Journal of Business and Management Studies
ISSN: 2709-0876
DOI: 10.32996/jbms
Journal Homepage: www.al-kindipublisher.com/index.php/jbms

JBMS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Safeguarding Federal Payment Infrastructure: Trustworthy AI for Improper-Payment Prevention, Synthetic Identity Detection, and Public-Benefit Disbursement Integrity in the United States

Yusuf Oli Rahat
Brac University
Email: yusuf.rahat@gmail.com
ORCID: https://orcid.org/ 0009-0001-8834-7759

Md Kamrul Islam
University of New Haven, Business Analytics
Email: misla22@unh.newhaven.edu
ORCID: https://orcid.org/0009-0001-8906-630X

Shah Farhan Rabbani
North South University
Email: shah.rabbani@northsouth.edu
ORCID: https://orcid.org/0009-0001-8434-223X
Corresponding Author: Yusuf Oli Rahat, **E-mail**: yusuf.rahat@gmail.com

| ABSTRACT
Federal payment integrity has become a high-stakes governance problem in the United States because public-benefit disbursement systems must now operate at digital scale while defending against increasingly adaptive fraud, identity manipulation, and data-quality failures. Recent federal estimates show that improper payments remain material across major programmes, while agencies and oversight bodies have increasingly explored data matching, risk analytics, and automated screening to strengthen prevention and recovery. Yet the policy challenge is not simply whether artificial intelligence can detect more anomalies. In public disbursement settings, analytical performance must be reconciled with due process, transparency, data governance, and the need to avoid delaying legitimate payments to eligible recipients. This paper develops a literature-based conceptual framework for trustworthy AI in federal payment infrastructure, with particular attention to improper-payment prevention, synthetic identity detection, and public-benefit disbursement integrity. Using an integrative review of peer-reviewed scholarship and authoritative U.S. government sources published, the paper synthesises four literatures that are often treated separately: payment integrity, digital identity assurance, AI-enabled fraud analytics, and public-sector AI governance. The paper proposes a multilayer framework that combines identity-proofing controls, multimodal risk scoring, graph-based network analysis, human-centred adjudication, and continuous governance monitoring. Its central argument is that effective federal payment integrity requires not only predictive accuracy but institutional trustworthiness: auditable models, calibrated thresholds, fairness safeguards, contestability, and shared cross-programme infrastructure. The paper contributes a structured analytical model, implementation logic, policy implications, and an evaluation template for future empirical testing.

| KEYWORDS

Trustworthy AI; federal payments; improper payments; synthetic identity fraud; payment integrity; public-benefit disbursement; digital identity; fraud analytics; public-sector AI; FinTech

| ARTICLE INFORMATION

ACCEPTED: 01 August 2024 **PUBLISHED:** 29 August 2024 **DOI:** 10.32996/jbms.2024.6.5.25

1. Introduction

Federal agencies disburse funds across some of the most consequential payment environments in the United States, including tax refunds, retirement benefits, income support, disaster assistance, healthcare reimbursements, and a wide range of programmatic transfers. Because these systems sit at the intersection of public finance, administrative law, and digital infrastructure, payment errors and fraud are not merely operational inefficiencies; they are failures of fiscal stewardship and public trust. The Government Accountability Office reported that improper payments remained substantial in fiscal year 2023, and GAO separately estimated direct annual financial losses from federal fraud at between $233 billion and $521 billion based on fiscal years 2018-2022 (Government Accountability Office 2024a; Government Accountability Office 2024b). Even allowing for definitional and measurement differences between fraud and improper payments, the scale of the problem places federal payment integrity squarely within the national-interest domain of fiscal resilience, service legitimacy, and administrative capacity.

The policy environment has changed in two ways that make the issue more urgent. First, digitisation has increased the speed and scale at which agencies originate, certify, and reconcile payments. Second, the fraud environment has become more adaptive. Synthetic identity schemes, account takeovers, altered checks, credential abuse, and cross-programme duplications exploit both fragmented data and institutional lag. In benefit and refund contexts, synthetic or partially synthetic identities may interact with eligibility systems, payment instructions, and bank-account information in ways that are difficult to detect through rules alone.

Federal institutions have responded with more ambitious analytical tools. By the first half of 2024, federal oversight and management guidance had already placed stronger emphasis on data quality, shared controls, and risk-based governance for AI-enabled screening and decision support (Government Accountability Office 2021; Office of Management and Budget 2024). Yet the emergence of stronger analytics does not dissolve governance concerns. In the public sector, the relevant question is not whether AI can be used, but how it can be used in ways that remain lawful, contestable, accurate, and proportionate.

This paper argues that the federal payment-integrity challenge is best understood as a design problem in trustworthy AI. Existing literatures provide important but incomplete pieces of the puzzle. Research on financial fraud detection shows the value of anomaly detection, graph analytics, and multimodal pattern recognition (West and Bhattacharya 2016; Ryman-Tubb, Krause and Garn 2018). Research on public-sector AI and accountability highlights transparency, discretion, and governance concerns in administrative settings. U.S. government practice offers identity-proofing and payment-integrity tools such as Do Not Pay, account verification, and cross-agency screening. What is still missing is an integrated framework designed specifically for federal payment infrastructure that combines fraud analytics with public-law safeguards and operational payment logic.

The contribution of this paper is fourfold. First, it synthesises the fragmented literatures on improper payments, synthetic identity risk, fraud analytics, and public-sector AI governance into a single analytical frame. Second, it identifies the principal research gap: the absence of a payment-system-specific model for trustworthy AI deployment in federal benefit and transfer environments. Third, it develops a multilayer framework for pre-disbursement and post-disbursement integrity management, combining identity assurance, multimodal risk scoring, graph-based threat detection, human review, and continuous monitoring. Fourth, it outlines an implementation and evaluation template suitable for future pilot testing by federal agencies or shared-services entities. The remainder of the paper reviews the literature, defines the research gap, explains the review-based methodology, presents the proposed framework, and discusses theoretical, policy, and managerial implications.

2. Background and literature review

The literature relevant to federal payment integrity spans at least four domains: payment-integrity administration, digital identity assurance, AI-based fraud detection, and public-sector AI governance. These literatures have developed largely in parallel. Payment-integrity scholarship and official reporting focus on programme errors, documentation failures, and control environments. Digital identity work concentrates on proofing, authentication, and credential assurance. Financial-fraud research emphasises classification, anomaly detection, graph learning, and increasingly multimodal analytics. Public-administration and AI-governance scholarship is more concerned with legitimacy, transparency, discretion, bias, and redress. A convincing framework for

federal disbursement integrity must bring these strands together rather than treating them as separate technical or policy conversations.

2.1 Federal payment integrity and public-benefit disbursement

A useful starting point is the distinction between improper payments and fraud. Improper payments are payments that should not have been made, were made in an incorrect amount, or cannot be verified because of insufficient documentation. Fraud, by contrast, involves obtaining value through willful misrepresentation. Not all improper payments are fraudulent, but all fraudulent payments are improper. GAO's 2024 reporting shows that federal payment errors are driven by a combination of overpayments, underpayments, unknown payments, and technically improper payments, with missing or insufficient documentation remaining an important cause across programmes (Government Accountability Office 2024b). This matters because a federal payment-integrity system cannot be designed exclusively as a fraud-catching mechanism; it must also address eligibility verification, documentation quality, and process discipline.

The scale and concentration of improper payments further clarify where analytical systems matter most. GAO's review shows that high improper-payment rates remain concentrated in a relatively small number of large programme areas, including healthcare and income-support programmes, even though these environments differ substantially in rules, evidence standards, and payment logic (Government Accountability Office 2024b). These are not homogeneous environments. Some involve recurring beneficiary payments, others claims or reimbursements, and still others episodic or emergency disbursements. A trustworthy AI approach therefore cannot rely on a single universal risk score. It must be modular enough to respect programme-specific rules, evidence standards, disbursement timing, and harm asymmetries.

Recent federal practice illustrates the growing role of shared payment-integrity infrastructure. Existing tools such as Do Not Pay, account verification, death-data matching, and cross-government analytics show that the federal challenge is increasingly infrastructural: how to embed reliable shared services into programme-specific payment workflows. These developments suggest that payment-integrity modernization should be understood not only as a modelling task, but also as a systems-integration and governance task.

Table 1. Selected literature streams informing the proposed framework

Domain	Representative source	Key contribution	Relevance to this paper
Fraud and improper payments	Government Accountability Office (2024a; 2024b)	Estimates the scale of federal fraud and improper payments; identifies persistent causes such as overpayments and insufficient documentation.	Establishes why payment integrity is a national fiscal and administrative problem.
Identity assurance	National Institute of Standards and Technology (2020)	Defines digital identity assurance requirements and identifies synthetic identity fraud as a concrete risk to verification systems.	Supports the identity-proofing layer and synthetic-identity focus of the framework.
Fraud analytics	West and Bhattacharya (2016); Ryman-Tubb, Krause and Garn (2018)	Shows the shift from rules to AI/ML and the particular value of graph structures for complex fraud detection.	Supports multimodal and network-based risk detection in federal payment environments.
Public-sector trustworthiness	National Institute of Standards and Technology (2023); Office of Management and Budget (2024); Government Accountability Office (2021)	Specifies trustworthiness, accountability, and governance requirements for AI in public settings.	Supports the claim that accuracy alone is insufficient in rights-affecting systems.

Domain	Representative source	Key contribution	Relevance to this paper
Human use of AI	Public-administration and human oversight literature	Highlights automation bias, transparency, accountability, and contestability challenges.	Justifies human-centred review, audit trails, and redress mechanisms.
Financial infrastructure analogues	FAHIM et al. (2023); Hasan et al. (2023); Rasel et al. (2023)	Shows governance, multimodal analytics, and infrastructure-resilience logic in adjacent financial domains.	Provides transferable design ideas for public payment infrastructure.

Note: The table summarises the main bodies of scholarship and official guidance synthesised in this paper.

2.2 Digital identity assurance and synthetic identity risk

Identity assurance is a central but under-integrated part of payment integrity. NIST's Digital Identity Guidelines provide the federal baseline for identity proofing, authentication, and federation, emphasising that agencies should calibrate digital identity assurance to risk rather than treat identity verification as a one-time administrative formality (National Institute of Standards and Technology 2020). This is especially relevant where public benefits are claimed remotely, account information is updated online, or documentation is submitted through digital channels. High-volume remote interactions create opportunities for identity recycling, synthetic profile creation, credential compromise, and coordinated mule-account behaviour.

Identity assurance is a central but under-integrated part of payment integrity. NIST's Digital Identity Guidelines provide the federal baseline for identity proofing, authentication, and federation, emphasising that agencies should calibrate digital identity assurance to risk rather than treat identity verification as a one-time administrative formality (National Institute of Standards and Technology 2020). This is especially relevant where public benefits are claimed remotely, account information is updated online, or documentation is submitted through digital channels. High-volume remote interactions create opportunities for identity recycling, profile fabrication, credential compromise, and coordinated mule-account behaviour.

The payment-integrity implications are substantial. If identity proofing is weak, downstream transaction analytics will be forced to detect fraud late, after false beneficiaries have already been admitted into the payment flow. If identity proofing is too rigid or opaque, legitimate beneficiaries may be excluded or delayed. Identity assurance therefore works best as a layered control environment rather than a single gate. That logic is consistent with federal digital identity guidance and supports an AI design philosophy based on corroboration across data modalities rather than overreliance on any one signal.

2.3 AI-enabled fraud analytics and multimodal detection

The financial-fraud literature offers several well-established lessons that can be adapted to the federal payment context. First, fraud detection has shifted from purely rule-based systems toward hybrid architectures combining rules, supervised classification, unsupervised anomaly detection, and increasingly graph-based learning. West and Bhattacharya (2016) showed in an early comprehensive review that financial fraud detection involves heterogeneous behaviours and domain-specific constraints, making single-method approaches inadequate. Ryman-Tubb et al. (2018) further argued that industrial fraud settings require methods that balance accuracy, adaptability, latency, and interpretability. These insights remain relevant because public-benefit fraud, like payment-card fraud, occurs in dynamic environments with imbalanced data, evolving tactics, and high costs of both false negatives and false positives.

Research on fraud analytics further emphasises the value of relational and multimodal detection for complex fraud ecosystems. Prior survey work shows that financial-fraud detection increasingly benefits from hybrid architectures that combine business rules, supervised models, anomaly detection, and network-aware analysis rather than relying on a single classifier (West & Bhattacharya, 2016; Ryman-Tubb et al., 2018). This is highly relevant to synthetic identity and organised disbursement abuse because many risk signals only become visible when entities are analysed relationally. A beneficiary record may appear plausible in isolation while still belonging to a suspicious cluster connected through mailing addresses, IP addresses, employer identifiers, bank accounts, phone numbers, or timing patterns.

Adjacent scholarship in FinTech, predictive analytics, and financial-integrity governance also contributes useful transferable ideas. FAHIM et al. (2023) emphasise governance mechanisms in high-stakes digital financial systems, while Rasel et al. (2023) demonstrate the value of multimodal and graph-based data in moving beyond narrow legacy scoring systems. Hasan et al. (2023) likewise show that AI-driven fraud detection and risk analytics are most effective when they are embedded within broader financial and cybersecurity infrastructure rather than deployed as isolated scoring tools. Ibrahim et al. (2022) further show that predictive analytics can improve risk visibility and strategic decision quality under uncertainty, which is useful for thinking about early-warning and prioritisation layers in public payment systems. Ibrahim et al. (2024) argue that financial-integrity frameworks are stronger

when transaction monitoring, entity resolution, and infrastructure protection are integrated rather than treated as separate control functions. Jahan et al. (2024) extend this logic by demonstrating the value of early-warning analytics for detecting irregular patterns in fast-moving financial environments, while Pritty et al. (2024) show that advanced AI can also support transparency, narrative scrutiny, and disclosure-oriented monitoring in complex reporting ecosystems. Although these studies address different financial domains, they collectively support a broader proposition: high-stakes financial integrity systems benefit from multimodal, network-aware, governance-sensitive, and transparently auditable AI designs.

The same body of literature also highlights limitations. Fraud data are severely imbalanced; labels can be noisy or delayed; adversaries adapt to detection strategies; and highly accurate black-box systems may be institutionally unacceptable in regulated or rights-sensitive settings. For federal payments, these limitations are magnified by legal requirements, documentation standards, beneficiary rights, and the political salience of service denial. A model that performs well on retrospective fraud labels may still be unsuitable if it is not auditable, contestable, or operationally aligned with how agencies certify and release funds.

2.4 Trustworthy AI, accountability, and public-sector governance

The governance literature clarifies why trustworthiness must be treated as a design requirement rather than an afterthought. NIST's AI Risk Management Framework defines trustworthiness in terms such as validity, reliability, accountability, transparency, explainability, privacy enhancement, fairness, and security (National Institute of Standards and Technology, 2023). NIST's bias guidance further warns that AI systems can reproduce or amplify harmful outcomes when data, labels, deployment contexts, or institutional assumptions are poorly understood (National Institute of Standards and Technology, 2022). Related work on financial reporting integrity and accountable FinTech governance reinforces the same point: strong model performance is not sufficient unless institutions can explain decisions, monitor misuse, and document controls around high-impact AI deployment (FAHIM et al., 2023; Pritty et al., 2024). In a federal payment setting, such risks are not abstract. False positives can delay subsistence benefits, create administrative burden, or trigger investigations against eligible recipients. False negatives can allow large-scale losses and erode programme legitimacy.

Public-sector scholarship similarly emphasises that AI in administrative systems interacts with bureaucratic discretion, accountability structures, and citizen trust. In public-benefit disbursement, these concerns are especially consequential because even supportive screening tools can shape human action in rights-sensitive environments. Trustworthy AI therefore requires procedural safeguards around the human use of algorithmic advice, not merely technical improvements in model performance.

The U.S. federal policy framework now makes these governance demands more explicit. OMB Memorandum M-24-10 requires agencies to establish governance, innovation, and risk-management practices for AI uses that affect rights and safety, while GAO's AI accountability framework provides practical guidance on governance, data, performance, monitoring, and risk management. When these principles are read alongside recent finance-oriented studies on AML, market surveillance, and reporting integrity, a consistent message emerges: high-risk analytical systems require documented controls, transparent escalation logic, and continuous monitoring rather than one-time model deployment (Government Accountability Office, 2021; Ibrahim et al., 2024; Jahan et al., 2024; Pritty et al., 2024).

3. Research gap

Despite substantial work in each relevant domain, the literature leaves an important gap. Payment-integrity studies and official reports document the scale and causes of improper payments but rarely specify how AI systems should be designed for programme-sensitive disbursement environments. Fraud-detection research offers powerful analytical methods, yet most studies are oriented toward private-sector transactions, card payments, or general financial fraud rather than federal benefit disbursement. Digital identity guidance addresses proofing and authentication, but not how identity signals should be fused with payment and network analytics inside a public-sector integrity architecture. Public-administration and AI-governance scholarship, meanwhile, highlights trust, fairness, discretion, and transparency, but often without payment-system detail.

As a result, there is no widely articulated framework that connects four elements in a single model for the U.S. federal context: (1) pre-disbursement identity assurance, including synthetic identity prevention; (2) multimodal risk analytics over beneficiary, account, transaction, document, and network signals; (3) human-centred operational decisioning that distinguishes between payment release, stepped verification, review, and post-payment recovery; and (4) trustworthiness safeguards such as auditability, fairness testing, contestability, and continuous monitoring. The lack of such a framework matters because agencies are increasingly under pressure to modernise integrity controls without creating opaque, delay-prone, or inequitable systems.

This paper addresses that gap by proposing a federal-payment-specific framework for trustworthy AI. The framework is not presented as a completed empirical model. Rather, it is a conceptually grounded architecture derived from the available literature

and official guidance. Its value lies in integrating fragmented insights into a coherent design logic that agencies, researchers, and policymakers can adapt and test.

4. Research objectives and questions

The paper pursues three research objectives. First, it identifies the principal integrity risks facing federal payment infrastructure, with emphasis on improper payments, synthetic identities, and public-benefit disbursement vulnerabilities. Second, it develops a trustworthy AI framework that links identity assurance, multimodal analytics, and operational governance. Third, it outlines a practical evaluation template that future agency pilots could use to assess both technical performance and institutional trustworthiness.

These objectives lead to three research questions:

RQ1. What design requirements distinguish trustworthy AI for federal payment infrastructure from generic fraud-detection models?

RQ2. How should identity, payment, document, and network signals be combined to improve improper-payment prevention and synthetic identity detection in public-benefit disbursement systems?

RQ3. What governance and evaluation mechanisms are necessary to ensure that AI-enabled payment-integrity systems remain accurate, auditable, fair, and operationally legitimate?

5. Methodology

Because the paper does not rely on proprietary agency data, it adopts an integrative literature-review and conceptual framework-development design. An integrative review is suitable where the research problem spans multiple literatures with different methods, units of analysis, and publication traditions. Here, the relevant evidence includes peer-reviewed articles on fraud detection, network analytics, AI governance, and public administration; official U.S. government reports on improper payments, fraud risk, AI policy, and payment integrity; and selected authoritative standards on digital identity and AI risk management.

The review prioritised sources so that the paper is temporally plausible for a publication setting. The search strategy focused on combinations of terms such as 'improper payments', 'federal fraud', 'payment integrity', 'public-benefit disbursement', 'synthetic identity', 'digital identity assurance', 'graph fraud detection', 'trustworthy AI', 'public-sector AI accountability', and 'federal AI governance'. Peer-reviewed and official sources were prioritised over trade commentary. Appendix A summarises the search logic and Appendix B the inclusion and exclusion criteria.

The synthesis proceeded in four stages. First, the literature was mapped into four thematic domains: payment integrity, digital identity, AI-enabled fraud analytics, and governance. Second, recurring concepts and tensions were extracted, including latency versus accuracy, false-positive harm, identity uncertainty, networked fraud structures, and human oversight. Third, these concepts were assembled into a multilayer design framework oriented to federal payment infrastructure rather than general financial services. Fourth, the emerging framework was checked against external governance baselines, particularly the NIST AI RMF, NIST bias guidance, OMB M-24-10, and GAO's accountability framework. The result is a theory-informed and policy-grounded analytical model rather than a statistically estimated system.

This methodological choice imposes clear limitations. The paper cannot claim empirical performance gains, causal effects, or operational savings from the proposed framework. It can, however, offer a defensible structure for future pilots, programme-specific tailoring, and evaluation design. In that sense, the methodology is deliberately honest: it provides a strong conceptual and implementation template while reserving empirical claims for subsequent research.

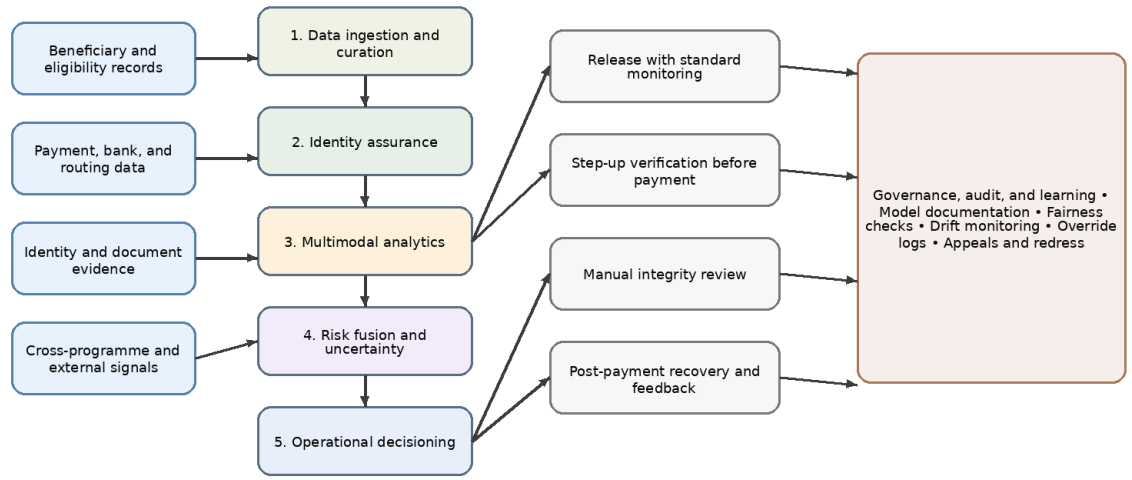
6. Proposed analytical framework

This section presents the proposed framework. It is designed for federal payment environments in which agencies must decide whether to release, hold, verify, or investigate a payment or beneficiary record under conditions of uncertainty. The framework is modular so that agencies with different programme types can adjust thresholds, data inputs, and review intensity. Its core premise is that trustworthy AI in federal payments should not be a single model; it should be a layered decision system combining pre-disbursement identity controls, multimodal analytics, network-level risk detection, human-centred adjudication, and continuous monitoring.

6.1 Design principles

Five design principles anchor the framework. First, corroboration over singularity: no consequential decision should depend on a single weak or opaque signal. Second, proportionality: the intensity of AI intervention should be matched to payment risk, programme context, and the harm of false positives. Third, traceability: every model score, threshold decision, and review action should be logged and interpretable for oversight purposes. Fourth, human accountability: the system should support, not obscure, human responsibility for consequential decisions. Fifth, adaptive governance: both fraud tactics and administrative rules evolve, so models, thresholds, and review practices must be continuously monitored and recalibrated.

Figure 1. Proposed trustworthy AI architecture for federal payment integrity



Illustrative blueprint only. Not a completed empirical deployment.

6.2 Framework architecture and operating logic

Figure 1 summarises the architecture. The first layer is data ingestion and curation. Relevant inputs include beneficiary master data, enrolment and recertification records, payment histories, bank-account and routing information, address and contact fields, death and incarceration indicators where lawfully available, device or channel metadata for online interactions, document-based evidence, cross-programme duplication signals, and investigative or adjudication outcomes. Data governance at this stage is foundational because poor entity resolution, stale records, or inconsistent beneficiary identifiers will degrade every downstream model.

The second layer is identity assurance. Here the system performs checks designed to assess whether the claimant or payee is real, current, and contextually consistent. This includes identity-proofing and authentication evidence, death-data and do-not-pay screening, account verification, SSN/TIN consistency checks, name-date-of-birth alignment, address validation, and device-account consistency where appropriate. Synthetic identity detection is not treated as a separate downstream classifier but as part of an identity-authenticity assessment. This reflects the insight that fraud often begins at the enrolment or payee-maintenance stage, not only at transaction execution.

The third layer is multimodal risk analytics. At this stage, structured models evaluate several dimensions of risk simultaneously. A beneficiary-level model estimates identity-authenticity risk. A payment-level model estimates disbursement-legitimacy risk based on timing, amount, change events, account novelties, and programme-specific indicators. A graph or network model evaluates relational risk by identifying clusters linked through shared bank accounts, addresses, phone numbers, devices, employers, vendors, or submission patterns. Where programmes rely on textual or document evidence, a document-integrity model can flag suspicious narratives, repeated templates, or inconsistent supporting information. The logic of this multimodality is consistent with research showing that narrow legacy scoring systems miss meaningful heterogeneity and network structure (Rasel et al. 2023).

The fourth layer is risk fusion and uncertainty management. Rather than collapsing all information into a single unqualified score, the framework proposes at least four outputs: an identity-authenticity score, a payment-legitimacy score, a network-risk

score, and an uncertainty indicator. The uncertainty indicator is critical because high model confidence and high risk are not the same as low confidence and high risk. In federal administration, uncertainty often means the appropriate response is additional verification rather than denial. This distinction helps avoid using AI scores as substitutes for adjudicative reasoning.

The fifth layer is operational decisioning. Four decision routes are proposed. Low-risk, high-confidence cases proceed to payment release. Moderate-risk or high-uncertainty cases are routed to stepped verification, such as additional document requests, payee confirmation, or account confirmation before release. High-risk cases with corroborating signals are sent to manual integrity review prior to payment, while post-payment anomaly flags trigger recapture, investigation, or future-prevention actions rather than retroactive blame by default. This decision architecture treats the payment system as a staged control environment, not a binary accept-reject gate.

Finally, the sixth layer is governance, audit, and learning. All consequential models should maintain documentation on intended use, training data provenance, validation boundaries, monitoring thresholds, override practices, fairness checks, and escalation paths. Human review outcomes feed back into the system as labels for recalibration, but this feedback loop must itself be audited because biased or inconsistent adjudication can contaminate training data. Figure 2 presents this operational oversight cycle.

Table 2. Multilayer framework for trustworthy AI in federal payment infrastructure

Framework layer	Core function	Illustrative data inputs	Primary integrity objective
1. Data ingestion and curation	Create a clean, auditable analytical base	Beneficiary records, payment histories, bank data, addresses, death data, claim records, review outcomes	Reduce identity and documentation uncertainty
2. Identity assurance	Validate that the claimant/payee is real and contextually consistent	Identity-proofing evidence, SSN/TIN checks, DNP matches, account verification, contact-channel consistency	Prevent synthetic, stale, or misdirected payees
3. Multimodal analytics	Estimate risk from multiple perspectives	Tabular features, event sequences, graph links, document signals, channel metadata	Detect anomalous behaviour, coordinated abuse, and suspicious updates
4. Risk fusion and uncertainty management	Combine outputs without oversimplification	Identity score, payment score, network score, uncertainty/confidence score	Separate high risk from high uncertainty
5. Operational decisioning	Translate analytics into action	Thresholds, business rules, reviewer notes, programme policy constraints	Support release, step-up verification, review, or recovery
6. Governance and learning	Maintain trustworthiness over time	Audit logs, fairness metrics, drift reports, override records, appeals outcomes	Ensure accountability, calibration, and continuous improvement

Note: The layers are modular and should be calibrated to programme-specific legal authority, data access, and payment timing.

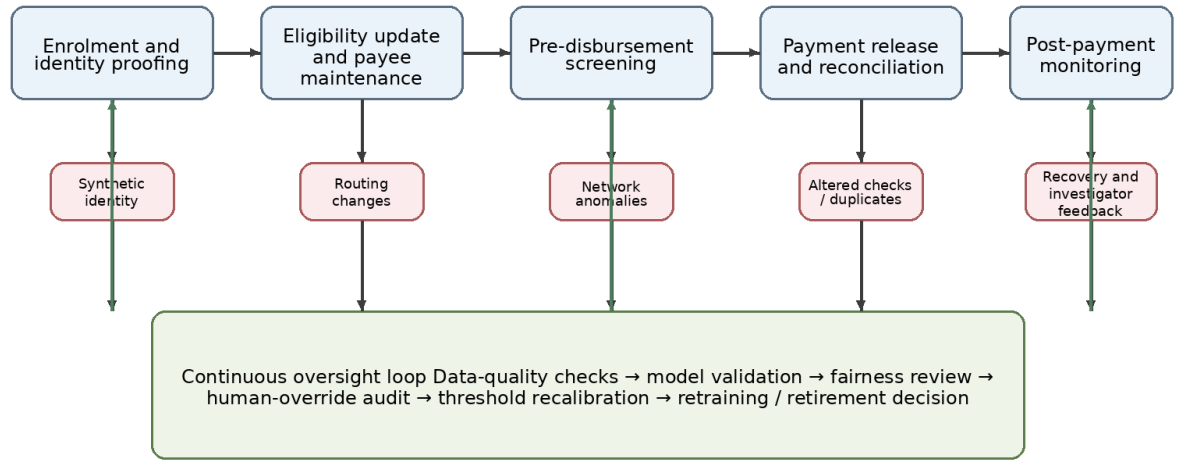
Table 3. Illustrative risk indicators for improper-payment prevention and synthetic identity detection

Risk domain	Illustrative indicators	Why the signal matters	Suggested response
Synthetic identity / enrolment risk	SSN reused across multiple households; inconsistent name–DOB combinations; recent beneficiary creation followed by rapid account change	Synthetic or partially fabricated identities often create cross-record inconsistencies or compressed activation patterns	Step-up verification or manual identity review before payment
Payee-account risk	Multiple beneficiaries mapped to one bank account; new account added	Mule accounts and misdirection schemes often	Account verification and

Risk domain	Illustrative indicators	Why the signal matters	Suggested response
	immediately before release; account closed or unverifiable	appear as account concentration or abrupt routing changes	hold pending confirmation
Death / eligibility risk	Death-data match; contradictory incarceration or eligibility indicators; missing recertification evidence	Many improper payments arise from stale eligibility or unverified status, not only intentional fraud	Eligibility re-check or administrative review
Network / organised-fraud risk	Shared address, device, employer, phone, or submission pattern across unrelated claims	Relational signals reveal coordinated abuse invisible in isolated records	Escalate to graph review and investigative triage
Document / narrative risk	Repeated text templates, inconsistent attachments, mismatch between narrative and structured fields	Generative or templated content may support fraudulent or low-integrity claims	Document-integrity review and corroboration request
Post-payment anomaly risk	Unusual recapture patterns, repeated nonreceipt claims, altered-check clusters	Some threats are only visible after payment issuance and require recovery-oriented controls	Post-payment recovery and feedback into model training

Note: These indicators are illustrative and should not be used as standalone reasons for adverse action.

Figure 2. Risk transmission and oversight cycle across the federal payment lifecycle



The framework assumes human accountability at each escalation point and does not treat model scores as final benefit decisions.

7. Results or analytical framework

Because the paper is conceptual rather than empirical, this section does not report measured results. Instead, it presents an expected-results and evaluation framework showing how the proposed architecture would be assessed in a future pilot. A rigorous federal pilot would need to evaluate not only model discrimination but also operational effects on payment timing, reviewer burden, and beneficiary treatment.

The first evaluation dimension is detection performance. Agencies should separately assess identity-authenticity, payment-legitimacy, and network-risk modules before evaluating fused outputs. This is important because composite scores can conceal weaknesses in individual components. For example, a strong transaction model may mask an unreliable identity model, encouraging premature deployment in high-risk settings.

The second dimension is operational effectiveness. A payment-integrity system that identifies more risk but imposes long delays or overwhelming review queues may not be acceptable in practice. Agencies should therefore measure manual-review rates, payment delays, step-up verification completion times, recovery yield, and the share of cases resolved without investigation. In benefit environments, speed and integrity are joint outcomes rather than separate ones.

The third dimension is governance quality. At minimum, agencies should monitor explanation availability, override rates, audit-log completeness, drift alerts, and fairness indicators. A technically strong model that cannot be audited or explained to internal reviewers is a fragile control, especially in programmes exposed to oversight, appeals, or litigation. Table 4 and Table 5 summarise the proposed decision matrix and evaluation dimensions.

Table 4. Proposed decision matrix for AI-supported federal payment integrity workflows

Decision category	Risk configuration	Operational action	Trustworthiness safeguard
Routine release	Low risk and high confidence	Release payment with standard logging	Periodic sampling and post-payment monitoring
Step-up verification	Moderate risk or elevated uncertainty	Request additional proof, confirm account details, or revalidate identity before payment	Time-bound review and beneficiary notice where appropriate
Manual integrity review	High risk supported by multiple corroborating signals	Route to trained reviewer before release or certification	Reason codes, override documentation, supervisor review
Post-payment recovery / investigation	Payment already issued but subsequent anomaly or network alert emerges	Initiate recapture, investigation, or future-payment blocks subject to law and programme rules	Audit trail, legal basis, and feedback-loop quality checks
Model governance intervention	Drift, unexplained disparity, data-quality degradation, or surge in overrides	Recalibrate, suspend, or retrain model component	Independent validation and governance-board review

Note: The decision matrix is a governance template and should be adapted to agency authority and programme rules.

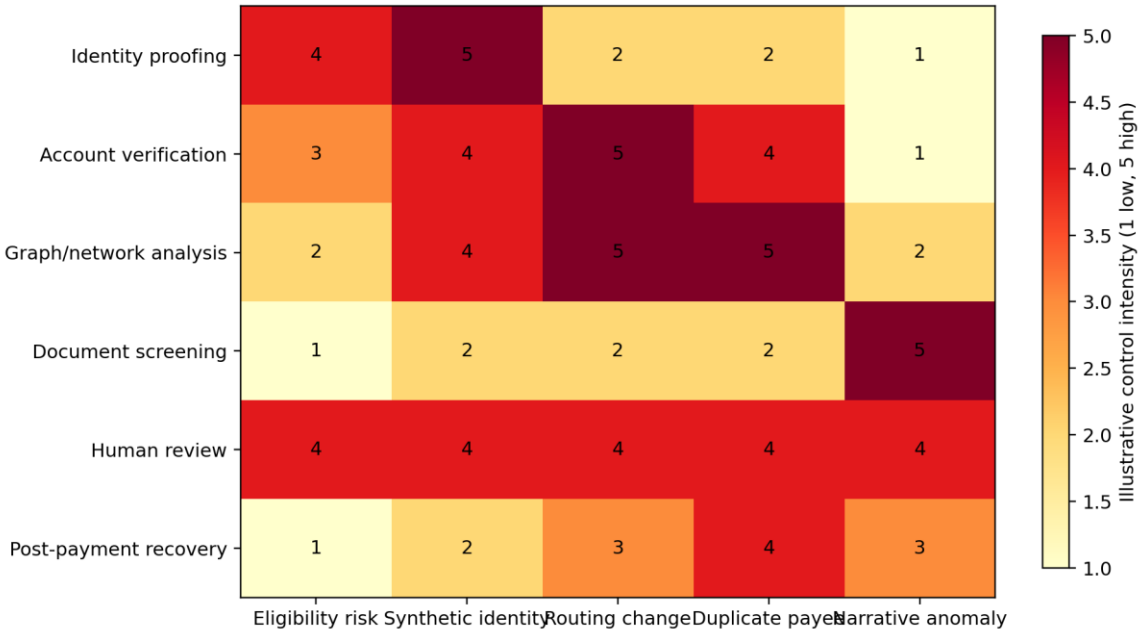
Table 5. Evaluation template for future agency pilots

Evaluation dimension	Illustrative metric	Why it matters in federal payments
Detection performance	Precision, recall, false-positive rate, false-negative rate, area under ROC/PR curves	Shows whether models identify meaningful risk without overwhelming reviewers or missing fraud
Operational efficiency	Manual-review rate, average review time, payment-release delay, recovery yield	Connects model outputs to service delivery and resource burden
Trustworthiness and governance	Explanation availability, override rate, audit-log completeness, drift alerts resolved	Assesses whether the system is auditable and governable rather than merely accurate

Evaluation dimension	Illustrative metric	Why it matters in federal payments
Equity and fairness	Error-rate disparities, escalation-rate disparities, verification burden by group where lawful and appropriate	Helps identify whether the system imposes disproportionate friction or risk on certain populations
Identity-integrity effectiveness	Synthetic identity detection rate, duplicate-payee discovery rate, account-verification hit rate	Measures whether upstream identity controls are reducing downstream improper payments
Programme integrity impact	Improper payments prevented or recovered; confirmed fraud referrals; reduction in repeat anomalies	Links AI outputs to core public-value objectives

Note: This table is an evaluation blueprint pending agency data and programme-specific validation.

Figure 3. Illustrative control-intensity heatmap for federal payment-risk scenarios



8. Discussion

The proposed framework contributes to the literature by reframing federal payment integrity as a joint problem of analytics, administration, and public legitimacy. Too often, the debate is framed as a choice between more aggressive fraud detection and faster benefit delivery. The literature reviewed here suggests that this is the wrong framing. Better design can improve both integrity and service quality when risk signals are layered, thresholds are calibrated, and review is targeted rather than indiscriminate. Shared services such as account verification, death-data matching, and cross-government analytics can reduce upstream uncertainty, thereby reducing the need for blunt downstream interventions.

The framework also clarifies why synthetic identity risk requires more than stronger biometric or document checks. Synthetic identity is best understood as a relational and temporal phenomenon. A fabricated or partially fabricated beneficiary may appear plausible in one record at one point in time, yet look suspicious once linked to other payments, accounts, or entities. That is why graph and network analytics are not peripheral to payment integrity; they are central. The same reasoning underpins work on multimodal graph approaches in adjacent financial contexts (Rasel et al. 2023).

At the same time, the discussion must remain institutionally grounded. In a public-benefit system, there is a meaningful difference between a fraud-alerting model and a rights-affecting decision system. A model that flags applicants or payments can

alter frontline behaviour even when the model does not make the final decision. Trustworthy AI therefore requires procedural safeguards around the human use of algorithmic advice, not merely technical improvements in model performance.

The paper further suggests that public-sector trustworthiness should be interpreted operationally. In this context, trustworthiness means that an agency can answer six practical questions: What signals were used? Why was a case flagged? What threshold triggered intervention? How often are false positives reviewed? Which groups experience more friction or delay? What happens when the model drifts or an adversary adapts? This operational interpretation helps translate abstract governance principles from GAO, NIST, and OMB into payment-system practice (Government Accountability Office 2021; National Institute of Standards and Technology 2023; Office of Management and Budget 2024).

9. Theoretical implications

The paper makes three theoretical contributions. First, it extends financial-fraud analytics into the domain of federal administrative infrastructure. Much existing fraud research assumes private-sector incentives, loss functions, and customer relationships. Federal payment systems differ because they combine anti-fraud objectives with statutory eligibility, public-law obligations, and service-delivery mandates. The framework proposed here therefore contributes to a more institutionally specific theory of AI-enabled financial integrity.

Second, the paper links digital identity assurance to payment-integrity theory. Rather than treating identity proofing and transaction monitoring as separate layers of administration, it shows that identity uncertainty is often the earliest detectable form of payment risk. This theoretical move helps explain why payment integrity should be organised around lifecycle controls, from enrolment to disbursement to post-payment monitoring, instead of being confined to ex post fraud investigation.

Third, the paper contributes to trustworthy-AI theory by showing that trustworthiness in public financial systems is inseparable from disbursement logic. In other words, fairness, transparency, accountability, and monitoring are not external ethics principles appended to a technical system; they are constitutive features of any payment-integrity model that is expected to operate legitimately in the public sector. This aligns with broader AI-governance scholarship while grounding it in a specific institutional use case (Government Accountability Office 2021; National Institute of Standards and Technology 2023).

10. Practical, policy, and managerial implications

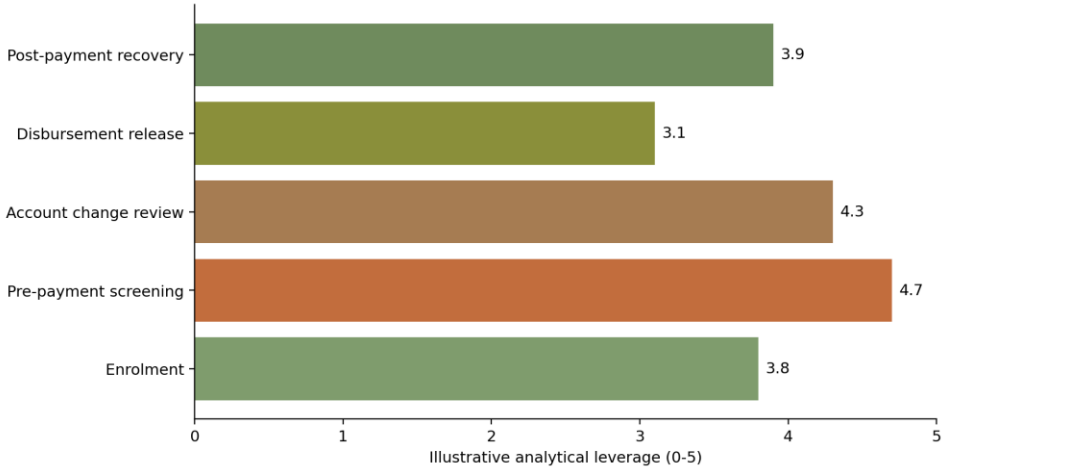
For practitioners and policymakers, the first implication is that payment integrity should be treated as shared infrastructure rather than programme-by-programme improvisation. Treasury's recent experience suggests that common services such as DNP, account verification, death-data matching, and cross-government analytics can create scale efficiencies and improve consistency. A trustworthy AI agenda should build on this shared-services model by creating reusable but programme-configurable components: entity resolution, graph linkage, model documentation, fairness dashboards, and reviewer workbenches.

Second, agencies should adopt a tiered intervention model. Not every anomaly should stop a payment, and not every identity inconsistency should trigger an adversarial investigation. A graded response—release, step-up verification, manual review, and post-payment monitoring—better aligns analytical outputs with due-process concerns and operational realities. This is especially important for programmes that serve vulnerable beneficiaries, where indiscriminate friction can itself become a form of administrative harm.

Third, agencies should invest in data quality and workforce capability before overcommitting to model complexity. In practice, this means improving canonical beneficiary records, match confidence, event timestamps, feedback labelling, and reviewer training. The model most likely to fail in a federal payment environment is not necessarily the least accurate one in theory; it is the one that is least aligned with data reality and institutional capacity.

Fourth, accountability mechanisms should be designed into procurement and operations. Model cards, audit logs, override tracking, periodic bias testing, drift analysis, and external review protocols should be required for all high-impact payment-integrity models. Work on algorithmic accountability and risk frameworks in digital financial systems reinforces the importance of governance mechanisms, not only predictive power, in high-stakes decisioning (FAHIM et al. 2023). This infrastructure-oriented perspective is also consistent with Hasan et al. (2023), who frame fraud analytics as part of a wider resilience architecture rather than a standalone compliance instrument.

Figure 4. Example staging of AI-assisted integrity leverage across the payment lifecycle



11. Limitations

This paper has several limitations. Most importantly, it is conceptual and literature-based. It does not estimate model performance, compare alternative classifiers on federal data, or quantify cost savings. The proposed framework should therefore be read as a design blueprint rather than evidence of realised operational superiority.

Second, federal payment systems vary widely in legal authority, data access, payment timing, evidentiary rules, and agency capacity. A model suitable for tax-refund integrity may be poorly calibrated for disaster relief, healthcare reimbursement, or social insurance administration. The framework must therefore be tailored at programme level rather than implemented as a one-size-fits-all system.

Third, the paper focuses on U.S. federal institutions and authoritative U.S. guidance. It does not attempt a comparative international analysis, even though similar issues arise in other jurisdictions. Fourth, it does not address every relevant legal issue in depth, such as specific Privacy Act constraints, constitutional due-process questions, or contractor liability in model procurement. These are important areas for future specialised work.

12. Future research directions

The most immediate next step is empirical validation. Future research should test the framework with agency or synthetic datasets that allow comparison of identity-only, payment-only, and multimodal network-aware models. Such studies should report not only accuracy metrics but also delay effects, reviewer burden, recapture yield, and disparity indicators across beneficiary groups.

A second avenue is adversarial testing. Because synthetic identity and disbursement fraud are adaptive, future work should simulate how offenders respond to stricter account verification, stronger graph detection, or changed thresholds. Red-team exercises, scenario analysis, and longitudinal drift studies would be especially useful.

Third, researchers should examine human-AI interaction in payment-integrity workflows. Reviewer behaviour matters as much as model output in high-stakes administrative systems. Field experiments or audit studies could examine how explanations, confidence intervals, reviewer interfaces, and escalation protocols affect outcomes.

Fourth, privacy-preserving architectures deserve greater attention. Shared infrastructure for payment integrity will increasingly require privacy-enhancing record linkage, secure multiparty computation, or federated learning approaches capable of combining signals across agencies without excessive data centralisation. Future work should also investigate how synthetic document generation and other rapidly evolving fraud techniques may change both offence and defence in document-intensive benefit environments.

13. Conclusion

Federal payment integrity is often discussed as a control problem, but it is more accurately a governance problem executed through control systems. The challenge is not only to detect more suspicious cases; it is to do so in a way that protects eligible recipients, preserves administrative legitimacy, and scales across heterogeneous programmes. Existing work on improper

payments, fraud risk, digital identity, and trustworthy AI contains most of the necessary components, but these components have rarely been integrated into a framework specific to U.S. public-benefit disbursement.

This paper has argued that trustworthy AI offers the right organising concept. A federal payment-integrity architecture should combine identity assurance, multimodal and graph-based analytics, graded operational responses, and continuous governance mechanisms around fairness, transparency, and accountability. Such a model is more institutionally realistic than a purely predictive one because it acknowledges that the public value of payment integrity lies not only in loss prevention but also in rightful payment, explainable intervention, and sustained public trust. This conclusion is consistent with recent scholarship showing that effective high-stakes analytics depend on the joint design of predictive performance, governance controls, and infrastructure resilience rather than on model accuracy alone (Hasan et al., 2023; Ibrahim et al., 2024; Jahan et al., 2024).

The proposed framework is intentionally forward-looking and non-empirical. Its contribution is to provide a rigorous, literature-based blueprint that agencies and researchers can adapt, test, and refine. In a period when the federal government is simultaneously modernising payments and tightening AI governance, that blueprint is both timely and practically consequential.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1]. FAHIM, A. S. M., Ibrahim, M., Pritty, A. A., & Tania, T. A. (2023). Algorithmic accountability in U.S. consumer FinTech: Governance mechanisms for credit risk, fair lending, and financial stability. Journal of Economics, Finance and Accounting Studies, 5(4), 80–93. https://doi.org/10.32996/jefas.2023.5.4.8

[2]. Government Accountability Office. (2021). Artificial intelligence: An accountability framework for federal agencies and other entities (GAO-21-519SP). https://www.gao.gov/products/gao-21-519sp

[3]. Government Accountability Office. (2024a). Fraud risk management: 2018–2022 data show federal government loses an estimated $233 billion to $521 billion annually to fraud, based on various risk environments (GAO-24-105833). https://www.gao.gov/products/gao-24-105833

[4]. Government Accountability Office. (2024b). Improper payments: Key concepts and information on programs with high rates or lacking estimates (GAO-24-107482). https://www.gao.gov/products/gao-24-107482

[5]. Hasan, M. N., Rasel, I. H., Arman, M., Ibrahim, M., & Jahan, N. (2023). Strengthening U.S. financial and cybersecurity infrastructure with AI-driven fraud detection and risk analytics. Journal of Computational Analysis and Applications, 31(2), 15–32. https://www.eudoxuspress.com/index.php/pub/article/view/3823

[6]. Ibrahim, M., Mahmud, S., Zadid, M. U., Jahan, N., Rahman, M. M., & FAHIM, A. S. M. (2024). AI-driven predictive analytics framework for anti-money laundering risk management and financial infrastructure protection in U.S. banking systems. Journal of Economics, Finance and Accounting Studies, 6(1), 155–166. https://doi.org/10.32996/jefas.2024.6.6.12

[7]. Ibrahim, M., Razib, M. N. H., Jahan, N., & Rahman, M. M. (2022). Climate risk, financial stability, and global capital allocation: A predictive analytics approach to assessing climate-related financial risk in international investment markets. Journal of Business and Management Studies, 4(4), 264–276. https://doi.org/10.32996/jbms.2022.4.4.34

[8]. Jahan, N., Pritty, A. A., Ibrahim, M., Zadid, M. U., FAHIM, A. S. M., & Mahmud, S. (2024). Machine learning-driven early warning analytics for identifying market manipulation, irregular trading activity, and suspicious market signals in U.S. stock markets. Journal of Computer Science and Technology Studies, 6(2), 257–283. https://doi.org/10.32996/jcsts.2024.6.2.26

[9]. National Institute of Standards and Technology. (2020). Digital identity guidelines (NIST Special Publication 800-63-3). https://pages.nist.gov/800-63-3/

[10]. National Institute of Standards and Technology. (2022). Towards a standard for identifying and managing bias in artificial intelligence (NIST Special Publication 1270). https://doi.org/10.6028/NIST.SP.1270

[11]. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0) (NIST AI 100-1). https://doi.org/10.6028/NIST.AI.100-1

[12]. Office of Management and Budget. (2024). Advancing governance, innovation, and risk management for agency use of artificial intelligence (Memorandum M-24-10). Executive Office of the President. https://www.whitehouse.gov/wp-content/uploads/2024/03/M-24-10-Advancing-Governance-Innovation-and-Risk-Management-for-Agency-Use-of-Artificial-Intelligence.pdf

[13]. Pritty, A. A., Ibrahim, M., FAHIM, A. S. M., & Zadid, M. U. (2024). Generative AI and U.S. financial reporting integrity: Detecting narrative manipulation, risk disclosure gaming, and fraud signals in 10-K filings. Journal of Economics, Finance and Accounting Studies, 6(4), 113–129. https://doi.org/10.32996/jefas.2024.6.4.11

[14]. Rasel, I. H., Ibrahim, M., Pritty, A. A., FAHIM, A. S. M., & Jahan, N. (2023). Beyond FICO: Enhancing mortgage default forecasting and inclusive lending via multimodal graph neural networks and urban mobility analytics. Frontiers in Computer Science and Artificial Intelligence, 2(2), 62–81. https://doi.org/10.32996/fcsai.2023.2.2.5

[15]. Ryman-Tubb, N. F., Krause, P., & Garn, W. (2018). How artificial intelligence and machine learning research impacts payment card fraud detection: A survey and industry benchmark. Engineering Applications of Artificial Intelligence, 76, 130–157. https://doi.org/10.1016/j.engappai.2018.07.008

[16]. West, J., & Bhattacharya, M. (2016). Intelligent financial fraud detection: A comprehensive review. Computers & Security, 57, 47–66. https://doi.org/10.1016/j.cose.2015.09.005