International Journal of Leading Research Publication (IJLRP)

E-ISSN: 2582-8010 ● Website: www.ijlrp.com ● Email: editor@ijlrp.com

The Convergent Cyber Threat Horizon in Financial Services:
Artificial Intelligence, Agentic Systems, Deepfakes, Quantum Computing, and the Mythos of Large Language Models

Stephen Coraggio

Independent Researcher, Cybersecurity & Financial Services Risk

Abstract:
Financial services institutions are entering a period of simultaneous and interacting technology shocks whose combined impact on the cyber threat landscape is qualitatively greater than that of any single force. This paper examines five such forces — generative artificial intelligence (AI), agentic AI, synthetic media (deepfakes), cryptographically-relevant quantum computing, and the institutional mythos surrounding large language models (LLMs) — and analyses their combined implications for banks, insurers, asset managers, payment providers, and market infrastructure operators. Drawing on primary standards documents, sector incident reports, and the emerging academic literature, we argue that the traditional cybersecurity program, built on assumptions of scarce adversary labour, durable cryptography, trustworthy media, and human-only identity, is structurally misaligned with the 2026–2032 threat environment. We contribute a five-vector threat taxonomy for financial services, a unified seven-domain defence framework mapping to NIST Cybersecurity Framework 2.0 and the EU Digital Operational Resilience Act (DORA), and a prioritised 24-month preparedness roadmap. We highlight under-researched risks including non-human identity proliferation, indirect prompt injection against agentic pipelines, harvest-now-decrypt-later (HNDL) exposure of long-retention financial data, and the governance gap introduced by uneven institutional understanding of LLM capabilities. The paper closes with research and policy recommendations including the establishment of sector-wide cryptographic inventories, standardised AI red-teaming protocols, and content-provenance requirements for material financial communications.

Keywords: cybersecurity; financial services; generative AI; agentic AI; deepfakes; post-quantum cryptography; large language models; prompt injection; non-human identity; operational resilience; DORA; NIST AI RMF

1. INTRODUCTION
Financial services cybersecurity has, for more than two decades, rested on a relatively stable set of operational assumptions: adversary capability scales with scarce human expertise [1]; public-key cryptography is effectively immortal within any planning horizon that matters to risk committees [2]; identity is primarily a human construct verifiable through the canonical three factors [3]; and media entering the enterprise — emails, voice calls, documents, video conferences — is broadly authentic unless evidence suggests otherwise. Each of these assumptions has either failed or is on a trajectory to fail within the planning horizon of existing technology assets [4], [5].

Five interacting forces drive this dislocation. First, commoditised generative AI has collapsed the cost of producing persuasive malicious content and exploit code [6], [7]. Second, agentic AI systems — language models wrapped in planning loops with tool access — convert models that produce text into systems that produce actions, opening a new adversarial tier and, equally, a new class of privileged insider [8], [9]. Third, synthetic media generation has reached fidelity sufficient to defeat voice and video based authentication, with documented losses already in the tens of millions of U.S. dollars per incident [10], [11]. Fourth, the maturation of cryptographically-relevant quantum computing, while still projected to arrive no earlier than the early 2030s, has activated harvest-now-decrypt-later (HNDL) threats today [12], [13]. Fifth, and cross-cutting, the institutional mythos around LLMs — the uneven, under-examined folk beliefs by which organisations understand AI capability — is itself a governance risk surface that interacts multiplicatively with the prior four [14].

This paper's contribution is fourfold. (i) We synthesise the threat literature across these five forces into a unified five-vector taxonomy specific to financial services. (ii) We identify a set of sector-specific amplifiers — identity-as-value, interconnection, real-time irrevocability, and regulatory disclosure — that make the financial sector a leading target. (iii) We propose a seven-domain defence framework, mapping each domain to the NIST Cybersecurity Framework 2.0 [15] and the EU Digital Operational Resilience Act [16]. (iv) We offer a prioritised 24-month preparedness roadmap, including concrete outcome signals suitable for board-level oversight. Throughout, we highlight open research questions that merit sustained academic attention.

The remainder of the paper is organised as follows. Section 2 reviews related work and positions our contribution. Section 3 develops the threat taxonomy. Sections 4 through 8 examine each of the five convergent forces in turn. Section 9 analyses the differential impact on financial services subsectors. Section 10 surveys the regulatory landscape. Section 11 proposes the unified defence framework. Section 12 offers a 24-month preparedness roadmap. Section 13 identifies limitations and research directions. Section 14 concludes.

2. RELATED WORK AND BACKGROUND

2.1 Generative AI and Offensive Security

The capability jump associated with transformer-based large language models [17], [18] has been documented extensively in benchmarks of general reasoning [19] and, more recently, in offensive cyber capability evaluations [20], [21]. Brundage et al. [22] anticipated the malicious use of AI in a widely-cited foresight study; subsequent empirical work by Gupta et al. [7] and by the Microsoft Threat Intelligence Center [23] has confirmed operational use by both criminal and state-affiliated actors. Narayanan and Kapoor [24] caution against overstatement of AI capability while acknowledging the genuine implications for security.

2.2 Agentic Systems and Prompt Injection

The transition from chatbot to agent was formalised by Shinn et al. [25] and Yao et al. [26]; its security implications have been explored principally through the lens of prompt injection. Perez and Ribeiro [27] demonstrated the fragility of instruction-following systems to adversarial prompts, while Greshake et al. [28] identified indirect prompt injection via retrieved content as a near-universal weakness in retrieval-augmented generation pipelines. The OWASP Top 10 for Large Language Model Applications (2025 edition) [29] consolidates these findings into a practitioner-oriented taxonomy that we adopt and extend.

2.3 Synthetic Media

Deepfake generation via generative adversarial networks [30] and diffusion models [31] has progressed rapidly. Chesney and Citron's early legal-policy analysis [32] and Westerlund's review [33] framed the

societal risk; recent surveys by Mirsky and Lee [34] and by Masood et al. [35] document the detection-generation arms race. The operational reality for financial services is increasingly reflected in incident disclosures, most prominently the 2024 Hong Kong video-conference deepfake wire fraud [11].

2.4 Post-Quantum Cryptography

Shor's polynomial-time quantum algorithm for integer factorisation and discrete logarithms [36] is the foundational threat to existing public-key cryptography. Mosca's formulation of the migration timing problem [12] remains the canonical risk-management frame: organisations must migrate before the minimum of their data confidentiality horizon plus migration time exceeds the time to a cryptographically-relevant quantum computer. NIST's 2024 standardisation of ML-KEM, ML-DSA, and SLH-DSA [37], [38], [39] provides the practical target for migration; Bernstein and Lange's review [40] describes the candidate landscape. U.S. government direction via NSM-10 [41] and OMB M-23-02 [42] has established migration expectations for federal systems that, in practice, cascade to regulated financial institutions.

2.5 Financial Services Cyber Risk

The Financial Stability Board's 2024 report on AI in financial services [43] and the Basel Committee's principles for operational risk [44] frame the sector-specific policy context. Empirical cost and incident trends are reported annually by IBM [4], Verizon [45], and the FBI Internet Crime Complaint Center [46]. FS-ISAC threat intelligence products [47] remain an under-cited but operationally central source. Our contribution relative to this literature is to unify the generative-AI, agentic, synthetic-media, and quantum threads into a single analytical frame tailored to financial services decision-makers.

2.6 AI Governance and Risk Management

The NIST AI Risk Management Framework [48] and ISO/IEC 42001 [49] establish the governance scaffolding that we operationalise. The EU AI Act [50] imposes the most demanding ex-ante regulatory regime on high-risk AI uses, which include a substantial share of financial services applications. MITRE's ATLAS knowledge base [51] extends ATT&CK-style taxonomy to adversarial AI. Weidinger et al. [52] and Bommasani et al. [53] provide comprehensive treatments of the ethical and systemic risks of foundation models.

3. A FIVE-VECTOR THREAT TAXONOMY FOR FINANCIAL SERVICES

We propose five threat vectors, enumerated in Table 1, to organise the analysis. The vectors are neither exhaustive nor mutually exclusive; in practice, material attacks traverse multiple vectors, and the most consequential scenarios combine them. The taxonomy is intended as an analytic scaffold for risk registers, red-team scoping, and regulatory dialogue rather than a classification ontology.

Vector	Core Threat	Representative Manifestation in Financial Services
V1: Generative AI	Commoditised production of malicious content and code	AI-authored spear-phishing, malware polymorphism, synthetic KYC documents
V2: Agentic AI	Autonomous adversaries; privileged non-human insiders	Multi-step attack agents; prompt-injection of treasury-facing AI agents
V3: Synthetic Media	Collapse of voice/video as authentication primitive	Executive-impersonation wire fraud; deepfake KYC bypass; market-moving fakes

Vector	Core Threat	Representative Manifestation in Financial Services
V4: Quantum/HNDL	Retroactive decryption of long-horizon data	Harvest-now-decrypt-later against M&A archives, KYC, customer PII
V5: LLM Mythos	Governance failure due to uneven institutional belief	Shadow AI; over-reliance on AI output; ungoverned RAG data flows

Table 1. Five-vector threat taxonomy for financial services cybersecurity.

Within each vector, we further distinguish three capability tiers: Tier 1 is AI-unaugmented (traditional tooling); Tier 2 is AI-augmented (human operators using AI tools); Tier 3 is AI-autonomous (multi-agent systems executing kill chains at machine speed). As of early 2026, the majority of empirically observed incidents in financial services reflect Tier 2 activity [23], [47], with Tier 3 emerging but not yet dominant. The migration from Tier 2 to Tier 3 is the central near-term risk trajectory.

4. VECTOR V1: GENERATIVE AI AS AN OFFENSIVE CAPABILITY

4.1 Economic Inversion of Attack Cost

The single most important fact about generative AI, from a defender's perspective, is economic rather than technical. Prior to 2023, most attack campaigns against financial institutions faced a binding constraint on skilled human labour [22]. Producing convincing multilingual spear-phishing, reading thousands of pages of leaked corporate documents for exploitable relationships, or writing malware variants capable of evading signature detection required operator time priced at the market rate for the relevant skills. Commodity LLMs have largely removed this constraint [6], [7]. What remains is a throughput limit set by API cost and rate limits, both of which continue to fall by roughly an order of magnitude every 18 to 24 months [19].

Operationally, this manifests as a one-to-two-order-of-magnitude increase in the volume of high-quality — not low-quality, which was always cheap — attacks, and a thickening of the tail of bespoke attacks. Programs calibrated on pre-2023 base rates are systematically under-defended.

4.2 Observed AI-Enabled Attack Classes

We organise observed attack classes into six operationally meaningful categories. Hyper-personalised social engineering uses target-specific context drawn from public sources to produce phishing content with click rates reported at two to five times those of legacy spear-phishing [23]. Business email compromise, version two, augments traditional BEC with real-time conversational response and synthesised supporting documentation; FBI IC3 data put global BEC exposure above USD 55 billion in the 2013–2023 window and the trajectory has continued [46]. Malware polymorphism leverages LLM code-generation to produce behaviour-equivalent but syntactically distinct variants that degrade signature-based detection [54]. AI-assisted vulnerability discovery compresses the time from CVE disclosure to exploit availability [55]. Credential-harvesting infrastructure has industrialised via AI-generated domains and brand-faithful phishing pages [56]. AI-orchestrated reconnaissance synthesises technology stack, personnel, and third-party exposure from public sources in minutes rather than weeks [7].

4.3 Defensive Use and the Asymmetric Race

The same technology is being deployed defensively — in alert triage, fraud detection, incident response, and secure software development [47], [57]. Defensive use is essential but constrained: defenders must protect a broad surface under compliance, audit, and reliability requirements, while attackers need only find a single path. Dual-use amplification does not cancel out; it preserves and, on some analyses, widens the attacker's structural advantage [22], [58].

5. VECTOR V2: AGENTIC AI — AUTONOMOUS ADVERSARIES AND NON-HUMAN INSIDERS

5.1 From Chatbots to Agents

An agentic AI system is a language model wrapped in a planning-and-tool-use loop [25], [26]: given an objective, the system decomposes it into steps, invokes tools (browsers, APIs, shells, email, payment systems), observes outcomes, and iterates. The transition from chatbot to agent is structural for cybersecurity because it converts a text-producing system into an action-producing one. The same transition applies on the defensive side: in-enterprise agentic systems are privileged actors that require governance commensurate with their capability [59].

5.2 Adversary Capabilities

Capability thresholds across the agentic attacker kill chain differ materially as of early 2026. Reconnaissance agents are fully operational: an agent can enumerate a target's infrastructure and produce an attack plan in tens of minutes [21]. Exploitation agents are partially operational: chaining of known exploits and adaptation of payloads in test environments is demonstrated, though production effectiveness against hardened targets remains imperfect [20]. Social-engineering agents are increasingly operational, particularly for voice-based help-desk attacks combining deepfake audio with conversational agents [11]. Persistence and evasion agents that recognise defensive telemetry and adjust behaviour to avoid detection have been demonstrated in research settings [60].

5.3 The Non-Human Identity Problem

An enterprise-deployed agent is, from an identity-governance perspective, a new class of principal. It is not human; it is not a traditional service account; it has reasoning capability, tool access, persistent memory, and — critically — it can be influenced by its inputs. We identify five categories of risk specific to this class [9], [28], [61].

5.3.1 Direct and Indirect Prompt Injection

An agent that ingests external content is exposed to adversarial instructions embedded in that content [28]. Indirect prompt injection is the OWASP LLM01 top risk for 2025 [29] and has no fully general defence; mitigation relies on layered input classification, output validation, strict tool allow-lists, and human-in-the-loop gates for high-consequence actions.

5.3.2 Tool Confusion and Lateral Capability Creep

Agents tend to accrete tools over time through successive incremental change requests. Each increment is individually justified; the aggregate produces highly privileged actors governed by natural-language prompts rather than deterministic code. A formal tool-scope review process analogous to privileged access certification is warranted.

5.3.3 Memory and State Poisoning

Agents with persistent memory stores (vector databases, retrieval indexes, conversational state) are exposed to poisoning attacks that alter output distributions without producing crisp failure signals [62]. Detection is harder than for deterministic code injection because the effect is statistical.

5.3.4 Model Supply Chain

Upstream model providers, fine-tuning pipelines, and hosted inference infrastructure each constitute supply-chain risk vectors. NIST SP 800-218A [63] and ISO/IEC 42001 [49] address the discipline, but adoption across the third-party landscape is uneven.

5.3.5 Accountability Gaps

When an agent takes a harmful action, legal and regulatory accountability remains unsettled. Boards and chief risk officers should assume that a duty-of-reasonable-care standard equivalent to that applied to human employees will attach retrospectively once supervisory expectations crystallise [43].

6. VECTOR V3: SYNTHETIC MEDIA AND THE COLLAPSE OF SENSORY TRUST

6.1 Technical State of the Art

Generative adversarial networks [30] and, more recently, diffusion models [31] have pushed synthetic audio, image, and video generation to fidelity levels at which human detection is unreliable under realistic conditions [34], [35]. Voice cloning from under 30 seconds of sample audio is commoditised; single-image full-motion portrait generation is widely available; real-time video-conferencing avatars that maintain lip-sync and emotional affect are demonstrated in multiple commercial products.

6.2 Operational Impact on Financial Services

Financial services has historically used recognition of voice, face, and recorded media as authentication and authorisation primitives — the callback procedure for wire authorisation, the recorded trade confirmation, the voice-biometric, and the video-based KYC liveness check are each, in effect, claims that an audiovisual signal is evidence of identity. Synthetic media has eroded each of these claims [32], [33]. Documented impact includes executive-impersonation wire fraud, most prominently the February 2024 Hong Kong incident in which a finance professional transferred approximately USD 25 million after attending a video conference whose participants were synthetic [11]. Synthetic-identity KYC enables account opening at scale under fabricated personas cultivated across social media and credit history [46]. Synthetic-content-driven market manipulation, illustrated by the May 2023 fabricated image of an explosion at the Pentagon that briefly moved the S&P 500 by approximately 0.3% [64], is a standing risk for a sector whose algorithmic participants react in milliseconds.

6.3 Detection and Its Limits

A substantial industry has formed around deepfake detection [34], [35], [65], with laboratory accuracies commonly reported above 95%. Three cautions apply. Detection is an adversarial game; any published detector becomes training signal for better generators. Laboratory accuracy does not translate to production against targeted attacks. Detection is necessary but insufficient: even a reliable detector does not reconstitute the lost authentication primitive.

6.4 Provenance as a Durable Answer

Cryptographic content-provenance standards, most prominently the C2PA specification and Content Credentials ecosystem [66], offer a more durable architectural response by binding signed provenance metadata to media from the point of capture onward. Adoption across the financial services sector is nascent; we argue that proactive deployment of C2PA in executive communications, regulatory disclosures, and customer-facing channels should be a standard posture by 2027.

7. VECTOR V4: QUANTUM COMPUTING AND THE HNDL HORIZON

7.1 Threat Mechanics

A cryptographically-relevant quantum computer (CRQC) of sufficient qubit count and error-correction depth would efficiently solve the integer-factorisation and discrete-logarithm problems via Shor's algorithm [36], breaking RSA, Diffie–Hellman, and elliptic-curve cryptography, and therefore compromising the confidentiality and authentication primitives underlying TLS, SSH, VPNs, PKI, code signing, and the majority of financial messaging authentication. Grover's algorithm [67] provides only a quadratic speedup against symmetric primitives, such that AES-256 and SHA-3 remain acceptable after a straightforward doubling of key length. The urgent threat surface is therefore the public-key layer.

7.2 Timeline Uncertainty

Credible expert estimates for CRQC availability cluster between 2030 and 2040 [12], [68]. The direction of travel is, however, not in doubt: qubit counts, gate fidelities, and error-correction architectures have progressed on schedule across leading programmes at IBM, Google, Quantinuum, IonQ, PsiQuantum, and

Chinese state laboratories [69]. The risk asymmetry for a regulated financial institution is unfavourable: early migration is merely expensive, late migration is, for a subset of the institution's data, existential.

7.3 Harvest-Now-Decrypt-Later

Adversaries need not await a CRQC to attack quantum-relevant data. Encrypted data exfiltrated today can be retained and decrypted once the capability matures [12], [41]. Western intelligence services have publicly assessed that China, Russia, and others operate HNDL programmes against financial, governmental, and defence targets. The financial sector is a priority target because much of its data — account numbers, government identifiers, KYC documentation, M&A deal terms, portfolio holdings — has a confidentiality half-life measured in decades.

7.4 Post-Quantum Migration

NIST's August 2024 finalisation of FIPS 203 (ML-KEM, formerly Kyber) [37], FIPS 204 (ML-DSA, formerly Dilithium) [38], and FIPS 205 (SLH-DSA, formerly SPHINCS+) [39] provides the foundation algorithms for migration. A fourth candidate, HQC, has been selected for standardisation as a backup KEM [70]. The practical migration programme comprises five workstreams: cryptographic inventory, crypto-agility architecture, hybrid classical-plus-PQC deployment through the transition period [71], data-classification-driven prioritisation, and third-party and regulatory coordination. Each is multi-year; initiation in 2026 is consistent with responsible planning horizons and consistent with U.S. federal direction [41], [42].

8. VECTOR V5: THE LLM MYTHOS AND THE GOVERNANCE GAP

8.1 Institutional Folk Beliefs as a Risk Surface

Every institution we have examined has developed, usually unconsciously, a working theory of what large language models are. Common patterns include treating LLMs as "autocomplete on steroids," as "junior analysts," as "oracles," or as "adversaries." A smaller mature minority treats them as a new class of system with distinctive affordances and failure modes — neither human nor deterministic software. The gap between the first four views and the fifth is where most LLM governance risk resides [14], [24], [52].

We use the term "LLM mythos" — colloquially, the "Claude mythos" after a prominent exemplar of the helpful-harmless-honest assistant model class — to denote this set of institutional folk beliefs. The mythos is not equivalent to the technology; it is the narrative infrastructure through which the technology is understood and governed.

8.2 Six Mythic Errors

We identify six recurrent mythic errors. The "just a chatbot" error underestimates the action-productive capacity of tool-augmented LLMs. The symmetric "cannot be trusted for anything" error dismisses governed deployments that would otherwise generate material productivity gains. The "vendor handles the risk" error misallocates accountability, given that institutional deployment — prompts, retrieval sources, tools, users, and data flows — is the deploying institution's responsibility [43], [48]. The "jailbreaks are not our problem" error ignores prompt-injection and indirect-injection attacks, which do not require breaking the model [28], [29]. The "shadow AI is minor" error under-weights empirical findings that a large majority of knowledge workers in financial services have used public AI tools for work [72]. The "we will know it when we see it" error assumes problematic use will be obvious, when in practice LLM outputs are fluent, confident, and error-prone in ways that propagate silently downstream [24], [52].

8.3 LLM-Specific Threat Surface

Beyond the mythic errors, ten concrete technical threat surfaces warrant explicit governance: (i) direct prompt injection; (ii) indirect prompt injection via retrieved content; (iii) sensitive information disclosure

through fine-tuning and RAG; (iv) training-data poisoning [73]; (v) excessive agency granted to deployed systems; (vi) model denial-of-service and cost exhaustion; (vii) vector-database and retrieval-source compromise; (viii) supply-chain and model-integrity risks; (ix) output-driven downstream attacks through AI-generated code, SQL, or shell commands; and (x) overreliance and skill atrophy in human reviewers [29], [52]. Each surface maps to concrete, testable controls that belong in the institution's Information Security Management System.

9. DIFFERENTIAL IMPACT ON FINANCIAL SERVICES SUBSECTORS

9.1 Why the Sector Is Disproportionately Exposed

Four structural features of financial services amplify exposure to the five vectors. First, identity-as-value: in no other sector is successful impersonation so directly convertible to currency. Second, concentration and interconnection: a small set of cloud providers, payment rails, clearing houses, correspondent banks, model providers, and KYC utilities creates systemic contagion paths [43]. Third, real-time irrevocability: the migration to instant settlement compresses detection-and-response windows toward zero [45]. Fourth, regulated disclosure: rapid-disclosure obligations (SEC four-day rule [74], DORA incident reporting [16], NYDFS 72-hour notification [75]) convert successful attacks into public-market consequences that adversaries can monetise through extortion and short-and-distort strategies.

9.2 Quantitative Exposure

Headline indicators align across independent sources. IBM's 2025 Cost of a Data Breach Report puts the average breach cost in financial services at approximately USD 6.1 million, roughly 40% above cross-industry average [4]. Global authorised-push-payment and BEC-related losses exceeded USD 25 billion in 2025 [46]. Deepfake-involved fraud incidents targeting banks grew by over 700% between 2023 and 2025 [76]. More than half of material incidents reported under DORA in Q1 2026 originated at a third-party provider [77].

9.3 Impact by Business Line

Retail and commercial banking exposure centres on call centres, digital channels, and branch-based identity verification, with particular vulnerability in treasury-management BEC scenarios. Capital markets face market-manipulation via synthetic content and fake news, cryptographic exposure across market-data integrity and settlement, and compressed reaction windows that amplify consequences. Asset and wealth management carry concentrated regulatory disclosure obligations and advisory relationships with high-net-worth targets; LLM-generated misadvice is an emerging liability channel. Insurance faces AI-generated claims fraud and systemically-correlated cyber-insurance underwriting risk [78]. Payments and market infrastructure are priority nation-state targets with systemic consequences; they bear the sector's most urgent obligation to lead PQC migration and harden agentic authorisation pathways. Digital-first fintechs, while often more advanced on defensive AI, are exposed through synthetic-KYC attacks, API security, and upstream cloud and model-provider dependency.

10. REGULATORY AND COMPLIANCE LANDSCAPE

Financial services institutions operate across a multi-jurisdictional landscape whose regulators are themselves racing to keep pace. We summarise the most consequential developments by domain rather than jurisdiction.

10.1 Artificial Intelligence

The EU AI Act [50], adopted in 2024 with phased effectiveness through 2027, imposes risk-tiered obligations with substantial financial-services relevance, particularly for credit scoring, insurance pricing, fraud detection, and employment selection applications classified as high-risk. In the United States, federal banking agencies have extended model risk management guidance (SR 11-7 and successors) to cover AI [79], and the SEC has issued rules and risk alerts on predictive data analytics [80]. State-level action,

notably NYDFS Circular Letter No. 7 (2024) on AI and Insurance [81] and Colorado SB 21-169, adds obligations. The UK has pursued a principles-based cross-regulator approach coordinated across FCA, PRA, and ICO [82]. Asia-Pacific regimes include the Monetary Authority of Singapore's Veritas framework and FEAT principles [83], the Hong Kong Monetary Authority's generative-AI guidelines [84], and Japan's soft-law approach.

10.2 Operational Resilience and Incident Reporting

DORA, effective January 2025, imposes the sector's most demanding ICT-risk regime, including third-party oversight for critical providers and mandatory threat-led penetration testing [16], [77]. The SEC cybersecurity disclosure rule requires Form 8-K disclosure within four business days of material-incident determination [74]. NYDFS 23 NYCRR 500, as amended in 2023, raised governance and notification expectations for covered entities [75]. U.S. federal banking agency computer-security incident notification rules require 36-hour notification of notification incidents to primary federal regulators [85]. UK PRA/FCA operational resilience expectations require important business service mapping and tested impact tolerances [86].

10.3 Post-Quantum Cryptography

NIST's 2024 PQC standards [37], [38], [39] anchor global migration. U.S. federal direction via NSM-10 [41] and OMB M-23-02 [42] has imposed inventory and planning obligations that cascade through vendors to regulated financial institutions. ENISA [87] and national authorities in the EU and UK have issued advisory guidance. Formal financial-sector expectations are anticipated to crystallise in 2027–2028.

10.4 Synthetic Media

Deepfake-specific regulation remains nascent. The EU AI Act imposes labelling requirements for synthetic content in limited contexts [50]. The most consequential financial-services requirement is derived from existing AML, KYC, and consumer-protection obligations: regulators increasingly expect institutions to demonstrate that identity verification, transaction authorisation, and communications channels remain effective in the face of synthetic media [81].

11. A UNIFIED DEFENCE FRAMEWORK

11.1 Design Principles

We propose that defence programs for the 2026–2032 horizon be built on six principles. (i) Assume adversary AI parity: plan for adversaries with equivalent model access to defenders. (ii) Assume content is untrusted: treat all text, audio, image, and video crossing a trust boundary as potentially synthetic. (iii) Assume agentic adversaries: response times must be machine-appropriate, with human-in-the-loop controls calibrated to avoid becoming exploitable bottlenecks. (iv) Assume identity is compromised: every authentication moment should be defensible under the assumption that traditional factors have been phished, cloned, or synthesised. (v) Assume long-tail compromises: estimate data confidentiality windows and align cryptographic lifetimes accordingly [12]. (vi) Assume continuous surprise: adversary capability is improving faster than any vendor roadmap.

11.2 Seven Control Domains

The framework comprises seven interlocking domains, mapped in Table 2 to NIST CSF 2.0 [15] and DORA [16].

Domain	Focus	NIST CSF 2.0	DORA
D1: Identity and NHI	Phishing-resistant MFA; NHI governance; JIT access	PR.AA, PR.AC, DE.CM	Art. 9
D2: Cryptography / PQC	Inventory; crypto-agility; hybrid PQC migration	PR.DS, GV.SC	Arts. 9, 28
D3: Content Authentication	C2PA; out-of-band verification; deepfake detection	PR.DS, PR.AA	Arts. 9, 10
D4: AI Governance & Security	Governance body; AI-RMF controls; red-teaming	GV.OC, PR.PS, DE.AE	Arts. 5, 6, 9
D5: Detection & Response	AI-native SOC; agentic defenders; AI TTPs	DE.CM, DE.AE, RS.MA	Arts. 10, 11
D6: Third-Party & Supply Chain	TPRM for AI; model supply chain; concentration	GV.SC, ID.RA	Arts. 28–44
D7: People, Culture, Training	Executive education; AI risk roles; paranoia	PR.AT, GV.RR	Art. 13

Table 2. Seven-domain defence framework mapped to NIST CSF 2.0 and DORA.

11.3 Selected Controls

Within each domain, we highlight the controls that receive the highest leverage from the convergent threat environment. Phishing-resistant MFA (FIDO2/WebAuthn, PIV/CAC, smart-card) for all workforce and privileged accounts is now a baseline rather than a target state [88]. A non-human identity (NHI) programme extending IAM rigour to AI agents — canonical registry, scoped credentials, short-lived tokens, policy-based authorisation, observability, and documented kill-switch — should be treated as a necessary rather than optional capability [9]. Cryptographic inventory should be a continuous capability integrated with the CMDB and SBOM programme. Hybrid classical-plus-PQC deployments follow current IETF guidance [71]. Content-provenance adoption via C2PA [66] should cover executive communications, regulatory disclosures, and, progressively, customer-facing content. AI red-teaming, either in-house or contracted, should be mandatory pre-production for high-risk use cases, extending the MITRE ATLAS taxonomy [51] into institutional practice. Detection engineering should incorporate AI-specific TTPs, including prompt-injection patterns, anomalous model use, and exfiltration via model APIs. Third-party risk assessments should test for AI use, PQC roadmap, deepfake-resistant authentication, and AI security posture, with enforceable contractual rights to audit.

12. A 24-MONTH PREPAREDNESS ROADMAP

We propose a phased 24-month roadmap calibrated for a large institution with a mature cybersecurity baseline. Smaller institutions should compress where feasible but should not defer items marked foundational. Each phase is defined by its outcome signals rather than its activities.

Horizon	Theme	Outcome Signals
Months 0–6	Foundations and Discovery	AI use-case registry live; cryptographic inventory scoped; NHI pilot operational; deepfake playbook adopted; AI governance body chartered
Months 6–12	Controls and Pilots	Phishing-resistant MFA enforced for privileged access; AI red team standing; PQC POCs running for code signing and TLS; C2PA pilot in executive channels
Months 12–18	Scale and Integrate	First-wave PQC migration (PKI, code signing, long-retention signing); deepfake-resistant customer channels; AI-native SOC at scale; governed RAG in production
Months 18–24	Operationalise and Mature	Crypto-agility across 80%+ of in-scope systems; NHI at parity with human IAM; sector exercises with agentic/synthetic scenarios; external attestations (ISO/IEC 42001); regulator briefings complete

Table 3. 24-month preparedness roadmap with horizon-level outcome signals.

Indicative budget ranges for a globally significant institution are USD 150–400 million across the 24-month window, covering cryptography, AI governance, identity modernisation, and AI-native security capabilities. Workforce implications include material expansion of AI-security, AI-risk, cryptography, and IAM teams; the market for qualified AI-security talent is tight and institutions should plan to build as much as to buy [53]. Leadership implications include explicit board-level engagement with AI and PQC programmes and standing risk-committee reporting covering AI risk, PQC progress, and deepfake exposure.

13. LIMITATIONS AND FUTURE RESEARCH

This paper has three limitations that we wish to make explicit. First, the empirical base for several of the trends described — particularly for Tier 3 agentic attacks and for HNDL exfiltration — remains thin, reflecting both recency and disclosure hesitancy. We anticipate that this picture will sharpen rapidly as DORA and SEC disclosures mature and as sector ISACs publish aggregated incident data. Second, the framework is a strategic orientation rather than an implementation specification; institutional maturity, regulatory environment, and risk appetite all condition the concrete controls that are appropriate. Third, we have deliberately avoided vendor and product recommendations, for reasons of neutrality and of expected obsolescence.

We identify six research directions that merit sustained attention. (i) Empirical measurement of prompt-injection prevalence and impact across production agentic deployments in regulated industries. (ii) Development of cost-effective, continuously updated cryptographic-inventory tooling, particularly for legacy and third-party-hosted systems. (iii) Operational evaluation of content-provenance standards at

scale, including usability in customer-facing contexts and resilience under adversarial attack. (iv) Formal accountability models for agent-originated harm, bridging model risk management, operational risk, and privacy law. (v) Quantitative modelling of sector-level contagion risk under AI-augmented attack scenarios, extending existing systemic-risk frameworks [43]. (vi) Comparative studies of the institutional LLM mythos across jurisdictions and institutional types, drawing on organisational sociology as well as computer science.

14. CONCLUSION

The 2020s will be remembered as the decade in which financial-services cybersecurity left its analogue assumptions behind. Generative AI, agentic AI, deepfakes, quantum computing, and the LLM mythos are each individually inflections; together they constitute a qualitatively new risk environment that is arriving on a timeline shorter than most institutional technology lifespans.

Our argument is not that any single one of these technologies is uniquely threatening. It is that the combination of them is, that it is arriving on a tight timeline, and that the institutions that begin their response programmes now will not merely avoid penalty but will define the sector standard to which regulators, customers, and counterparties will hold others. The financial services sector has previously risen to systemic technology challenges — Y2K, the internet, mobile banking, real-time payments. The discipline required to meet the present moment is of comparable ambition; what it lacks in novelty it makes up for in the compressed timeline and interdependence of the work.

We close by noting that security is a property not of systems but of the relationships among systems. When the systems change, the relationships must be renegotiated. That is the work of this decade.

Acknowledgements

The author thanks the anonymous practitioners in banking, insurance, capital markets, and market infrastructure who provided off-the-record context during the preparation of this manuscript. Any errors remain the author's own.

Conflict of Interest Statement

The author declares no financial or non-financial competing interests.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Data Availability

No new empirical datasets were generated in the course of this work. All sources cited are publicly available at the locations indicated in the references.

REFERENCES:

[1] B. Schneier, Click Here to Kill Everybody: Security and Survival in a Hyper-Connected World. New York: W. W. Norton, 2018.
[2] W. Diffie and M. E. Hellman, "New directions in cryptography," IEEE Transactions on Information Theory, vol. 22, no. 6, pp. 644–654, 1976.
[3] National Institute of Standards and Technology, "Digital Identity Guidelines," NIST SP 800-63-4 (Initial Public Draft), Gaithersburg, MD, 2024.
[4] IBM Security and Ponemon Institute, Cost of a Data Breach Report 2025, Armonk, NY: IBM, 2025.
[5] World Economic Forum, Global Cybersecurity Outlook 2025, Geneva: WEF, 2025.
[6] M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Praharaj, "From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy," IEEE Access, vol. 11, pp. 80218–80245, 2023.
[7] Europol, ChatGPT: The impact of Large Language Models on Law Enforcement, Europol Innovation Lab Tech Watch Flash, The Hague, 2023.

[8] L. Weng, "LLM-powered autonomous agents," Lil'Log, June 2023. [Online].
[9] S. Abdelnabi, K. Greshake, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection," in Proc. 16th ACM Workshop on Artificial Intelligence and Security (AISec), 2023, pp. 79–90.
[10] B. Chesney and D. K. Citron, "Deep fakes: A looming challenge for privacy, democracy, and national security," California Law Review, vol. 107, pp. 1753–1820, 2019.
[11] K. Chen, "Finance worker pays out $25 million after video call with deepfake 'chief financial officer'," CNN Business, February 4, 2024.
[12] M. Mosca, "Cybersecurity in an era with quantum computers: Will we be ready?," IEEE Security & Privacy, vol. 16, no. 5, pp. 38–41, 2018.
[13] Cloud Security Alliance, Practical Preparations for the Post-Quantum World, Seattle, WA: CSA, 2024.
[14] A. Narayanan and S. Kapoor, AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference. Princeton, NJ: Princeton Univ. Press, 2024.
[15] National Institute of Standards and Technology, The NIST Cybersecurity Framework (CSF) 2.0, NIST CSWP 29, Gaithersburg, MD, February 2024.
[16] European Parliament and Council, "Regulation (EU) 2022/2554 on digital operational resilience for the financial sector (DORA)," Official Journal of the European Union, L 333, 2022.
[17] A. Vaswani et al., "Attention is all you need," in Advances in Neural Information Processing Systems 30, 2017, pp. 5998–6008.
[18] T. B. Brown et al., "Language models are few-shot learners," in Advances in Neural Information Processing Systems 33, 2020, pp. 1877–1901.
[19] R. Bommasani et al., "On the opportunities and risks of foundation models," arXiv:2108.07258, 2021.
[20] R. Fang, R. Bindu, A. Gupta, Q. Zhan, and D. Kang, "LLM agents can autonomously exploit one-day vulnerabilities," arXiv:2404.08144, 2024.
[21] Google Threat Intelligence Group, Adversarial Misuse of Generative AI, Mountain View, CA: Google, 2025.
[22] M. Brundage et al., The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation, Future of Humanity Institute, Oxford, 2018.
[23] Microsoft Threat Intelligence, "Staying ahead of threat actors in the age of AI," Microsoft Security Blog, February 14, 2024.
[24] A. Narayanan and S. Kapoor, "GPT-4 and professional benchmarks: The wrong answer to the wrong question," AI Snake Oil Substack, March 2023.
[25] N. Shinn, F. Cassano, E. Berman, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," in Advances in Neural Information Processing Systems 36, 2023.
[26] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing reasoning and acting in language models," in Proc. ICLR, 2023.
[27] F. Perez and I. Ribeiro, "Ignore previous prompt: Attack techniques for language models," in NeurIPS ML Safety Workshop, 2022.
[28] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz, and M. Fritz, "Not what you've signed up for," arXiv:2302.12173, 2023.
[29] OWASP Foundation, OWASP Top 10 for Large Language Model Applications, Version 2025, 2025. [Online].
[30] I. J. Goodfellow et al., "Generative adversarial nets," in Advances in Neural Information Processing Systems 27, 2014, pp. 2672–2680.

[31] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in Advances in Neural Information Processing Systems 33, 2020, pp. 6840–6851.
[32] R. Chesney and D. K. Citron, "Deepfakes and the new disinformation war," Foreign Affairs, vol. 98, no. 1, pp. 147–155, 2019.
[33] M. Westerlund, "The emergence of deepfake technology: A review," Technology Innovation Management Review, vol. 9, no. 11, pp. 39–52, 2019.
[34] Y. Mirsky and W. Lee, "The creation and detection of deepfakes: A survey," ACM Computing Surveys, vol. 54, no. 1, pp. 1–41, 2021.
[35] M. Masood, M. Nawaz, K. M. Malik, A. Javed, A. Irtaza, and H. Malik, "Deepfakes generation and detection: State-of-the-art, open challenges, countermeasures, and way forward," Applied Intelligence, vol. 53, pp. 3974–4026, 2023.
[36] P. W. Shor, "Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer," SIAM Journal on Computing, vol. 26, no. 5, pp. 1484–1509, 1997.
[37] National Institute of Standards and Technology, Module-Lattice-Based Key-Encapsulation Mechanism Standard, FIPS PUB 203, Gaithersburg, MD, August 2024.
[38] National Institute of Standards and Technology, Module-Lattice-Based Digital Signature Standard, FIPS PUB 204, Gaithersburg, MD, August 2024.
[39] National Institute of Standards and Technology, Stateless Hash-Based Digital Signature Standard, FIPS PUB 205, Gaithersburg, MD, August 2024.
[40] D. J. Bernstein and T. Lange, "Post-quantum cryptography," Nature, vol. 549, pp. 188–194, 2017.
[41] The White House, National Security Memorandum on Promoting United States Leadership in Quantum Computing While Mitigating Risks to Vulnerable Cryptographic Systems (NSM-10), Washington, DC, May 2022.
[42] Office of Management and Budget, Migrating to Post-Quantum Cryptography, OMB M-23-02, Washington, DC, November 2022.
[43] Financial Stability Board, The Financial Stability Implications of Artificial Intelligence, Basel: FSB, November 2024.
[44] Basel Committee on Banking Supervision, Principles for the Sound Management of Operational Risk (revised), Basel: BIS, 2021.
[45] Verizon Business, 2025 Data Breach Investigations Report, Basking Ridge, NJ: Verizon, 2025.
[46] Federal Bureau of Investigation, Internet Crime Report 2024, Internet Crime Complaint Center, Washington, DC, 2025.
[47] FS-ISAC, Navigating Cyber 2025: Financial Services Cybersecurity Trends, Reston, VA: FS-ISAC, 2025.
[48] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1, Gaithersburg, MD, January 2023.
[49] International Organization for Standardization, ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system, Geneva, 2023.
[50] European Parliament and Council, "Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," Official Journal of the European Union, L series, 2024.
[51] MITRE Corporation, ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems, McLean, VA: MITRE, 2024. [Online].
[52] L. Weidinger et al., "Taxonomy of risks posed by language models," in Proc. ACM FAccT, 2022, pp. 214–229.
[53] R. Bommasani et al., "The Foundation Model Transparency Index," Stanford CRFM, 2023.
[54] M. Beckerich, L. Plein, and S. Coronado, "RatGPT: Turning online LLMs into proxies for malware attacks," arXiv:2308.09183, 2023.

[55] Mandiant (Google Cloud), M-Trends 2025: Cyber Front Lines, Reston, VA: Mandiant, 2025.
[56] Anti-Phishing Working Group, Phishing Activity Trends Report, 4th Quarter 2024, Cambridge, MA: APWG, 2025.
[57] Cybersecurity and Infrastructure Security Agency (CISA), AI Security Collaboration Playbook, Washington, DC: CISA, 2025.
[58] B. Buchanan, A. Lohn, M. Musser, and K. Sedova, Truth, Lies, and Automation: How Language Models Could Change Disinformation, Center for Security and Emerging Technology (CSET), Washington, DC, 2021.
[59] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems 33, 2020.
[60] A. Zou, Z. Wang, J. Z. Kolter, and M. Fredrikson, "Universal and transferable adversarial attacks on aligned language models," arXiv:2307.15043, 2023.
[61] N. Carlini et al., "Extracting training data from large language models," in Proc. USENIX Security, 2021, pp. 2633–2650.
[62] M. Jagielski et al., "Poisoning web-scale training datasets is practical," arXiv:2302.10149, 2023.
[63] National Institute of Standards and Technology, Secure Software Development Practices for Generative AI and Dual-Use Foundation Models, NIST SP 800-218A, Gaithersburg, MD, 2024.
[64] S. Kharpal, "AI fake image of Pentagon explosion briefly sends stocks lower," CNBC, May 23, 2023.
[65] Content Authenticity Initiative and Partnership on AI, State of Deepfake Detection Report, New York: CAI, 2025.
[66] Coalition for Content Provenance and Authenticity, C2PA Technical Specification v2.0, 2024. [Online].
[67] L. K. Grover, "A fast quantum mechanical algorithm for database search," in Proc. ACM STOC, 1996, pp. 212–219.
[68] National Academies of Sciences, Engineering, and Medicine, Quantum Computing: Progress and Prospects, Washington, DC: The National Academies Press, 2019.
[69] I. H. Deutsch, "Harnessing the power of the second quantum revolution," PRX Quantum, vol. 1, no. 2, 020101, 2020.
[70] National Institute of Standards and Technology, "NIST selects HQC as fifth post-quantum algorithm for standardization," NIST News, March 2025.
[71] D. Stebila, S. Fluhrer, and S. Gueron, "Hybrid key exchange in TLS 1.3," IETF Internet-Draft draft-ietf-tls-hybrid-design, 2024.
[72] Microsoft and LinkedIn, Work Trend Index Annual Report 2024: AI at Work Is Here, Now Comes the Hard Part, 2024.
[73] A. Wan, E. Wallace, S. Shen, and D. Klein, "Poisoning language models during instruction tuning," in Proc. ICML, 2023.
[74] U.S. Securities and Exchange Commission, Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure, Final Rule, 17 CFR Parts 229, 232, 239, 240, 249, Release Nos. 33-11216; 34-97989, July 2023.
[75] New York State Department of Financial Services, 23 NYCRR Part 500 — Cybersecurity Requirements for Financial Services Companies, Second Amendment, November 2023.
[76] Sumsub, Identity Fraud Report 2024, London: Sumsub, 2025.
[77] European Banking Authority, European Securities and Markets Authority, and European Insurance and Occupational Pensions Authority, Joint ESAs Report on DORA ICT-related Incident Reporting, Paris/Frankfurt, Q1 2026.

[78] Geneva Association, Cyber Insurance and Systemic Risk, Geneva: Geneva Association, 2023.
[79] Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, and Federal Deposit Insurance Corporation, Interagency Guidance on Model Risk Management (SR 11-7 and successors), 2011–2024.
[80] U.S. Securities and Exchange Commission, "Conflicts of interest associated with the use of predictive data analytics by broker-dealers and investment advisers," Proposed Rule, Release No. 34-97990, July 2023.
[81] New York State Department of Financial Services, Insurance Circular Letter No. 7 (2024): Use of Artificial Intelligence Systems and External Consumer Data and Information Sources in Insurance Underwriting and Pricing, Albany, NY, July 2024.
[82] HM Government, A Pro-Innovation Approach to AI Regulation, White Paper CP 815, London: HMSO, 2023.
[83] Monetary Authority of Singapore, Veritas Document 3A: Principles to Promote FEAT in the Use of AI and Data Analytics in the Financial Sector, Singapore: MAS, 2022.
[84] Hong Kong Monetary Authority, Use of Generative Artificial Intelligence by Authorized Institutions, Circular, Hong Kong: HKMA, August 2024.
[85] Board of Governors of the Federal Reserve System, Office of the Comptroller of the Currency, and Federal Deposit Insurance Corporation, Computer-Security Incident Notification Requirements for Banking Organizations and Their Bank Service Providers, Final Rule, Federal Register, vol. 86, no. 223, November 2021.
[86] Bank of England Prudential Regulation Authority and Financial Conduct Authority, Building Operational Resilience: Impact Tolerances for Important Business Services, Policy Statements PS6/21 and PS21/3, London, March 2021.
[87] European Union Agency for Cybersecurity (ENISA), Post-Quantum Cryptography — Integration Study, Athens: ENISA, 2022.
[88] Cybersecurity and Infrastructure Security Agency (CISA), Implementing Phishing-Resistant MFA, Fact Sheet, Washington, DC: CISA, 2022.