophisticated form of integrity attack. Attackers embed hidden triggers in training data. Models trained on poisoned data behave normally for standard inputs. However, inputs containing trigger patterns produce attacker-specified outputs. Government fraud detection systems could be compromised to ignore specific patterns of fraudulent activity. Benefits eligibility models could be manipulated to approve or deny specific applicant profiles. Backdoors remain dormant until activated, evading detection during testing and initial deployment. Analysis of backdoor attack methodologies reveals their resistance to standard defenses [3].

The threat landscape continues evolving as attack techniques mature. Clean-label attacks poison training data without changing ground truth labels, proving difficult to detect through manual review. Federated learning systems face unique poisoning vulnerabilities where malicious participants contribute corrupted model updates. Data validation techniques provide partial mitigation but cannot eliminate poisoning risks entirely. Defense requires layered approaches combining data sanitization, anomaly detection, and robust training algorithms. Ongoing research indicates that comprehensive defenses against data poisoning remain an open problem [3].

3.2 Inference Phase Attacks

Deployed models face distinct threat categories during operational use. Model extraction attacks seek to replicate proprietary algorithms through systematic querying. Research demonstrates practical extraction attacks against production machine learning systems [4]. Attack methodology involves submitting carefully chosen inputs and observing model outputs. Sufficient query-response pairs enable training

10.48047/jocaaa.2026.35.01.96

substitute models that approximate target model behavior. Studies show that extraction attacks succeed against multiple model types including decision trees, logistic regression, and neural networks [4]. Government AI systems present attractive targets for extraction attacks. Fraud detection algorithms represent significant investments in feature engineering and model development. Adversaries who extract these models gain insight into detection criteria. This knowledge enables designing fraud schemes that evade detection. Public safety risk assessment tools similarly become less effective once adversaries understand their decision boundaries. Research demonstrates that prediction APIs provide sufficient information for successful model extraction even without access to confidence scores [4]. Model inversion attacks pursue different objectives. Rather than replicating model functionality, inversion attacks reconstruct training data characteristics. Sensitive attributes of training subjects can be inferred from model outputs. Government health surveillance models trained on patient data could leak information about specific individuals. Benefits eligibility models might reveal patterns in approved applicant profiles. These privacy violations carry legal implications under data protection regulations. Attack surface analysis encompasses both extraction and privacy threats facing deployed systems [4]. Adversarial input attacks manipulate model behavior at inference time. Small perturbations to input data can trigger misclassification. Image recognition systems misidentify objects when pixels are modified imperceptibly. Document processing systems can be deceived by carefully crafted formatting. Perturbations often appear random to human observers while reliably triggering incorrect model outputs. Membership inference attacks determine whether specific records appeared in training data. Confirming that an individual appeared in a health surveillance training dataset reveals information about their medical history. Comprehensive threat modeling provides the foundation for understanding inference-phase vulnerabilities [4].

Threat Category	Attack Phase	Description	Government Impact
Data Poisoning	Training	Injection of malicious samples into training datasets	Systematic biases in eligibility determinations
Backdoor Attacks		Embedding hidden triggers that activate malicious behavior	Fraud detection bypass for specific patterns
Clean-Label Attacks		Poisoning without changing ground truth labels	Undetectable manipulation of model behavior
Model Extraction	Inference	Replication of algorithms through systematic querying	Exposure of fraud detection criteria
Model Inversion		Reconstruction of training data characteristics	Privacy violations revealing citizen information
Adversarial Inputs		Perturbations triggering misclassification	Manipulation of risk assessment outputs
Membership Inference		Determination of training data inclusion	Disclosure of individual participation in datasets

Table 1. AI/ML Threat Categories in Government Operations [3, 4].

4. Secure MLOps Architecture Design

4.1 Data Provenance and Integrity

Secure machine learning pipelines require comprehensive tracking of data lineage. Every dataset entering the training process must have documented origins and transformation history. This provenance information enables verification of data integrity and supports forensic investigation when anomalies occur. Government agencies aggregating data from multiple sources face particular challenges in maintaining complete provenance records. Research articulates the critical role of provenance in establishing trustworthy machine learning systems [5].

The connection between machine learning provenance and FAIR data principles provides actionable implementation guidance [5]. Findability requires unique identifiers for datasets and models at each pipeline stage. Accessibility demands clear protocols for retrieving provenance information. Interoperability necessitates standardized formats for representing lineage relationships. Reusability depends on sufficient documentation for independent verification. These principles guide the design of provenance systems for government AI applications [5].

Data validation forms the primary defense against poisoning attacks. Statistical profiling establishes baseline characteristics for expected data distributions. Incoming data undergoes comparison against these baselines. Significant deviations trigger alerts for manual review. Schema validation ensures structural consistency. Type checking prevents injection of malformed records. Referential integrity verification confirms relationships between related datasets. Reproducibility requirements necessitate validation mechanisms that ensure data quality throughout the pipeline [5].

Cryptographic techniques provide tamper-evidence for stored data. Hash values computed over dataset contents detect unauthorized modifications. Digital signatures authenticate data sources. Merkle trees enable efficient verification of large datasets. Immutable storage systems prevent retrospective alteration of historical data. Government agencies must balance security requirements against storage costs and performance constraints. Access controls restrict data modification privileges to authorized personnel. Role-based access models define permissions based on job responsibilities. Data integrity mechanisms support both security and reproducibility objectives [5].

4.2 Model Development and Deployment Controls

Secure training environments isolate model development from production systems. Containerized workloads provide consistent execution environments. Network segmentation limits communication pathways. Resource quotas prevent denial-of-service through excessive computation. These controls contain the impact of potential compromises during the development phase. Security requirements for enterprise MLOps pipelines emphasize isolation as a foundational control [6].

Model versioning maintains complete histories of all development iterations [6]. Each version captures model architecture, hyperparameters, training data references, and evaluation results. Version control systems designed for machine learning handle large binary artifacts efficiently. Branching strategies support parallel experimentation while maintaining clear lineage to production models. Secure MLOps frameworks position version control as essential for both security and operational integrity [6].

Cryptographic signing establishes authenticity for model artifacts. Private keys held by authorized personnel generate signatures during model approval. Deployment systems verify signatures before loading models into production. Hardware security modules protect signing keys from extraction. Certificate authorities manage key lifecycle and revocation. This infrastructure prevents deployment of unauthorized or tampered models. Enterprise AI security architectures detail cryptographic controls for model integrity verification [6].

10.48047/jocaaa.2026.35.01.96

Deployment pipelines enforce security validation gates. Automated security scanning checks for known vulnerabilities in model dependencies. Bias testing evaluates model fairness across protected demographic categories. Robustness testing assesses model behavior under adversarial conditions. Models failing validation cannot proceed to production deployment. Manual override requires documented approval from security leadership. Continuous monitoring tracks model behavior in production environments. Inference latency metrics detect potential denial-of-service conditions. Output distribution monitoring identifies drift from expected patterns. MLOps security frameworks integrate continuous monitoring with deployment controls [6].

Component	Security Function	Implementation Mechanism
Data Provenance	Lineage tracking and integrity verification	Cryptographic hashing and digital signatures
Statistical Profiling	Baseline establishment for anomaly detection	Distribution comparison and deviation alerts
Schema Validation	Structural consistency enforcement	Type checking and referential integrity
Training Isolation	Containment of development compromises	Containerization and network segmentation
Model Versioning	Complete iteration history maintenance	Version control with hyperparameter tracking
Cryptographic Signing	Model artifact authenticity verification	Hardware security modules and certificate authorities
Deployment Gates	Security validation before production release	Automated scanning and bias testing
Continuous Monitoring	Production behavior analysis	Inference pattern tracking and drift detection

Table 2. Secure MLOps Architecture Components [5, 6].

5. Control Framework Alignment

5.1 NIST AI Risk Management Framework Integration

Effective AI security requires holistic risk management rather than piecemeal approaches. The NIST AI Risk Management Framework provides structured guidance for comprehensive risk management. Analysis of how existing risk management standards apply to high-consequence AI systems informs government implementation strategies [7]. Government agencies can leverage established frameworks while addressing AI-specific considerations. Research provides actionable guidance for organizations implementing AI risk management programs [7].

The Governance function establishes organizational context for AI risk management. Senior leadership commitment ensures proper funding and enablement of security programs. Risk tolerance statements define acceptable boundaries for AI system behavior. Policies codify security requirements and compliance obligations. Organizational culture shapes how personnel perceive and respond to AI-related threats. Government agencies must align AI governance with existing information security management systems. Governance structures determine the effectiveness of technical controls [7].

10.48047/jocaaa.2026.35.01.96

The Map function characterizes AI systems and their operational contexts. Threat modeling identifies potential adversaries, attack vectors, and impact scenarios. Data flow diagrams document information movement through AI pipelines. Dependency analysis reveals supply chain risks in model development. Stakeholder analysis identifies affected populations and their interests. Comprehensive mapping enables targeted control selection based on specific risk profiles. Risk management guidance highlights mapping as a prerequisite for effective control implementation [7].

The Measure function assesses AI system trustworthiness characteristics. Security testing evaluates resilience against known attack techniques. Performance metrics track model accuracy and reliability over time. Fairness assessments examine outcome distributions across demographic groups. Transparency evaluations assess explainability and interpretability. Regular measurement enables trend analysis and early problem detection. The Manage function implements risk response activities. Control implementation addresses identified vulnerabilities. Incident response procedures define actions when security events occur. Measurement and management functions interact to maintain security posture throughout system lifecycles [7].

5.2 Fairness and Bias Considerations

Government AI systems must satisfy fairness requirements beyond technical security controls. Algorithmic decisions affecting citizens carry equity obligations. Bias in training datasets or model architectures can produce discriminatory effects. Security frameworks must incorporate fairness objectives alongside confidentiality, integrity, and availability. Research surveys the landscape of fairness in machine learning and documents the complexity of achieving equitable outcomes [8].

Multiple fairness definitions exist, and mathematical constraints prevent simultaneous satisfaction of all criteria [8]. Demographic parity requires equal positive outcome rates across groups. Equalized odds demand equal true positive and false positive rates. Individual fairness requires similar treatment for similar individuals. Government agencies must select appropriate fairness criteria based on application context and legal requirements. Analysis demonstrates inherent tradeoffs between different fairness definitions [8].

Fairness testing integrates into secure MLOps pipelines alongside security validation. Training data undergoes analysis for demographic representation and historical bias. Model outputs receive systematic evaluation across protected categories. Disparate impact analysis compares outcome rates between groups. Intersectional analysis examines outcomes for individuals belonging to multiple protected categories. Detected disparities trigger review and potential remediation before deployment. Fairness research informs testing methodologies for government systems [8].

Ongoing fairness monitoring addresses drift and changing populations. Model behavior may shift as input distributions evolve. Demographic composition of affected populations changes over time. Periodic fairness audits verify continued compliance with equity requirements. Feedback mechanisms enable affected individuals to report perceived unfairness. Integration with complaint management systems supports systematic response to fairness concerns. Fairness requires ongoing attention rather than one-time validation [8].

RMF Function	Primary Objective	Associated Controls	Government Application
Govern	Organizational context establishment	Risk tolerance statements and policy codification	Leadership commitment and resource allocation
Map	System and context characterization	Threat modeling and data flow documentation	Adversary identification and impact assessment
Measure	Trustworthiness assessment	Security testing and fairness evaluation	Bias detection and robustness validation
Manage	Risk response implementation	Incident response and continuous improvement	Remediation workflows and compliance verification

Table 3. NIST AI RMF Function Mapping [7, 8].

6. Governance Models and Audit Readiness

6.1 Accountability Structures

Effective governance requires explicit definition of accountability relationships. AI system ownership must be formally designated rather than assumed. Accountability gaps create vulnerabilities where security requirements fall between organizational boundaries. Government agencies benefit from formal governance structures that define roles, authorities, and reporting relationships for AI security. Research examines mechanisms for establishing trustworthy AI development practices and emphasizes institutional accountability [9].

Institutional factors shape organizational dynamics related to AI safety and security [9]. Third-party audits provide independent security assurance. Bug bounties and red teaming enable external security research. Incident reporting mechanisms facilitate learning from security events. Government agencies can adapt these mechanisms for public sector contexts. Trustworthy AI frameworks provide taxonomies of accountability mechanisms [9].

AI system owners bear primary responsibility for security outcomes. Owners authorize system deployment and accept associated risks. Resource allocation decisions reflect owner priorities. Performance accountability motivates investment in security controls. Clear ownership prevents diffusion of responsibility that undermines security programs. Government agencies should document ownership assignments in system inventories and authorization records. Accountability requires both formal designation and meaningful authority [9].

Independent model risk review provides objective assessment separate from development teams. Reviewers evaluate security controls, fairness characteristics, and compliance status. Independence requires organizational separation and reporting relationships outside development management. Review findings inform deployment decisions and remediation priorities. Escalation procedures address disagreements between reviewers and system owners. Oversight committees integrate perspectives across security, privacy, ethics, and operations. Verification mechanisms support independent review processes [9].

6.2 Audit Trail and Explainability Requirements

Government AI systems operate under legal and regulatory requirements for transparency and accountability. Administrative procedures assume decision-makers can explain their reasoning. Citizens affected by adverse decisions possess rights to review and appeal. These requirements create obligations

10.48047/jocaaa.2026.35.01.96

for AI system documentation and explainability that exceed typical commercial applications. Research examines requirements for explainable artificial intelligence in organizational contexts [10].

Explainability serves multiple purposes beyond regulatory compliance [10]. System developers use explanations to debug model behavior. Operators rely on explanations to verify appropriate system function. Affected individuals need explanations to understand decisions and identify errors. Auditors require explanations to verify compliance with policies and regulations. Explainability frameworks address the needs of multiple stakeholder groups [10].

Comprehensive audit trails document AI system behavior for retrospective review. Input logging captures data submitted to models for inference. Output logging records model predictions and confidence scores. Decision logging tracks how model outputs influenced final determinations. Version logging identifies which model version produced each output. Timestamp logging enables reconstruction of system state at any historical point. Organizational requirements for explainability necessitate comprehensive logging infrastructure [10].

Reproducible training environments support independent verification of model characteristics. Infrastructure-as-code practices capture environment configurations. Dependency management ensures consistent software versions. Random seed documentation enables deterministic reproduction of training runs. Data snapshots preserve training datasets at specific points in time. Documentation standards ensure audit artifacts remain interpretable over time. Retention policies ensure documentation availability throughout system lifecycles and beyond. Explainability requirements extend throughout the entire system lifecycle [10].

Governance Element	Responsibility	Key Activities
AI System Owners	Primary security accountability	Deployment authorization and risk acceptance
Independent Reviewers	Objective assessment	Security control and fairness evaluation
Oversight Committees	Cross-functional integration	Portfolio risk review and policy recommendations
Input Logging	Decision traceability	Capture of inference request data
Output Logging	Prediction documentation	Recording of model outputs and confidence scores
Version Logging	Model identification	Tracking of deployed model versions
Reproducible Environments	Independent verification	Infrastructure-as-code and dependency management

Table 4. Governance and Accountability Framework [9, 10].

Conclusion

Securing AI and ML pipelines within the State government sector demands lifecycle-wide security integration rather than reactive post-deployment measures. Adversarial threats targeting government algorithmic systems span multiple attack categories. Data poisoning compromises model training integrity. Model extraction enables adversaries to replicate proprietary detection algorithms. Adversarial inputs manipulate inference outputs. Supply chain vulnerabilities introduce risks through compromised

10.48047/jocaaa.2026.35.01.96

dependencies. Conventional cybersecurity measures designed for traditional information systems cannot adequately counter machine learning-specific attack vectors.

Embedded technical safeguards must address vulnerabilities at every pipeline stage. Data provenance mechanisms verify dataset origins and transformation histories. Cryptographic signing authenticates model artifacts before production deployment. Isolated training environments contain potential compromises during development phases. Continuous monitoring detects anomalous inference patterns and model drift.

Governance structures enable proper accountability through defined ownership and independent performance evaluation. Oversight committees monitor performance across security, privacy, and ethics domains. Alignment with risk management best practices streamlines compliance efforts. Fairness testing ensures equity of outcomes across demographic categories affected by algorithmic decisions. Comprehensive audit trails satisfy transparency obligations inherent in public sector operations. Explainability requirements enable citizens to understand and contest adverse determinations.

Public trust in government AI depends on demonstrable security and fairness measures. Future advancement should prioritize automated compliance validation mechanisms. Security frameworks must evolve alongside expanding AI capabilities to counter emerging threats while enabling beneficial public service applications.

References

- [1] Brent Daniel Mittelstadt et al., "The ethics of algorithms: Mapping the debate," Big Data & Society, 2016. [Online]. Available: https://journals.sagepub.com/doi/pdf/10.1177/2053951716679679?source=post_page
- [2] Nicolas Papernot et al., "SoK: Towards the Science of Security and Privacy in Machine Learning," arXiv, 2016. [Online]. Available: <https://arxiv.org/pdf/1611.03814>
- [3] Antonio Emanuele Cina et al., "Machine Learning Security against Data Poisoning: Are We There Yet?," arXiv, 2024. [Online]. Available: <https://arxiv.org/pdf/2204.05986>
- [4] Florian Tramèr et al., "Stealing Machine Learning Models via Prediction APIs," 25th USENIX Security Symposium, 2016. [Online]. Available: https://www.usenix.org/system/files/conference/usenixsecurity16/sec16_paper_tramer.pdf
- [5] Sheeba Samuel et al., "Machine Learning Pipelines: Provenance, Reproducibility and FAIR Data Principles," arXiv, 2020. [Online]. Available: <https://arxiv.org/pdf/2006.12117>
- [6] Dave Noel et al., "Advancing Enterprise AI: From Generative Models to Privacy-Preserving Systems and Secure MLOps Pipelines," [Online]. Available: <https://www.researchgate.net/profile/Blessing-Idowu/publication/394979387>
- [7] Anthony M. Barrett et al., "Actionable Guidance for High-Consequence AI Risk Management: Towards Standards Addressing AI Catastrophic Risks," arXiv, 2023. [Online]. Available: <https://arxiv.org/pdf/2206.08966>
- [8] ALEXANDRA CHOULDECHOVA AND AARON ROTH, "A Snapshot of the Frontiers of Fairness in Machine Learning," Communications of the ACM, 2020. [Online]. Available: <https://dl.acm.org/doi/pdf/10.1145/3376898>
- [9] Miles Brundage et al., "Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims," arXiv, 2020. [Online]. Available: <https://arxiv.org/pdf/2004.07213>

10.48047/jocaaa.2026.35.01.96

[10] Julie Gerlings et al., "Reviewing the Need for Explainable Artificial Intelligence (xAI)," Proceedings of the 54th Hawaii International Conference on System Sciences, 2021. [Online]. Available: <https://arxiv.org/pdf/2012.01007>