
Deepfake Enabled Financial Fraud and Consumer Trust in Digital Finance: Can Vulnerability–Trust–Resilience Framework Be Applied in the United States of America?

Nenbee, Simeon Gbimoie (PhD)¹, Golley, Israel Tega², Mgbataogu, Ifeanyi Sunday (PhD)³

¹Department of Economics, University of Port Harcourt

²Graduate School of Economics and Management University of Port Harcourt, <https://orcid.org/0009-0008-2609-8612>

³Department of Finance and Banking, University of Port Harcourt

ABSTRACT: Generative artificial intelligence has made convincing synthetic identities, voices, images, and video cheaper and easier to produce. Financial institutions now face attacks directed not only at technical authentication, but also at the social cues through which customers and employees decide whom to trust. This study combines descriptive analysis of public United States fraud statistics with directed qualitative content analysis of 18 (eighteen) authoritative regulatory, law-enforcement, standards, and policy documents published between 2019 and 2025. Eight response mechanisms were coded: accountable ownership, identity verification, deepfake detection and provenance, transaction monitoring, human authentication and escalation, customer information, incident response and information sharing, and third-party governance. Incident response appeared in about 94% of the corpus while identity verification stood at 89%. Human authentication and transaction monitoring were present in 83% and 72%, respectively. Half of the documents addressed provenance or deepfake detection, whereas only 39% connected counter-fraud controls with customer communication. Public loss data confirm a rapidly worsening fraud environment but do not identify losses caused specifically by deepfakes, producing an attribution-denominator problem. The proposed Vulnerability–Trust–Resilience (VTR) framework shows how synthetic media can turn weaknesses in authentication and judgment into financial loss and wider distrust. Resilience therefore depends on layered verification, transaction-context controls, empowered human escalation, rapid recovery, and carefully calibrated communication—not on a single detector. It is therefore recommended that there is need to urgently develop testable propositions, a managerial scorecard, and a reproducible engagement agenda by the government to forestall the growing incidence of deepfake enabled financial fraud in the financial sector of the economy.

KEYWORDS: Deepfake fraud; consumer trust; digital finance; identity verification and generative artificial intelligence

1. INTRODUCTION

Digital finance rests on trust that must function at scale to various categories of customers. Customers often times open accounts remotely, authorize payments on mobile devices, obtain advice through electronic channels, and increasingly communicate with automated agents. Most customers are employees in the digital finance space and make consequential decisions using video meetings, telephone calls, emailed instructions, identity documents, and digital signatures. In each instance, the institution converts incomplete identity and behavioral signals into judgments about access, credit, payment, and responsibility. In recent years, generative Artificial Intelligence (AI) alters the economics of attacking those judgments. Faces, voices, documents, and persuasive dialogue can now be synthesized or manipulated cheaply, across languages, and at operational scale thereby posing threat to consumption patterns and habits.

The threat extends well beyond conspicuous fabricated video. FinCEN (2024a) reported increasing suspicious activity involving suspected deepfake media in schemes directed at institutions and their customers, including altered or fabricated identity documents intended to defeat verification and authentication. Synthetic voices can even make impersonation more credible in such scenarios.

Deepfake Enabled Financial Fraud and Consumer Trust in Digital Finance: Can Vulnerability–Trust–Resilience Framework Be Applied in the United States of America?

Manipulated faces may target onboarding or liveness controls. Generated documents can facilitate account takeover, mule-account creation, and payment fraud, while personalized messages make social engineering harder to recognize. The appropriate unit of analysis is therefore the fraud process—not the media artifact in isolation—from reconnaissance and persuasion through identity compromise, transaction execution, and concealment.

The Federal Bureau of Investigation’s Internet Crime Complaint Center (IC3, 2025) did received 859,532 complaints for the year 2024 and \$16.6 billion in losses—33% above the 2023 total. Those figures are not a measure of deepfake losses; incident reports do not consistently record whether synthetic media played a role. They do, however, describe the environment into which scalable impersonation technology has entered.

One sense about the IC3(2025) report is that the aggregate fraud statistics establish urgency and exposure, but they cannot estimate the marginal causal contribution of deepfakes. The damage can extend beyond direct victimization. Once customers believe that voices, faces, documents, or institutional messages can be fabricated at will, legitimate interactions become harder and more expensive to verify. Some may delay valid transfers or retreat from remote channels; others may encounter a “liar’s dividend”, the term Chesney & Citron (2019) use for the way public awareness of synthetic media lets wrongdoers dismiss genuine, incriminating evidence as fake. More specifically, institutional responses can create their own harms if added friction excludes customers without advanced devices, stable connectivity, conventional identity records, or accessible biometric options. Governance must therefore pursue two objectives at once: prevent successful fraud and preserve justified confidence and equitable access.

Arising from the above, latitudes of research on deepfakes are dispersed across various disciplines like computer science, cybersecurity, criminology, psychology, information systems, and sorts. These fields frequently examine different parts of the same problem. Detection studies tend to report benchmark accuracy under controlled conditions, whereas fraud advisories concentrate on operational red flags. Trust research on the hand explores reliance on technology but seldom follows the consumer through post-incident recovery while financial-risk documents acknowledge cyber and model risk without always separating synthetic-media controls from conventional anti-fraud practice. This fragmentation leaves three gaps: First, detection alone cannot carry the control burden. A classifier may perform strongly on familiar generation methods and then deteriorate after compression, adversarial alteration, model change, or deployment in a different population. Losses also arise when a customer or employee follows a persuasive instruction that no detector ever evaluates. Effective defense must span identity, device, behavior, transaction context, human challenge, and recovery. Second, consumer trust is often treated as an attitude rather than an outcome shaped by governance. Confidence may be damaged by the original fraud, but also by the institution’s response, unclear warnings, mistaken account restrictions, or a burdensome reimbursement process. Proportionate and understandable controls can reinforce trust. Repeated false positives and unequal exclusion can erode it. Third, the empirical denominator is missing. Public systems record complaint categories and losses, while institutions apply heterogeneous internal labels. “Deepfake-enabled” may mean that synthetic media was essential to the scheme, that it merely strengthened persuasion, or that its use was suspected but unconfirmed. Without a stable taxonomy, prevalence, institutional comparison, and control effectiveness remain difficult to estimate.

Accordingly, this paper sets to provide answers to the following questions and more: (a) Which organizational and technical response mechanisms recur in authoritative guidance on synthetic-media and digitally enabled financial fraud? (b) Where should the current frameworks under-specify consumer trust, content provenance, and outcome be measured? and (c) Through what pathways do deepfake-enabled attacks and institutional responses affect adoption, confidence, inclusion, and resilience in digital finance? These questions and more are interrogated in this study. The rest of the paper addresses literature review, methodology, results and discussion and concluding remarks.

2. LITERATURE REVIEW

2.1 Survey of Theoretical Literature

Routine activity theory explains predatory events as a convergence of motivated offenders, suitable targets, and insufficient guardianship (Cohen & Felson, 1979). Digital finance broadens the reach of potential targets and permits remote contact; generative AI adds credibility and scale. Guardianship, however, is distributed across the system. A customer may question an unusual request, a bank may detect anomalous device behavior, a telecommunications provider may restrict spoofing, and law enforcement may combine reports across victims. This perspective directs analysis away from the artifact alone and toward the opportunity structure surrounding it.

Deepfake Enabled Financial Fraud and Consumer Trust in Digital Finance: Can Vulnerability–Trust–Resilience Framework Be Applied in the United States of America?

Protection motivation theory helps explain why some warnings prompt useful action and others do not. People assess both the seriousness of a threat and whether an effective, manageable response is available (Rogers, 1975). Deepfake warnings may improve behavior when customers recognize a credible risk and know how to respond. Alarm without a practical verification step may instead create helplessness or generalized distrust. Customer education should therefore strengthen response efficacy and self-efficacy. Telling people to distrust everything is unworkable; instructing them to verify unusual payment requests through a known, independent channel is actionable.

Trust in automation research distinguishes appropriate reliance from maximal reliance. Lee and See (2004) define trust as an attitude that an agent will help achieve goals under uncertainty and show that calibration matters. Under-trust wastes useful automation; over-trust creates misuse. Deepfake fraud produces a related calibration problem across channels: consumers and employees must neither accept every audiovisual cue nor reject all legitimate remote interaction. Institutions support calibration by displaying meaningful signals, explaining protective processes, and providing recovery when signals fail.

Organizational resilience concerns the capacity to anticipate, absorb, recover from, and adapt after disruption. Synthetic-media fraud is adaptive: controls alter offender behavior, new generation tools appear, and attacks migrate between channels. Resilience therefore requires feedback from complaints, confirmed cases, false positives, customer outcomes, and external intelligence. A static annual policy is insufficient when attack signatures change weekly.

Institution-based trust adds the rules, safeguards, and structural assurances surrounding a transaction (McKnight *et al.*, 2002). Customers need not understand a detection model to judge whether a bank provides credible safeguards, understandable notices, accessible escalation, and fair remediation. This perspective makes trust an outcome of governance design rather than merely communication or brand reputation.

The *Vulnerability–Trust–Resilience* (VTR) framework developed in this article joins these perspectives. Vulnerability occurs when synthetic content exploits dependence on familiar signals, procedural gaps, urgency, or authority. Trust is shaped by the attack and by what the institution visibly does afterward. Resilience appears when controls break the fraud chain, contain loss, restore legitimate access, and turn incidents into learning. The proposed routes to value are therefore interruption, calibration, recovery, and adaptation. Inferencing the expressed views in literature, a joint advisory from the Cybersecurity and Infrastructure Security Agency, the Federal Bureau of Investigation, and the National Security Agency (2023) framed deepfakes as a cross-sector organizational risk rather than a purely technical curiosity, recommending governance, detection, and incident-response measures in combination. It is pertinent to ascertain the Case-based scholarship which reinforces this organizational framing. In de Rancourt-Raymond & Smaili's (2023) work that examined publicly reported corporate deepfake-fraud incidents, it argued that the ethical and governance failures exposed by these cases extend well beyond any single technical detection gap. Early deepfake research concentrated on generation and detection. The technical literature describes an arms race: artifacts that reveal one generation method can disappear in later models, while distribution through platforms alters resolution and strips metadata. NIST (2024b) accordingly treats provenance, watermarking, detection, testing, and incident disclosure as complementary controls rather than promising a definitive classifier. Accuracy measured on a balanced laboratory dataset also does not translate directly into positive predictive value when attacks are rare. Even a strong detector may overwhelm reviewers when genuine interactions vastly outnumber fraudulent ones.

Identity research emphasizes document authenticity, biometrics, presentation-attack detection, and liveness. These controls address remote onboarding and account access, but each has boundary conditions. Biometric systems may perform differently across demographic groups, accessibility conditions, devices, and environments. Active liveness challenges add friction and may be imitated or socially engineered. Document checks can fail when synthetic identities combine real and fabricated elements. Layering is therefore essential: identity evidence should be connected to device reputation, behavioral patterns, account history, and transaction purpose. The scale of the underlying problem is already substantial: FinCEN's (2024b) analysis of Bank Secrecy Act filings found that identity-related concerns accounted for 42% of suspicious activity reports in 2021, representing \$212 billion in suspicious activity, most commonly involving impersonation intended to defeat verification.

Fraud management adds transaction monitoring and case investigation. An authenticated person can still be manipulated into authorizing a payment, and a deepfake can target an employee outside customer-login controls. Contextual signals—new beneficiary, unusual amount, time pressure, channel switching, device change, atypical location, or inconsistent narrative—may remain informative even when audiovisual content looks authentic. FinCEN's (2024a) red flags reflect this system view by combining media anomalies, identity inconsistencies, device indicators, and transaction behavior.

Deepfake Enabled Financial Fraud and Consumer Trust in Digital Finance: Can Vulnerability–Trust–Resilience Framework Be Applied in the United States of America?

Cybersecurity guidance emphasizes preparation and information sharing. The U.S. Department of the Treasury (2024a) identified data supply-chain, third-party, and capability gaps affecting financial-sector management of AI-specific cyber risks, while a companion report on broader AI adoption in financial services documented similar concerns around data quality, third-party dependence, and workforce capability (U.S. Department of the Treasury, 2024b). NIST's AI Risk Management Framework organizes activity around govern, map, measure, and manage (NIST, 2023), while the generative-AI profile applies those functions to risks including confabulation, information integrity, privacy, and security (NIST, 2024a). These frameworks support lifecycle governance but are not specialized fraud-loss measurement standards. The International Organization for Standardization's (2023) AI management-system standard complements this lifecycle view by specifying organizational processes for risk assessment and continual improvement, while the European Union Agency for Cybersecurity's (2022) threat-landscape analysis situates deepfake-enabled fraud within a broader pattern of AI-enabled cybercrime observed across European institutions. Reflecting similar concerns domestically, the Federal Reserve Board (2025) has highlighted deepfakes as an emerging element of the bank cybersecurity arms race, requiring institutions to treat synthetic-media risk as a standing agenda item rather than a periodic review topic.

Financial-stability analysis broadens the scope. The FSB (2024) identifies AI-related vulnerabilities involving third-party dependencies, cyber risks, market correlations, and model risk, data quality, and governance. It also recognizes that generative AI can increase fraud and disinformation. The systemic concern is not simply that one customer is deceived, but that common providers, shared authentication assumptions, or synchronized misinformation could produce correlated losses and runs on confidence. The Financial Action Task Force (2023) situates this concern within the broader landscape of cyber-enabled fraud, noting that illicit proceeds increasingly move through layered digital channels that blend social engineering, compromised credentials, and rapid fund transfers, complicating both detection and recovery.

Consumer trust scholarship suggests that competence, benevolence, integrity, predictability, and structural assurance matter. After fraud, customers observe whether an institution responds promptly, communicates respectfully, explains decisions, restores access, and allocates loss fairly. A technically successful control that repeatedly locks out legitimate customers can lower trust, especially among vulnerable groups. Conversely, transparent verification can signal competence. Security and customer experience are not separate workstreams; they jointly determine perceived reliability. Experimental evidence from adjacent digital-trust research supports this pattern: Muzatko & Bansal (2024) find that prompt, transparent breach disclosure combined with concrete protective measures restores consumer trust more effectively than delayed or vague communication, reinforcing that the manner and speed of institutional response, not only the incident itself, shapes post-incident trust.

Public fraud statistics establish scale but offer weak attribution to a specific mechanism. IC3 reports losses across cyber-enabled crime, and the Federal Trade Commission (2025) aggregates consumer reports across fraud categories. Underreporting, awareness, duplicate or incomplete records, and classification decisions all affect these systems. Their proper interpretation is as indicators of reported loss, not estimates of population prevalence. Most importantly, historical records do not contain a consistent deepfake field. A precise national estimate of deepfake-caused losses is therefore not credible without case-level classification.

The literature seemingly reveals five gaps. They are: (i) technical detection is weakly integrated with transaction and organizational controls. (ii), trust is rarely measured before and after an incident using comparable instruments. (iii), deepfake involvement lacks standardized reporting levels. (iv), false-positive, accessibility, and inclusion costs are underdeveloped. (v), few studies connect control maturity to verified loss reduction using longitudinal or quasi-experimental designs. The present review organizes convergent mechanisms and makes these missing quantities explicit.

3. METHODOLOGY

This paper uses a structured integrative review with directed qualitative content analysis and descriptive statistics secondary-data analysis. This approach is appropriate because the research question concerns institutional convergence and conceptual integration across heterogeneous authoritative sources. It does not estimate causal effects or population prevalence. The unit of analysis for qualitative coding is one public document. The quantitative component consists of code frequencies and reported-loss trends reproduced from official sources.

The analysis followed five stages: problem formulation; transparent source identification; eligibility screening; deductive coding with documented decision rules; and synthesis through frequencies, co-occurrence, negative cases, and theory building. The review is

Deepfake Enabled Financial Fraud and Consumer Trust in Digital Finance: Can Vulnerability–Trust–Resilience Framework Be Applied in the United States of America?

structured rather than claimed as a systematic review of all academic literature. Its corpus intentionally prioritizes documents with direct authority over financial, cybersecurity, identity, fraud, or consumer-protection practice.

3.2 Population and Sampling

The conceptual population comprises English-language public guidance, standards, reports, alerts, and strategy documents addressing synthetic media, generative-AI security, digital identity, cyber-enabled fraud, or financial-sector AI risk. Purposive maximum-variation sampling was used across issuer type and jurisdiction. Sources were eligible if they (a) were issued by a regulator, law-enforcement body, standards institution, intergovernmental body, or financial authority; (b) contained operational recommendations or documented threat evidence; (c) were available in full; and (d) were published from 2019 through 2025, with foundational identity or cybersecurity guidance retained when necessary.

Documents were excluded when they were promotional vendor material, duplicated another source, offered commentary without traceable evidence, or mentioned AI without a relevant fraud, identity, cyber, or trust mechanism. Eighteen documents were retained. U.S. sources were emphasized because the intended contribution concerns U.S. financial practice, while international sources were included to test convergence and improve transferability.

The sampling strategy creates clear boundaries. It does not measure the whole scholarly literature, confidential bank practice, or criminal prevalence. Authoritative documents may reproduce one another and may privilege institutional perspectives. Those limitations are handled through explicit claims: frequency indicates documentary emphasis, not real-world effectiveness.

3.3 Data Collection and Variables

Bibliographic fields included document identifier, issuer, title, year, jurisdiction, issuer type, and URL. Eight binary mechanism variables were defined before final coding:

- **OW—Accountable ownership:** named governance responsibility, executive oversight, defined decision rights, or formal accountability.
- **IV—Identity verification:** identity proofing, authentication, biometrics, liveness, document validation, or independent identity corroboration.
- **DD—Deepfake detection and provenance:** synthetic-content detection, watermarking, provenance, authenticity metadata, or media forensics.
- **TM—Transaction monitoring:** behavioral analytics, device intelligence, payment anomaly detection, holds, or risk-based transaction review.
- **HA—Human authentication and escalation:** call-back procedures, challenge protocols, dual authorization, employee escalation, or authority to pause.
- **CI—Customer information and trust:** warnings, consumer education, understandable communication, complaint access, inclusion, or trust restoration.
- **IR—Incident response and information sharing:** case reporting, suspicious activity reporting, recovery, law-enforcement referral, threat intelligence, or post-incident learning.
- **TP—Third-party and platform governance:** vendor controls, cloud/model dependencies, telecommunications or platform coordination, contractual assurance, or supply-chain risk.

A code was assigned 1 only where the document contained a substantive recommendation, control, risk discussion, or implementation expectation. Passing mention was coded 0. Coding was conservative to reduce interpretive inflation. A document could receive multiple codes because layered controls are not mutually exclusive.

Public trend data were transcribed from annual IC3 reports. The measures are complaint counts, reported dollar losses, loss per complaint, year-over-year loss growth, and compound annual growth. The series is contextual rather than a deepfake-specific outcome. Dollar values are nominal. No imputation was used.

3.4 Data Analysis

For mechanism j across N documents, prevalence was calculated as:

$$Prevalence(j) = \sum code(ij) / N \times 100.$$

Thereafter, the pairwise co-occurrence was examined to identify control bundles. A co-occurrence means only that two mechanisms appear in the same document; it does not prove complementarity or effectiveness. Temporal comparisons separated 2019–2022 and

Deepfake Enabled Financial Fraud and Consumer Trust in Digital Finance: Can Vulnerability–Trust–Resilience Framework Be Applied in the United States of America?

2023–2025 sources to explore changes after rapid generative-AI diffusion. Because the sample is purposive and small, no null-hypothesis significance tests were applied. Inferential p values would imply a probability sample that does not exist.

For the IC3 series, annual loss growth was calculated as

$$(Loss_t / Loss_{t-1} - 1) \times 100.$$

Compound annual growth from 2021 through 2024 was calculated as

$$(Loss_{2024} / Loss_{2021})^{(1/3)} - 1.$$

Ratios were rounded after calculation. These transformations are descriptive.

Theory development used pattern matching. Codes were mapped to four VTR functions: interruption (IV, DD, TM, HA), calibration (CI and explainable verification), recovery (IR), and adaptation (OW, IR, TP). Negative-case analysis examined documents that discussed advanced detection without consumer recovery, or customer awareness without operational controls. These cases prevented the synthesis from equating any single mechanism with resilience.

3.5 Validity, Reliability, and Research Integrity

Construct validity was supported by operational definitions and a binary coding rule. Source triangulation covered law enforcement, regulation, standards, and international coordination. An audit trail is provided in Appendix A and the accompanying CSV. Quantitative calculations can be reproduced from Table 1.

Reliability is constrained by single-author coding. To reduce drift, coding was conducted in two passes separated by a full review of definitions. Ambiguous cases were resolved toward 0. A future study should use at least two independent coders, report Cohen’s kappa or Krippendorff’s alpha, and publish passage-level evidence. The current matrix should be treated as a transparent, revisable research instrument.

Internal validity is limited because document prevalence cannot demonstrate that mechanisms reduce losses or increase trust. The paper uses terms such as “supports,” “is consistent with,” and “plausibly contributes” rather than causal language. External validity is bounded by English-language public sources and institutional contexts. Public statistics understate unreported events and may include changing reporting practices.

Research integrity required a strict non-fabrication rule. No interviews, survey participants, bank-level incidents, or effect sizes were invented. Aggregate fraud figures are not relabeled as deepfake losses. The article is an original synthesis, but no author or software can promise a particular similarity-check percentage because databases and algorithms vary. Proper attribution, quotation avoidance, and independent phrasing were used throughout.

4. RESULTS

4.1 Public Fraud-Loss Context

Table 1: Reported Internet-Crime Complaints and Losses in the United States, 2021–2024

Year	Complaints	Reported losses (US\$ billions)	Loss per complaint (US\$)
2021	847,376	6.9	8,143
2022	800,944	10.3	12,860
2023	880,418	12.5	14,198
2024	859,532	16.6	19,313
Average	847,067.5	11.58	13,628.5

Note. (i) Complaint and loss data are from the FBI Internet Crime Complaint Center's annual reports covering 2021 through 2024 (Federal Bureau of Investigation, 2022, 2023, 2024, 2025). (ii) Values describe all reported internet crime, not deepfake-attributable loss. Loss-per-complaint ratios use rounded loss totals and therefore differ slightly from calculations using unrounded agency data.

Deepfake Enabled Financial Fraud and Consumer Trust in Digital Finance: Can Vulnerability–Trust–Resilience Framework Be Applied in the United States of America?

In Table 1, between 2021 and 2024, reported losses rose by approximately 140.6%, equivalent to a compound annual growth rate of about 33.9%. Complaint volume moved much less, from 847,376 to 859,532, so reported loss per complaint more than doubled. The change may reflect greater loss severity, a different composition of reported offenses, or both. It does not demonstrate that generative AI caused the increase.

The 2024 increase in Table 1 is especially material: losses rose 32.8% from the rounded 2023 total while complaints declined 2.4%. Institutions consequently face a loss environment in which preventing a comparatively small number of high-value social-engineering, investment, business-email-compromise, or account-takeover events may have substantial value. Deepfake capability can increase persuasion in these categories, but the public series cannot identify the share.

4.2 Prevalence of Response Mechanisms

Table 2: Mechanism Prevalence Across the 18-Document Corpus

Mechanism	Documents	Prevalence	Interpretation
Identity verification (IV)	16	88.9%	Identity evidence and authentication remain the principal control boundary.
Incident response/information sharing (IR)	17	94.4%	Reporting and adaptive response are treated as essential under changing attacks.
Human authentication/escalation (HA)	15	83.3%	Independent challenge and authority to pause complement technical tools.
Transaction monitoring (TM)	13	72.2%	Contextual payment and behavioral signals remain valuable when media is convincing.
Accountable ownership (OW)	12	66.7%	Responsibility is recognized but not consistently operationalized in threat alerts.
Third-party/platform governance (TP)	12	66.7%	Fraud defenses depend on vendors, platforms, telecommunications, and shared infrastructure.
Deepfake detection/provenance (DD)	9	50.0%	Technical authenticity mechanisms are emerging but uneven and nonconclusive.
Customer information/trust (CI)	7	38.9%	Consumer communication and recovery receive less attention than prevention and reporting.

Source: Author's content analysis of the 18-document corpus. Prevalence represents the percentage of documents in which each mechanism was identified (n = 18).

The dominant pattern in Table 2 was layered rather than detector-centered. Identity verification appeared in 16 documents, incident response or information sharing in 17, human challenge in 15, and transaction monitoring in 13. Only nine offered a substantive discussion of deepfake detection, watermarking, provenance, or media forensics. The distribution suggests that authoritative institutions regard authenticity classifiers as useful components, but not as sufficient defenses.

Customer information in Table 2 was least prevalent. Some technical and prudential documents reasonably fall outside consumer communication, but the gap matters at system level. A fraud-control architecture can reduce direct losses while weakening channel trust if customers do not understand verification requests, cannot reach an accountable human, or experience unexplained restrictions. Only six documents jointly contained CI, IV, and IR, the bundle most directly related to prevention, communication, and recovery.

Deepfake Enabled Financial Fraud and Consumer Trust in Digital Finance: Can Vulnerability–Trust–Resilience Framework Be Applied in the United States of America?

4.3 Control Bundles and Negative Cases

Identity verification and incident response co-occurred in fifteen documents. This pairing reflects a lifecycle logic: institutions must both establish identity and learn when proofing fails. Identity verification and transaction monitoring co-occurred in twelve. The latter bundle is particularly important because identity certainty is rarely absolute. A customer who passes login controls may be coerced; a criminal may control a legitimate account; or an employee may act on a fabricated executive request. Transaction context supplies an independent signal.

Human authentication co-occurred with transaction monitoring in ten documents. The operational implication is not generic “human in the loop” language. Effective challenge requires a defined trigger, an independent communication path, access to relevant context, authority to delay action, and protection from retaliation for pausing a senior request. Without these conditions, formal escalation may be ceremonial.

Deepfake detection and provenance co-occurred with identity verification in all nine DD-coded documents. This shows that content analysis is usually embedded in broader proofing rather than presented as a standalone gate. The negative case is equally informative: nine documents made no substantive DD recommendation. Their controls—reporting, behavioral monitoring, customer warnings, or governance—can still interrupt deepfake-enabled fraud without classifying the media itself.

Third-party governance appeared in twelve documents, especially recent AI and cyber-risk frameworks. The dependency extends beyond a bank’s model vendor. Identity proofing may rely on device manufacturers, data brokers, cloud services, biometric vendors, telecommunications carriers, and social platforms. A customer can be deceived on one platform and transfer funds through another. Contractual service levels, evidence access, incident notification, testing rights, and exit arrangements are therefore resilience controls.

4.4 The Attribution-Denominator Problem

No harmonized public denominator for deepfake-enabled financial fraud emerged from the review. Three classification questions remain unsettled. Was synthetic media technically confirmed, considered probable, merely suspected, or only alleged? Did it enable the fraud or play an incidental persuasive role? And where should loss be attributed when a multi-stage scheme combines generated content, stolen credentials, a mule account, and a victim-authorized transfer?

This produces both numerator and denominator problems. The numerator is uncertain because deepfake involvement may not be detected or recorded. The denominator is uncertain because opportunities for exposure—calls, onboarding attempts, payment requests, or media artifacts—are not consistently counted. Reports of percentage growth from small or proprietary baselines cannot substitute for population estimates.

A minimum event taxonomy should record: institution and channel; suspected or confirmed synthetic modality; attack stage; target type; verification controls encountered; transaction type and attempted amount; prevented, pending, and realized loss; customer reimbursement; time to containment; false-positive status; and reporting destination. Synthetic involvement should use confidence levels rather than a forced yes/no label. Publication can occur in aggregated form to protect privacy and investigations.

4.5 Why Vulnerability–Trust–Resilience Framework?

The evidence supports a mechanism with four stages. Offenders first locate procedural gaps and dependence on weak signals. They then combine synthetic cues with urgency, authority, familiarity, or fear to exploit trust. At the point of intervention, layered controls either interrupt the scheme or allow it to proceed. Resilience begins after that decision: the institution contains harm, restores legitimate access, communicates with affected parties, shares intelligence, and updates its controls.

Four outcomes should be measured: financial loss severity; decision accuracy, including false positives and false negatives; trust calibration, including continued use and perceived competence; and adaptation speed, measured from novel incident detection to control deployment. Resilience is not the absence of all incidents. It is the capacity to lower likelihood, limit impact, recover fairly, and learn faster than attackers adapt.

5. RESULTS AND DISCUSSION

5.1 Interpretation of Results

The findings challenge the narrow view that deepfake fraud is primarily a media-detection problem. Across the corpus, identity, transaction context, human challenge, and incident response received more attention than synthetic-content classification. That

Deepfake Enabled Financial Fraud and Consumer Trust in Digital Finance: Can Vulnerability–Trust–Resilience Framework Be Applied in the United States of America?

emphasis is strategically coherent. Generators change quickly; independent confirmation, payment limits, behavioral anomalies, dual authorization, and rapid payment recall can remain effective across multiple generations of attack technology.

Trust also appears to be under-governed. Fewer than half of the documents gave substantive attention to customer information or trust-related recovery, even though digital finance requires consumers to tolerate remote interaction and occasional security friction. Unexplained account blocks may prevent loss while weakening confidence. Reassuring communication unsupported by effective controls may encourage unsafe reliance. Calibrated trust requires visible competence, understandable procedures, and accessible remedy.

The attribution-denominator problem changes the research agenda. It is tempting to open an article with a large aggregate loss figure and imply that deepfakes explain it. Such framing may attract attention but weakens scientific credibility. The stronger contribution is to show precisely what is and is not measured, then propose a taxonomy that makes future causal work possible.

5.2 Comparison with Prior Literature

Routine activity theory predicts that reducing opportunity can be as important as identifying offender technology. The observed prevalence of IV, TM, HA, and IR supports this mechanism-level interpretation. Capable guardianship is distributed across customer behavior, institutional systems, employee authority, third parties, and public reporting.

Protection motivation theory helps explain why CI cannot be reduced to generic warnings. Threat communication should identify observable triggers and feasible coping actions. Messages should instruct customers to initiate verification through a known channel, establish family or organizational challenge phrases where appropriate, resist urgency, and report quickly. Excessive emphasis on technical sophistication can lower self-efficacy.

Trust-in-automation theory aligns with the need for calibration. An authenticity score should inform, not silently determine, consequential action. Reviewers need uncertainty information, relevant transaction context, and escalation options. Customers need signals that are difficult for attackers to reproduce, not cosmetic trust marks.

The findings extend resilience research by placing remediation and measurement inside the framework. Fraud defenses are often assessed through prevented loss, but realized customer outcomes also include access restoration, reimbursement, distress, time cost, and future channel use. Institutions that learn only from confirmed losses ignore near misses and harmful false positives.

5.3 Theoretical Implications

The VTR framework makes three theoretical claims. First, synthetic media is a capability multiplier rather than a self-contained offense category. Its impact depends on organizational vulnerabilities and transaction opportunities. Second, trust is both an attack surface and a performance outcome. Offenders exploit interpersonal and institutional trust; controls then reshape future trust in the institution and channel. Third, resilience depends on control diversity. Independent signals reduce common-mode failure, while recovery and learning determine whether an incident becomes persistent harm.

These claims generate measurable propositions. Future research can model loss severity as a function of control bundles, use mediation analysis to test response quality and procedural fairness, and examine the nonlinear effect of friction on adoption. Multilevel models could separate customer, institution, vendor, and jurisdiction effects. Event studies could assess changes following publicized deepfake incidents, although identification would require careful control for simultaneous fraud trends.

5.4 Managerial and Policy Implications

Financial institutions should begin with process mapping. High-risk points include remote onboarding, credential reset, beneficiary creation, high-value payment authorization, executive requests, customer-service account changes, and claims or loan documentation. Each process should identify which signals an attacker can synthesize and which independent signals remain available. Controls should be risk-tiered. Low-risk inquiries need minimal friction. High-value or anomalous instructions should trigger step-up verification through a separately established channel, cooling-off periods, dual approval, or direct confirmation. Voice or facial familiarity alone should never authorize an exceptional transaction. Employees need explicit authority to stop a request from a senior person without career penalty.

Institutions should evaluate detection tools using deployment-relevant base rates and demographic performance. Procurement should require disclosure of evaluation data, update practices, known limitations, incident notification, evidence retention, and audit rights. A vendor's headline accuracy is not enough. Testing should include compression, replay, injection, device variation, accessibility scenarios, and novel generator families.

Deepfake Enabled Financial Fraud and Consumer Trust in Digital Finance: Can Vulnerability–Trust–Resilience Framework Be Applied in the United States of America?

Customer communication should be designed as a control. Messages should be brief, specific, and actionable; offer independent contact paths; avoid blaming victims; and explain unusual verification. Post-incident handling should coordinate fraud, cybersecurity, complaints, legal, communications, and vulnerable-customer teams. Measures should include time to human contact, account restoration, provisional credit where applicable, repeat victimization, complaint resolution, and trust recovery.

Policy makers should standardize reporting without requiring public disclosure of exploitable details. A shared taxonomy could be embedded in suspicious-activity reporting, consumer complaint systems, and sector information-sharing organizations. Aggregate publication should distinguish suspected from confirmed use and attempted from realized loss. Regulators should also monitor whether anti-deepfake controls create disproportionate exclusion.

Table 3: VTR Managerial Scorecard

Domain	Example leading indicator	Example outcome indicator	Governance question
Vulnerability mapping	High-risk processes threat-modeled (%)	Control gaps overdue	Where can synthetic signals authorize value?
Verification	Step-up completion and failure rates	Confirmed impersonation escape rate	Are independent signals truly independent?
Transactions	Anomalous payments reviewed within target	Prevented and realized loss	Does context override apparent familiarity?
Human escalation	Staff completing simulations	Losses after ignored escalation	Can employees safely pause authority?
Customer trust	Warnings tested for comprehension	Trust, attrition, complaint outcomes	Do controls increase calibrated confidence?
Recovery	Recall and response time	Net loss and restoration time	How quickly is harm contained and repaired?
Learning	Novel cases coded within taxonomy	Time from incident to control update	Is the institution adapting measurably?
Third parties	Critical vendors tested	Vendor-linked incident severity	Can dependencies be observed and exited?

Source: Author’s proposed VTR managerial scorecard, developed through synthesis of findings from the 18-document corpus.

5.5 Limitations

The corpus is purposive, English-language, and weighted toward public institutions. Frequency reflects emphasis, not quality or implementation. Binary coding reduces nuance. Single-author recoding does not replace independent reliability assessment. Public fraud data are reported-event statistics affected by underreporting and classification change. Monetary figures are nominal. No causal link between generative AI and aggregate loss trends is estimated. Consumer trust is theorized from documentary and established trust literature rather than measured in a new sample.

These limitations define the next empirical step. A preregistered multi-country study could combine a consumer experiment, bank-level incident taxonomy, and longitudinal control-maturity survey. Ethical safeguards would be essential because realistic impersonation stimuli can distress participants or reveal exploitable procedures. Research partnerships should publish de-identified measures and preserve negative results.

Deepfake Enabled Financial Fraud and Consumer Trust in Digital Finance: Can Vulnerability–Trust–Resilience Framework Be Applied in the United States of America?

6. CONCLUSION AND RECOMMENDATIONS

6.1 Summary of Key Findings

Deepfake-enabled financial fraud is an adaptive socio-technical threat to identity, transaction integrity, and justified trust. Among the 18 authoritative documents analyzed, identity verification and incident response were most prevalent, followed by human escalation and transaction monitoring. Content detection and provenance appeared less consistently, reinforcing the case for defense in depth rather than reliance on a universal detector. Customer communication and trust recovery were the least integrated elements. U.S. reported fraud losses increased sharply from 2021 to 2024, but existing data cannot identify the portion caused or enabled by synthetic media.

The VTR framework traces how synthetic content exploits procedural weakness and familiar trust cues, how independent controls interrupt loss, and how fair recovery and institutional learning produce resilience. Its practical implication is straightforward: authenticate the consequential action, rather than merely asking whether a voice, face, or document appears genuine.

6.2 Recommendations for Research and Practice

Financial institutions should map synthetic-media exposure across high-value workflows, combine independent identity and transaction signals, institutionalize protected human escalation, test vendors under realistic conditions, and measure customer recovery alongside prevented loss. Authorities should adopt a common event taxonomy with confidence levels for suspected and confirmed synthetic involvement. Researchers should avoid attributing aggregate fraud losses to deepfakes without case-level evidence and should prioritize longitudinal, experimental, and quasi-experimental designs.

Future work should test the five propositions using matched institutional data and diverse consumer samples. A useful minimum design would measure exposure, control maturity, false-positive burden, attempted and realized loss, response quality, trust before and after incidents, inclusion outcomes, and adaptation time. Such evidence would allow the field to move from urgent but imprecise warnings toward accountable estimates of which controls protect both money and confidence.

REFERENCES

- 1) Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: A routine activity approach. *American Sociological Review*, 44(4), 588–608. <https://doi.org/10.2307/2094589>
- 2) Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review*, 107(6), 1753–1820. <https://doi.org/10.15779/Z38RV0D15J>
- 3) Cybersecurity and Infrastructure Security Agency, Federal Bureau of Investigation, & National Security Agency. (2023). *Contextualizing deepfake threats to organizations*. <https://www.cisa.gov/resources-tools/resources/contextualizing-deepfake-threats-organizations>
- 4) de Rancourt-Raymond, A., & Smaili, N. (2023). The unethical use of deepfakes. *Journal of Financial Crime*, 30(4), 1066–1077. <https://doi.org/10.1108/JFC-04-2022-0090>
- 5) European Union Agency for Cybersecurity. (2022). *ENISA threat landscape 2022*. <https://www.enisa.europa.eu/publications/enisa-threat-landscape-2022>
- 6) Federal Bureau of Investigation. (2022). *2021 Internet Crime Report*. Internet Crime Complaint Center. https://www.ic3.gov/AnnualReport/Reports/2021_IC3Report.pdf
- 7) Federal Bureau of Investigation. (2023). *2022 Internet Crime Report*. Internet Crime Complaint Center. https://www.ic3.gov/AnnualReport/Reports/2022_IC3Report.pdf
- 8) Federal Bureau of Investigation. (2024). *2023 Internet Crime Report*. Internet Crime Complaint Center. https://www.ic3.gov/AnnualReport/Reports/2023_IC3Report.pdf
- 9) Federal Bureau of Investigation. (2025). *2024 IC3 annual report*. Internet Crime Complaint Center. https://www.ic3.gov/AnnualReport/Reports/2024_IC3Report.pdf
- 10) Federal Reserve Board. (2025, April 17). *Deepfakes and the AI arms race in bank cybersecurity*. <https://www.federalreserve.gov/newsevents/speech/files/barr20250417a.pdf>
- 11) Financial Action Task Force. (2023). *Illicit financial flows from cyber-enabled fraud*. <https://www.fatf-gafi.org/en/publications/Methodsandtrends/illicit-financial-flows-cyber-enabled-fraud.html>

Deepfake Enabled Financial Fraud and Consumer Trust in Digital Finance: Can Vulnerability–Trust–Resilience Framework Be Applied in the United States of America?

- 12) Financial Crimes Enforcement Network. (2024a, November 13). *FinCEN alert on fraud schemes involving deepfake media targeting financial institutions* (FIN-2024-Alert004). <https://www.fincen.gov/system/files/shared/FinCEN-Alert-DeepFakes-Alert508FINAL.pdf>
- 13) Financial Crimes Enforcement Network. (2024b, January 9). *Identity-related suspicious activity: 2021 threats and trends*. https://www.fincen.gov/sites/default/files/shared/FTA_Identity_Final508.pdf
- 14) Financial Stability Board. (2024). *The financial stability implications of artificial intelligence*. <https://www.fsb.org/2024/11/the-financial-stability-implications-of-artificial-intelligence/>
- 15) Federal Trade Commission. (2025). *Consumer Sentinel Network data book 2024*. <https://www.ftc.gov/reports/consumer-sentinel-network-data-book-2024>
- 16) International Organization for Standardization. (2023). *ISO/IEC 42001:2023 information technology—Artificial intelligence—Management system*. <https://www.iso.org/standard/81230.html>
- 17) Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392
- 18) McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and validating trust measures for e-commerce: An integrative typology. *Information Systems Research*, 13(3), 334–359. <https://doi.org/10.1287/isre.13.3.334.81>
- 19) Muzatko, S., & Bansal, G. (2024). It pays to be forthcoming: Timing of data breach announcement, trust violation, and trust restoration. *Internet Research*, 34(5), 1629–1663. <https://doi.org/10.1108/INTR-12-2021-0939>
- 20) 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>
- 21) National Institute of Standards and Technology. (2024a). *Artificial intelligence risk management framework: Generative artificial intelligence profile* (NIST AI 600-1). <https://doi.org/10.6028/NIST.AI.600-1>
- 22) National Institute of Standards and Technology. (2024b). *Reducing risks posed by synthetic content: An overview of technical approaches to digital content transparency* (NIST AI 100-4). <https://doi.org/10.6028/NIST.AI.100-4>
- 23) Rogers, R. W. (1975). A protection motivation theory of fear appeals and attitude change. *The Journal of Psychology*, 91(1), 93–114. <https://doi.org/10.1080/00223980.1975.9915803>
- 24) U.S. Department of the Treasury. (2024a). *Managing artificial intelligence-specific cybersecurity risks in the financial services sector*. <https://home.treasury.gov/system/files/136/Managing-Artificial-Intelligence-Specific-Cybersecurity-Risks-In-The-Financial-Services-Sector.pdf>
- 25) U.S. Department of the Treasury. (2024b). *Artificial intelligence in financial services*. <https://home.treasury.gov/system/files/136/Artificial-Intelligence-in-Financial-Services.pdf>

Appendix A: Document-Level Coding Matrix

ID	Issuer and document (year)	OW	IV	DD	TM	HA	CI	IR	TP
D0 1	CISA/FBI/NSA, <i>Contextualizing Deepfake Threats</i> (2023)	1	1	1	0	1	1	1	1
D0 2	FinCEN, <i>Deepfake Fraud Alert</i> (2024)	0	1	1	1	1	0	1	0
D0 3	FBI, <i>2021 Internet Crime Report</i> (2022)	0	1	0	1	1	1	1	0
D0 4	FBI, <i>2022 Internet Crime Report</i> (2023)	0	1	0	1	1	1	1	0
D0 5	FBI, <i>2023 Internet Crime Report</i> (2024)	0	1	0	1	1	1	1	0

Deepfake Enabled Financial Fraud and Consumer Trust in Digital Finance: Can Vulnerability–Trust–Resilience Framework Be Applied in the United States of America?

ID	Issuer and document (year)	OW	IV	DD	TM	HA	CI	IR	TP
D0 6	FBI, <i>2024 IC3 Annual Report</i> (2025)	0	1	0	1	1	1	1	0
D0 7	NIST, <i>AI RMF 1.0</i> (2023)	1	0	0	0	1	0	1	1
D0 8	NIST, <i>Generative AI Profile</i> (2024)	1	1	1	0	1	0	1	1
D0 9	NIST, <i>Synthetic Content Transparency</i> (2024)	1	1	1	0	1	1	1	1
D1 0	U.S. Treasury, <i>AI-Specific Cybersecurity Risks</i> (2024)	1	1	1	1	1	0	1	1
D1 1	U.S. Treasury, <i>AI in Financial Services</i> (2024)	1	1	1	1	1	0	1	1
D1 2	FSB, <i>Financial Stability Implications of AI</i> (2024)	1	1	0	1	0	0	1	1
D1 3	FATF, <i>Illicit Financial Flows From Cyber-Enabled Fraud</i> (2023)	1	1	0	1	1	0	1	1
D1 4	FTC, <i>Consumer Sentinel Data Book 2024</i> (2025)	0	0	0	1	0	1	1	0
D1 5	ENISA, <i>Threat Landscape 2022</i> (2022)	1	1	1	1	1	0	1	1
D1 6	ISO/IEC, <i>AI Management System</i> (2023)	1	1	0	0	1	0	1	1
D1 7	Federal Reserve, <i>Deepfakes and Bank Cybersecurity</i> (2025)	1	1	1	1	1	0	1	1
D1 8	FinCEN, <i>Identity-Related Suspicious Activity</i> (2024)	1	1	1	1	0	0	0	1

Codes: OW = accountable ownership; IV = identity verification; DD = deepfake detection/provenance; TM = transaction monitoring; HA = human authentication/escalation; CI = customer information/trust; IR = incident response/information sharing; TP = third-party/platform governance. Coding reflects substantive document content, not assessed effectiveness.

Appendix B: Proposed Deepfake-Enabled Fraud Event Schema

1. **Event identification:** anonymous event identifier, reporting institution type, date detected, and jurisdiction.
2. **Synthetic-media status:** not assessed, suspected, probable, or technically confirmed; audio, image, video, text, or multimodal.
3. **Attack stage:** reconnaissance, onboarding, authentication, social engineering, transaction authorization, account recovery, or concealment.
4. **Target and channel:** customer, employee, executive, vendor, or public; branch, telephone, video, mobile, web, email, messaging, or platform.

Deepfake Enabled Financial Fraud and Consumer Trust in Digital Finance: Can Vulnerability–Trust–Resilience Framework Be Applied in the United States of America?

5. **Control encounter:** identity proofing, liveness, device, behavior, transaction monitoring, call-back, dual authorization, detector, or manual review.
6. **Outcome:** prevented, pending, or realized loss; attempted and realized amount; funds recovered; customer reimbursement.
7. **Operational impact:** time to detection, containment, human contact, account restoration, and control update.
8. **Trust and inclusion:** complaint, attrition, satisfaction, accessibility issue, false-positive determination, and vulnerable-customer indicator.
9. **External response:** suspicious activity report, law-enforcement referral, information-sharing alert, vendor notification, or platform takedown.

Appendix C: Illustrative Survey and Experimental Measures for Future Validation

Future researchers may validate the VTR framework through a preregistered factorial experiment. Participants would view a legitimate or synthetically manipulated payment request, combined with one of three bank responses: no step-up verification, unexplained step-up verification, or explained independent verification. Primary outcomes would include compliance intention, detection confidence, institutional trust, perceived competence, procedural fairness, continued-use intention, and perceived burden. Researchers should measure digital literacy, prior fraud exposure, age, disability/accessibility requirements, and financial vulnerability as moderators. Items may use seven-point agreement scales: “I believe this institution is capable of protecting my account”; “The verification process is proportionate to the risk”; “I understand why additional verification was requested”; “I would continue using this digital channel”; and “I know how to report a suspicious request.” Behavioral outcomes should be prioritized over self-report where ethical and feasible. Manipulation checks must distinguish perceived realism from actual authenticity. Debriefing should occur immediately, and no real credentials or funds should be involved.



There is an Open Access article, distributed under the term of the Creative Commons Attribution – Non Commercial 4.0 International (CC BY-NC 4.0) (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits remixing, adapting and building upon the work for non-commercial use, provided the original work is properly cited.