International Journal of
Science, Architecture, Technology and Environment

# Artificial Intelligence in Cyber Risk Management Across United States Banking: A Systematic Review and Case-Based Synthesis of Threat Intelligence, Fraud Detection, Security Operations Automation, and Cyber Resilience

Evidence on the Extent to Which Artificial Intelligence Improves Cyber Risk Detection and Incident Response in United States Financial Institutions

Ayomipo Alademehin[1*]

[1]*Lamar University, Beaumont, Texas, USA*

[*]*Corresponding author*

*DOI: https://doi.org/10.63680/ijsate062681.93*

## Abstract

Artificial intelligence has become central to how United States financial institutions detect cyber threats, prevent fraud, operate security operations centers, and pursue cyber resilience, yet the question of how much it improves cyber risk detection and incident response, as distinct from how much it is marketed as doing so, remains incompletely answered. This paper addresses that question through a systematic review and case-based synthesis of the peer-reviewed literature, the regulatory and supervisory record, the industry and vendor evidence base, and a set of documented institutional case studies, assembled and appraised according to a transparent and replicable protocol. The review is organized around four application domains identified as central to contemporary practice, namely threat intelligence, fraud detection, security operations automation, and cyber resilience, and it evaluates within each domain the evidence for the effect of artificial intelligence on the speed and accuracy of detection and on the speed and effectiveness of response.

The synthesis finds that the evidence for a substantial and favorable effect is strongest and most consistent in fraud detection, where adaptive machine-learning systems have repeatedly outperformed the static rule-based systems that preceded them on the dimensions of detection accuracy, false-positive reduction, and real-time operation, and where large institutions report meaningful and quantified loss avoidance. The evidence is strong but more qualified in security operations automation, where artificial intelligence has materially compressed the time required to detect, triage, and contain incidents and has relieved the alert-overload and skills-shortage pressures that have long degraded security operations, while introducing new dependencies and failure modes that the evidence is only beginning to characterize. The evidence in threat intelligence is favorable but harder to isolate from confounding factors, and the evidence on cyber resilience as a system-level outcome remains the least mature. Across all four domains, the synthesis identifies a consistent and

consequential countercurrent, namely that the same technology improving the defense is simultaneously available to the adversary, that artificial-intelligence systems introduce their own attack surface through adversarial manipulation and data poisoning, and that the opacity of the systems complicates the validation, explanation, and governance that the regulated banking context requires. The paper concludes that artificial intelligence improves cyber risk detection and incident response in United States banking to a degree that is real, domain-dependent, and meaningful but neither uniform nor unbounded, that the improvement is conditional on the data, the governance, and the human-machine integration that surround the technology, and that the net effect on cyber resilience depends on an adversarial co-evolution whose trajectory is not yet settled. A proposed agenda for the primary empirical research that the field still lacks is set out to guide the work that the systematic synthesis shows to be necessary.

***Keywords:*** Artificial intelligence; machine learning; cyber risk management; threat detection; incident response; fraud detection; threat intelligence; security operations center; cyber resilience; United States banking; financial institutions; adversarial machine learning; systematic review

## 1. Introduction

### 1.1 Background and Motivation

The United States banking system has become, over the past decade, one of the most intensively digitized and one of the most heavily targeted sectors of the economy, and the management of cyber risk has moved from a technical specialty at the periphery of the institution to a central concern of its boards, its regulators, and its customers. The financial sector concentrates the assets, the data, and the transactional infrastructure that make it an attractive target, and the consequences of a successful attack extend beyond the affected institution to the consumers whose financial lives it holds, the markets in which it operates, and the stability of the financial system itself. Against this backdrop, financial institutions have turned with increasing urgency to artificial intelligence, understood here to encompass the machine-learning systems that have matured over the past decade and the generative and foundation-model systems that have proliferated since 2023, as a means of detecting cyber threats more quickly, preventing fraud more accurately, operating their security functions more efficiently, and recovering from incidents more effectively than the human-driven and rule-based methods that preceded them could allow.

The scale of this turn is documented in industry evidence. Surveys conducted across 2024 and 2025 report that a substantial and rising majority of financial institutions now deploy artificial intelligence and machine learning in some part of their cyber risk and fraud functions, that spending on the detection of fraud and cyber threats has increased year over year at most institutions, and that the largest institutions now devote sums measured in the hundreds of millions of dollars annually to cybersecurity, a material fraction of which is directed at artificial-intelligence-enabled capability. The motivations for this investment are not difficult to identify. The volume of data that a modern financial institution must monitor for signs of compromise or fraud far exceeds what human analysts can review; the speed at which contemporary attacks unfold has compressed the time available for detection and response to minutes or hours; the shortage of skilled cybersecurity professionals has left security functions chronically understaffed; and the adversary has itself begun to deploy artificial intelligence, raising the sophistication and the scale of the attacks that the defender must counter. Artificial intelligence promises a response to each of these pressures, and the promise has been compelling enough to drive adoption across the industry at a remarkable pace.

Yet the very speed and breadth of this adoption raise a question that the enthusiasm surrounding it has tended to obscure. The claim that artificial intelligence improves the detection of cyber threats and the response to incidents is made constantly, by the institutions that deploy the technology, by the vendors that sell it, and by the consultants and analysts who advise on it, but the claim is made in many different forms, based on many different kinds of evidence, and with many different degrees of rigor. Some of the evidence is rigorous, peer-reviewed, and quantified; some is drawn from vendor materials whose methods are undisclosed and whose incentives are evident; and many lie between. For a financial institution deciding how much to invest in the technology, for a regulator assessing whether an institution's reliance on it is prudent, and for a scholar seeking to understand the technology's actual effect, the need to distinguish the well-supported claims from the poorly supported ones, and to characterize the conditions under which the technology delivers the improvement it promises, is pressing. It is the need that the present paper addresses.

### 1.2 The Evidentiary Problem

The difficulty confronting anyone who seeks to assess the effect of artificial intelligence on cyber risk management in banking is not a scarcity of claims but a surfeit of them, accompanied by a scarcity of the kind of rigorous, independent, and comparable evidence that would permit the claims to be evaluated. The evidence base is large, but it is heterogeneous in quality, fragmented across disciplines and sources, and shaped by incentives that complicate its interpretation. The voluminous vendor literature reports impressive improvements in detection accuracy, false-positive reduction, and response time, but its methods are typically undisclosed, its comparisons are frequently to unspecified baselines, and its authors have an evident interest in the favorable result. The institutional disclosures, found in regulatory filings and public statements, report adoption and investment and occasional outcomes, but they are constrained by the institutions' reluctance to disclose the details of their security postures and by the difficulty of attributing any particular outcome to any technology. The peer-reviewed literature is more rigorous, but it is dominated by studies that evaluate specific techniques on benchmark datasets under controlled conditions that may not reflect the operational banking environment, and studies that evaluate the effect of artificial intelligence on the actual cyber risk outcomes of actual financial institutions in operational conditions are comparatively rare.

This evidentiary problem is compounded by several features specific to the domain. The first is the difficulty of measurement, for the outcomes that matter, namely the threats detected, the frauds prevented, the incidents contained, and the resilience achieved, are difficult to observe directly and are frequently defined and measured differently across studies and institutions. The second is the problem of the counterfactual, for the effect of artificial intelligence on cyber risk is the difference between the outcomes with the technology and the outcomes that would have been obtained without it, and the latter is rarely observable, so that the apparent effect of the technology is confounded with the many other factors that change alongside its adoption. The third is the adversarial and dynamic character of the domain, for the cyber threat environment is not a fixed target against which a technology effect can be measured once and for all, but a co-evolving contest in which the adversary adapts to the defense, so that an improvement measured at one moment may erode as the adversary responds. The fourth is the secrecy that surrounds the domain, for institutions are reluctant to disclose the details of their cyber defenses, and the most informative evidence is frequently the least accessible. These features mean that the assessment of the technology effect cannot be a simple matter of tallying the reported improvements but must be a careful and critical synthesis that weighs the quality of the evidence, attends to the confounders and the incentives, and characterizes the conditions and the limits of the effect it finds.

### 1.3 Research Question and Objectives

This paper is organized around a single research question: to what extent does artificial intelligence improve cyber risk detection and incident response in United States financial institutions? The question is framed deliberately in terms of extent rather than in terms of a simple presence or absence of effect, because the interesting and consequential question is not whether artificial intelligence has any effect, which is not seriously in doubt, but how large the effect is, how it varies across the domains of application, how confident the available evidence permits one to be in it, and under what conditions and within what limits it obtains. The question is framed in terms of detection and response because these are the two functions at the heart of cyber risk management, the detection of the threat or the fraud or the incident and the response that contains and remediates it, and because the focus areas that motivate the paper, namely threat intelligence, fraud detection, security operations automation, and cyber resilience, all bear on these two functions.

The paper pursues this question through four objectives. The first is to assemble and appraise, according to a transparent and replicable protocol, the heterogeneous evidence base that bears on the question, distinguishing the well-supported claims from the poorly supported ones and weighing the quality of the evidence in each domain. The second is to synthesize the appraised evidence within each of the four focus domains, characterizing the extent of the effect of artificial intelligence on detection and response, the confidence that the evidence permits, and the conditions and limits of the effect. The third is to integrate the documented institutional case studies, drawn from the public record of named and anonymized institutions, with the synthesized evidence, grounding the general findings in the specific experience of the institutions that have deployed the technology. The fourth is to identify, from the gaps and weaknesses that the systematic review reveals in the existing evidence base, an agenda for the primary empirical research that the field still lacks, including the interview-based and quantitative studies that would provide the rigorous, independent, and operationally grounded evidence that the existing base does not supply.

### 1.4 Scope and Definitions

The scope of the paper is delimited along several dimensions. The sectoral scope is United States banking and the closely related financial institutions, including the commercial banks, the savings institutions, and the financial-technology firms that operate within or alongside the regulated banking system, chosen because this sector concentrates the cyber risk, the regulatory attention, and the artificial-intelligence adoption that the paper examines, and because confining the scope to a single regulatory and market context permits a more coherent synthesis than a global scope would allow. The functional scope is cyber risk detection and incident response. The two functions identified in the research question were examined across the four focus domains of threat intelligence, fraud detection, security operations automation, and cyber resilience. The temporal scope emphasizes the evidence of the period from approximately 2015 to 2026, during which machine learning matured into operational deployment and generative systems arrived, while drawing on earlier work that establishes the foundations.

Several definitions frame the analysis. Artificial intelligence is understood broadly, as noted above, to encompass both the machine-learning systems that learn patterns from data and the generative and foundation-model systems that have recently proliferated, in recognition that the cyber risk functions of contemporary banks deploy both. Cyber risk is understood as the risk of loss, disruption, or harm arising from the compromise of the institution's information systems, encompassing external attacks, fraud, and operational failures that the institution's cyber defenses address. Detection is understood as the identification of a threat, a fraud, an intrusion, or an incident, and is assessed along the dimensions of speed,

measured by metrics such as the mean time to detect, and accuracy, measured by the rates of true and false detection. Response is understood as the containment, the remediation, and the recovery that follow detection, and is assessed along the dimensions of speed, measured by metrics such as the mean time to respond or contain, and effectiveness, measured by the degree to which the response limits the harm. Cyber resilience is understood, following the regulatory usage, as the capacity of the institution to anticipate, withstand, recover from, and adapt to cyber incidents, a system-level outcome to which detection and response contribute but which they do not exhaust.

### 1.5 Contributions and Structure

The paper makes four contributions. First, it provides a systematic and transparently conducted synthesis of the heterogeneous evidence on the effect of artificial intelligence on cyber risk management in United States banking, appraising the quality of the evidence and distinguishing the well-supported from the poorly supported claims, where much of the existing discussion proceeds without such appraisal. Second, it organizes the synthesis around the four focus domains and characterizes the extent, the confidence, and the conditions of the effect within each, providing a differentiated rather than a uniform account of a technology whose effect varies considerably across its applications. Third, it integrates documented institutional case studies with synthesized evidence, grounding the general findings in specific and verifiable institutional experience drawn entirely from the public record. Fourth, it identifies, from the gaps the review reveals, a concrete agenda for the primary empirical research that the field requires, converting the limitations of the existing evidence base into a forward program of work. Throughout, the paper observes strict discipline with respect to evidence, relying only on documented and attributable sources, refraining from the manufacture of data, and clearly labelling the boundary between what the evidence establishes and what it leaves open.

The remainder of the paper proceeds as follows. Section 2 sets out the methodology of the systematic review, including the protocol, the sources and search strategy, the inclusion criteria, the evidence-appraisal approach, the case-study component, and the limitations of the method. Section 3 establishes the context, examining the changing cyber threat environment in banking, the limits of the rule-based and signature-based defenses that artificial intelligence is displacing, and the operational pressures of alert overload and the skills gap that motivate its adoption. Sections 4 through 7 present the domain syntheses, treating threat intelligence, fraud detection, security operations automation, and cyber resilience in turn, and within each, presenting the role of the technology, the evidence of its effect, and the case evidence and synthesis. Section 8 examines the adversarial counter current and the limits of technology, including offensive artificial intelligence, attacks on the defensive systems themselves, and opacity, false-positivity, and governance burdens. Section 9 integrates the domain findings into a cross-domain synthesis and discussion. Section 10 sets out the proposed agenda for primary empirical research. Section 11 concludes.

## 2. Methodology of the Systematic Review

The credibility of a synthesis depends on the transparency and the rigor of the method by which it is conducted, and this section sets out the methodology of the present review in sufficient detail to permit its appraisal and, in principle, its replication. The review is a systematic review in the sense that it follows a defined protocol for the identification, the selection, the appraisal, and the synthesis of the evidence, adapted from the established conventions of systematic review to the particular character of the present subject, in which much of the relevant evidence lies outside the peer-reviewed literature in the regulatory, industry, and institutional records. The section describes the review protocol and approach, the sources and the

search strategy, the inclusion criteria, the appraisal of the evidence and the assessment of its quality, the case-study component, and the limitations of the method.

### 2.1 Review Protocol and Approach

The review was conducted according to a protocol defined in advance of the synthesis, following the spirit of the established reporting standards for systematic reviews while adapting their procedures to a subject in which the evidence is heterogeneous in type and quality and distributed across the scholarly, regulatory, and industry literatures. The protocol specified the research question, the four focus domains, the sources to be searched, the criteria for the inclusion and the exclusion of the evidence, the framework for the appraisal of the quality of the evidence, and the approach to the synthesis. The protocol distinguished, in particular, among the several types of evidence that bear on the question, namely the peer-reviewed scholarly evidence, the regulatory and supervisory evidence, the industry and survey evidence, the vendor evidence, and the institutional case evidence, and it specified for each type the weight to be accorded to it in the synthesis and the caution with which it was to be treated.

The approach to the synthesis is best characterized as a narrative and thematic synthesis rather than a meta-analysis, a choice dictated by the nature of the evidence. A meta-analysis, which would pool the quantitative results of comparable studies into a single estimate of effect, is not feasible for the present subject, because the studies that bear on the question are too heterogeneous in their outcomes, their measures, their settings, and their methods to be pooled meaningfully, and because much of the relevant evidence is not of the form that a meta-analysis requires. The synthesis instead proceeds thematically, organizing the evidence within each focus domain, characterizing the direction and the magnitude and the consistency of the effect that the evidence indicates, appraising the quality and the weight of the evidence, and arriving at a judgment, appropriately hedged, regarding the extent of the effect and the confidence that the evidence permits. This approach sacrifices the apparent precision of a pooled quantitative estimate, which the evidence could not in any case support, in Favor of a transparent and critical synthesis that attends to the quality and the limits of the evidence as a pooled estimate would not.

### 2.2 Sources, Search Strategy, and Inclusion Criteria

The review drew on five categories of sources. The peer-reviewed scholarly literature was identified through searches of the major scholarly databases and indexes for the terms associated with artificial intelligence, machine learning, cyber risk, threat detection, fraud detection, security operations, and cyber resilience in the financial and banking context, supplemented by the examination of the references of the identified works and of the works that cited them. The regulatory and supervisory literature was identified through the publications of the United States financial regulators and the interagency bodies, including the supervisory guidance, the examination handbooks, the reports, and the bulletins that bear on the use of artificial intelligence and on cyber risk and resilience in financial institutions. The industry and survey literature were identified through the reports of the established industry research organizations, the surveys of financial institutions and security operations, and the analyses of the consultancy and industry bodies. The vendor literature was identified through the published materials of the principal providers of artificial-intelligence-enabled security and fraud technology. The institutional case evidence was identified through the public disclosures, the regulatory filings, the public statements, and the documented incidents of named and anonymized financial institutions.

The inclusion criteria required that a source bear directly on the effect of artificial intelligence on cyber risk detection or incident response in the financial or banking context, that it postdate approximately 2015 except where an earlier work establishes a necessary foundation, and that it provide evidence of a quality and a specificity sufficient to contribute to the synthesis. The exclusion criteria removed the sources that addressed artificial intelligence or cyber risk in general without bearing on their intersection in the financial context, the sources that merely asserted the effect of the technology without providing evidence of it, and, in the appraisal, the sources whose quality was insufficient to support the weight that reliance on them would require. The vendor evidence was subjected to scrutiny, including where it provided specific and verifiable information but was discounted where its methods were undisclosed and its incentives evident, and never relied on as the sole support for a finding. The aim throughout was not to assemble the largest possible body of evidence but to assemble a body of evidence of sufficient quality to support a credible synthesis, and to appraise the quality of each piece of evidence as part of the synthesis itself.

### 2.3 Evidence Appraisal and Quality Assessment

The appraisal of the evidence proceeded along several dimensions adapted to the heterogeneous character of the evidence base. The first dimension was the independence of the source, the degree to which the source was free of interest in the result it reported, with the peer-reviewed and the regulatory evidence having the greatest independence and the vendor evidence the least. The second dimension was the transparency of the method, the degree to which the source disclosed how its results were obtained, with the studies that disclosed their data, their measures, and their methods, according to greater confidence than the sources that reported results without disclosing their basis. The third dimension was the specificity and the verifiability of the evidence, the degree to which the source provided specific, quantified, and verifiable information rather than general assertions. The fourth dimension was the operational realism of the evidence, the degree to which it reflected the actual conditions of operational banking rather than the controlled conditions of a benchmark evaluation, a dimension on which much of the rigorous scholarly evidence is paradoxically weak and much of the institutional evidence comparatively strong.

The appraisal along these dimensions informed the weight accorded to each piece of evidence in the synthesis and the confidence attached to the findings it supported. A finding supported by multiple independent sources of high quality across several of the dimensions was accorded high confidence; a finding supported only by sources weak on one or more of the dimensions, such as vendor evidence of undisclosed method or scholarly evidence of limited operational realism, was accorded lower confidence and reported with the appropriate hedging. The appraisal was not reduced to a single numerical score, which the heterogeneity of the evidence would not support, but was conducted as a structured and transparent judgment that the reader can examine and contest. The synthesis reports, for each principal finding, not only the direction and the magnitude of the effect that the evidence indicates but the confidence that the appraisal of the evidence permits, distinguishing throughout the findings that the evidence establishes from those that it merely suggests.

### 2.4 The Case-Study Component

The systematic review is complemented by a case-study component that grounds the general findings of the synthesis in the specific and documented experience of financial institutions. The case studies were selected to illustrate the deployment of artificial intelligence across the four focus domains, to span the range of institutional types from the largest global banks to the smaller and financial-technology institutions, and to

draw on the public record of documented institutional experience. The cases rely entirely on the public and attributable record, including the institution's own disclosures and public statements, the regulatory filings, the documented incidents, and the analyses of the institutions by independent parties, and they refrain from any reliance on undisclosed or unverifiable information. Where an institution is named, the information attributed to it is drawn from the public record and is attributable to an identified source; where the evidence concerns an institution that cannot be named, it is drawn from documented but anonymized accounts and is treated with the additional caution that anonymity requires.

The case-study component serves several functions in the analysis. It grounds the general findings of the synthesis in specific experience, demonstrating that the effects the synthesis identifies are realized, or are not realized, in the actual practice of actual institutions. It surfaces the conditions and the limits of the effects, for the case studies reveal the data, the governance, and the human-machine integration on which the realization of the effects depends. And it illustrates the variation across institutional types, for the experience of the largest institutions, which deploy the most sophisticated capability at the greatest scale, differs from that of the smaller institutions, which frequently access the technology through third-party vendors and confront the constraints of more limited resources. The case studies are presented within the domain sections, each illustrating the domain it accompanies, and they are integrated into the cross-domain synthesis of Section 9.

### 2.5 Limitations of the Method

The method has limitations that must be acknowledged and that condition the findings it supports. The first and most fundamental is that the review synthesizes existing evidence and does not generate new primary evidence, so that its findings are only as strong as the evidence base permits, and the weaknesses of that base, including the scarcity of independent and operationally realistic evidence and the prevalence of evidence shaped by the incentives of its producers, constrain the confidence of the synthesis. The review is candid throughout about these constraints, and the proposed research agenda of Section 10 is directed precisely at the generation of the primary evidence that the existing base lacks, but the present synthesis cannot transcend the limits of the evidence it reviews. The second limitation is the difficulty of the appraisal itself, for the judgment of the quality and the weight of heterogeneous evidence is inevitably a matter of judgment on which reasonable assessors might differ, and the review mitigates this difficulty through the transparency of its appraisal but cannot eliminate it.

The third limitation is the dynamism of the subject, for the technology and the threat environment are both evolving rapidly, and a synthesis conducted at a moment captures the state of the evidence at that moment and may be overtaken by subsequent developments, a limitation particularly acute for the generative and agentic systems whose deployment is recent and whose evidence base is still forming. The fourth limitation is the secrecy of the domain, for the most informative evidence, concerning the actual cyber defenses and the actual incidents of actual institutions, is frequently the least accessible, and the synthesis is constrained to the evidence that is in the public record, which may not be representative of the experience that is not. The fifth limitation is the focus on the United States banking context, which permits a coherent synthesis but limits the generalizability of the findings to other sectors and jurisdictions. These limitations do not vitiate the synthesis, which remains the most rigorous assessment that the available evidence supports, but they condition its findings and underscore the need for the primary research that the review identifies.

## 3. The Threat Landscape and the Case for Artificial Intelligence

The case for artificial intelligence in cyber risk management rests on the proposition that the contemporary threat environment, and the operational pressures that accompany it, have outrun the capacity of the human-driven and rule-based methods that preceded the technology, and that artificial intelligence offers a response to pressures that the prior methods cannot meet. This section establishes the context for the domain syntheses that follow by examining the changing cyber threat environment in banking, the limits of the rule-based and signature-based defenses that artificial intelligence is displacing, and the operational pressures of alert overload and the skills gap that motivate the adoption of technology. The examination establishes both the genuine need that artificial intelligence addresses and the baseline against which its effect must be measured.

### 3.1 The Changing Cyber Threat Environment in Banking

The cyber threat environment confronting United States financial institutions has changed in several respects that bear directly on the case for artificial intelligence. The volume of attacks has risen, as the digitization of banking has expanded the attack surface and as the tools of attack have become more widely available, so that institutions confront a constant and rising flow of attempted intrusions, frauds, and compromises. The speed of attacks has increased, with evidence indicating that the interval between the initial compromise and the consequential harm has compressed dramatically, in some documented cases to a matter of minutes, leaving a correspondingly compressed window for detection and response. The sophistication of attacks has grown, as adversaries have adopted more advanced techniques, including the artificial intelligence that the defenders also deploy, raising the difficulty of distinguishing the malicious from the benign. And the character of the attacks has shifted, with the evidence indicating that the unauthorized-party fraud driven by credential theft and account takeover has come to account for a large majority of fraud incidents and losses, and that the involvement of third parties in breaches has risen as the supply chains and the vendor relationships of institutions have expanded the perimeter that must be defended.

The financial consequences of this environment are substantial and rising. The industry evidence projects that the losses from fraud enabled by generative artificial intelligence in the United States will rise several-fold over the period to 2027, that the volume of the synthetic media used in fraud is doubling at a rapid cadence, and that the global costs of cybercrime have reached a scale measured in the trillions of dollars annually. The specific evidence for banking indicates that the average cost of a data breach has risen year after year and that the financial sector bears costs among the highest of any sector, reflecting the value of the assets and the data it holds, and the regulatory consequences of their compromise. This rising consequence raises the stakes of detection and response and sharpens the case for any technology that can improve them, while also raising the cost of the failures, including the false positives and the missed detections, that an imperfect technology can produce. The changing threat environment thus establishes both the need that artificial intelligence addresses and the consequences against which its successes and failures must be weighed.

### 3.2 The Limits of Rule-Based and Signature-Based Défense

The methods that artificial intelligence is displacing in cyber risk management are predominantly rule-based and signature-based, and an understanding of their limits is necessary both to the case for the technology and to the assessment of its effect, for the effect of artificial intelligence is in large part the difference between

what it can do and what these prior methods could. The rule-based methods detect threats and fraud by the application of predefined rules, flagging the transactions or events that match the patterns that the rules encode, and the signature-based methods detect threats by matching them against a database of the signatures of known threats. Both methods are effective against the threats they are designed to detect, and both have the virtues of transparency and of predictability, for the basis of every detection is explicit in the rule or the signature that produced it. These virtues account for the long dominance of the methods and for their continuing role in contemporary defense.

The methods suffer, however, from limits that the changing threat environment has rendered increasingly consequential. The first limit is their inability to detect the threats they were not designed to detect, for a rule-based method detects only the patterns its rules encode, and a signature-based method detects only the threats whose signatures it holds, and both are therefore blind to the novel threats, the zero-day exploits, and the previously unseen fraud patterns that the adaptive adversary continually produces. The second limit is their rigidity, for the rules and the signatures must be defined and updated by human analysts, a process that cannot keep pace with the volume and the velocity of the evolving threats, so that the methods are perpetually behind the adversary. The third limit is their tendency, when tuned to detect a wide range of threats, to generate a high volume of false positives, flagging the benign events that happen to match the patterns the rules encode and burdening the human analysts who must adjudicate the flags. These limits are not incidental but structural, arising from the fundamental character of the methods as the application of predefined patterns, and they establish the space that artificial intelligence, with its capacity to learn the patterns of the threats from the data rather than to apply the patterns defined in advance, is deployed to fill.

### 3.3Operational Pressures: Alert Overload and the Skills Gap

Beyond the limits of the prior methods, the adoption of artificial intelligence is driven by two operational pressures that the changing threat environment has intensified and that bear directly on the capacity of the institution to detect and respond. The first is the pressure of alert overload, the phenomenon by which the security functions of institutions are inundated with a volume of alerts that far exceeds the capacity of their human analysts to investigate. The evidence indicates that the security operations of institutions receive thousands of alerts daily, that the human analysts can investigate only a fraction of them, and that the necessity of choosing which alerts to investigate and which to ignore introduces the risk that the consequential alerts will be lost among the inconsequential ones. The alert overload arises in part from the false positives that the rule-based methods generate and in part from the sheer volume of the events that the digitized institution produces, and it degrades the detection that the security function exists to provide, for the threat that generates an alert that is never investigated is not detected.

The second pressure is the shortage of skilled cybersecurity professionals, a shortage that has left the security functions of institutions chronically understaffed and that the rising volume and sophistication of the threats has rendered increasingly acute. The evidence indicates that institutions struggle to recruit and retain the skilled analysts that their security operations require, that many operate with security teams smaller than their threat environment warrants, and that the shortage degrades both the detection and the response that the security function provides. The two pressures compound each other, for the alert overload demands more analyst capacity precisely as the skills gap constrains the supply of it, and together they establish the operational case for a technology that can automate the routine work of the security function, triage the flood of alerts, and multiply the effective capacity of the scarce human analysts. It is against these pressures, as much as against the limits of the prior methods, that the effect of artificial intelligence must be

measured, for much of the value that the technology is claimed to provide consists precisely in the relief of the alert overload and the augmentation of the scarce analyst capacity that these pressures create.

### 3.4 The Economics of the Cyber Threat and the Stakes of Detection

The case for artificial intelligence in cyber risk management rests not only on the changing character of the threats but on the economics of the contest between the attacker and the defender, and an understanding of that economics is necessary to the assessment of the technology effect. The economics of the cyber threat have shifted in the attacker's favor along several dimensions that bear on the case for the technology. The marginal cost to the attacker of mounting an attack has fallen, as the tools of the attack have become more widely available and as artificial intelligence has automated the work that the attack formerly required, so that the attacker can mount more attacks at a lower cost. The marginal return to the attacker of a successful attack has risen, as the digitization of banking has concentrated on the assets and the data that the attack targets, so that the successful attack yields a greater reward. And the asymmetry between the attacker and the defender has widened, for the attacker needs to succeed only once while the defender must succeed always, and the falling cost and the rising return of the attack multiply the attempts against which the defender must succeed.

This shifting economics raises the stakes of detection and sharpens the case for any technology that can improve it, while also raising the cost of the failures that an imperfect technology can produce. The value of improvement in detection is the harm that the improved detection averts, which grows as the economics of the threat rise the frequency and the consequence of the attacks, so that the value of the artificial-intelligence-enabled detection grows with the worsening of the threat environment. The cost of a failure of detection, whether a missed attack or a false alarm, also grows with the worsening environment, for the missed attack yields the attacker a greater reward and the false alarm consume the scarce defensive capacity that the rising volume of the attacks demands. The economics of the threat thus establishes both the growing value of the detection that artificial intelligence promises and the growing cost of the failures that an imperfect technology can produce, and it frames the assessment of the technology effect as a question not merely of whether the technology improves the detection but of whether the improvement is sufficient to keep pace with the worsening economics of the threat against which it is deployed.

### 3.5 The Third-Party and Supply-Chain Dimension

A dimension of the changing threat environment that warrants particular attention, both because of its growing prominence and because of its implications for the artificial-intelligence-enabled defense, is the rising involvement of third parties in the breaches that financial institutions suffer. The evidence indicates that the involvement of third parties in breaches has risen substantially, that the supply chains and the vendor relationships of institutions have expanded the perimeter that the institution must defend, and that the institution's security now depends not only on its own defenses but on the defenses of the third parties on which it relies. This dimension complicates the artificial-intelligence-enabled defense, for the institution that deploys the technology to defend its own systems may remain exposed through the third parties whose systems it does not control, and the technology that detects the threats to the institution's own systems may not extend to the threats that enter through the third parties.

The third-party dimension also bears on the institution's reliance on artificial intelligence itself, for the institution that accesses the technology through the third-party vendors, as many of the smaller institutions

do, introduces a dependence on the vendors whose systems it does not control and whose failures, whether through error or through compromise, become a source of the institution's risk. The concentration of the institutions reliance on a small number of the providers of the artificial-intelligence-enabled security and fraud technology introduces, moreover, a systemic dimension, for the compromise or the failure of a widely used provider could propagate across the many institutions that rely on it, a concern that the evidence on the concentration of the technology provision raises and that the supervisory attention to the third-party and the concentration risk reflects. The third-party and supply-chain dimension thus both complicate the artificial-intelligence-enabled defense, by extending the perimeter beyond the institution's own systems, and introduce a new dependence and a new systemic risk, through the institution's reliance on the third-party providers of the technology, and it figures in the assessment of the technology's net effect that the synthesis develops.

### 3.6 Establishing the Baseline: What Artificial Intelligence Must Improve Upon

The assessment of the effect of artificial intelligence requires a clear baseline, the level of detection, and the response that the prior methods achieved, against which the effect of the technology can be measured, for the effect of the technology is the difference between what it achieves and what the prior methods achieved, and the assessment requires the baseline as much as the achievement. The baseline is established by the rule-based and the signature-based methods examined in the preceding sections, and by the human-driven processes that surrounded them, and it is characterized by the limits that those methods and processes exhibited, the blindness to the novel threats, the rigidity in the face of the evolving threats, the false positives that burdened the operations, and the alert overload and the skills gap that degraded the detection and the response. The baseline is not a level of zero detection, for the prior methods detected the threats they were designed to detect, and the human analysts investigated the alerts they had the capacity to investigate, but a level of detection and response constrained by the limits that the prior methods and the operational pressures imposed.

The clarity of the baseline matters to the assessment in two respects. The first is that it establishes the magnitude of the improvement that the technology must achieve to be worth its cost, for the technology is worth its cost only if its improvement over the baseline exceeds the cost of its deployment and the new risks it introduces, and the assessment of its worth requires the baseline against which the improvement is measured. The second is that it cautions against the attribution to the technology of the improvements that other factors produce, for the detection and the response improve over time for many reasons, including the improvement of the prior methods, the investment in the human analysts, and the maturation of the security practices, and the attribution of the improvement to the technology requires the baseline that isolates the technology effect from the other factors. The establishment of the baseline is, in consequence, a necessary part of the assessment, and the synthesis attends throughout to the baseline against which the effect of the technology is measured, cautioning against the attribution to the technology of the improvements that the baseline itself, through the improvement of the prior methods and the other factors, would have produced.

## 4. Threat Intelligence

The first of the four focus domains is threat intelligence, the function by which an institution gathers, analyzes, and acts upon information about the threats it faces, anticipating the attacks that are coming rather than merely reacting to the attacks that have arrived. This section examines the role of artificial intelligence in threat intelligence, the evidence for its effect on the speed and the accuracy of detection, and the anticipation of threats, and the case evidence that grounds the synthesis in documented institutional experience. The domain is treated first because threat intelligence is, in a sense, logically before the others, informing the detection of fraud, the operation of the security function, and the pursuit of resilience with the knowledge of the threats against which each must defend.

### 4.1 The Role of Artificial Intelligence in Threat Intelligence

Threat intelligence in its traditional form is a labor-intensive function in which human analysts gather information about threats from a wide range of sources, including the technical indicators of compromise, the communications of the threat actors, the disclosures of vulnerabilities, and the reports of incidents at other institutions, and synthesize this information into the actionable intelligence that informs the institution's defenses. The function is constrained by the volume of the information, which far exceeds what human analysts can review, by the speed at which the information must be processed to remain actionable, and by the difficulty of distinguishing the signal from the noise in a vast and heterogeneous flow of data. Artificial intelligence is deployed in threat intelligence to address each of these constraints, automating the gathering and the initial analysis of the information, processing volumes that exceed human capacity at speeds that preserve the actionability of the intelligence, and applying the pattern-recognition capacity of machine learning to the separation of the signal from the noise.

The specific applications of artificial intelligence in threat intelligence span several functions. Technology is applied to the automated collection and correlation of the indicators of compromise from the many sources that produce them, assembling a coherent picture of the threat environment from the fragments distributed across the sources. It is applied to the analysis of the communications and the behavior of the threat actors, including the monitoring of the venues in which the actors operate and the identification of the emerging threats that the actors discuss. It is applied to the prediction of the threats that an institution is likely to face, drawing on the patterns of past attacks and the characteristics of the institution to anticipate the attacks that are coming. And, with the arrival of generative and large language model systems, it is applied to the synthesis and the communication of intelligence, generating the summaries and the analyses that the human analysts and the decision-makers require from the underlying data. Across these applications, the role of technology is to multiply the capacity of the threat-intelligence function, processing more information more quickly and surfacing the patterns that the human analysts, constrained by the volume and the velocity of the data, would miss.

### 4.2 Evidence of Effect

The evidence for the effect of artificial intelligence on threat intelligence is favorable in its direction but is harder to isolate and to quantify than the evidence in the other domains, for the effect of threat intelligence is mediated through the other functions it informs and is therefore difficult to measure directly. The evidence that does bear on the question indicates that the application of artificial intelligence to threat intelligence increases the volume of the information that can be processed, compresses the time required to process it,

and improves the identification of the relevant threats from the flow of data, and the institutions that have deployed the technology report that it has enhanced their capacity to anticipate and to prepare for the threats they face. The acquisition by a major payment network of a specialized provider of artificial-intelligence-driven threat intelligence, at a cost measured in the billions of dollars, is itself evidence of the value that a sophisticated and well-resourced institution attaches to the capability, for the acquisition reflects a considered judgment that the capability is worth a substantial investment.

The confidence that this evidence permits is, however, tempered by several considerations that the appraisal must weigh. The first is the difficulty of the counterfactual, for the effect of the artificial- intelligence-enabled threat intelligence is the difference between the threats anticipated and prevented with it and those that would have been anticipated and prevented without it, and the latter is not observable, so that the apparent effect is confounded with the many other factors that bear on the institution's security. The second is the mediation of the effect, for threat intelligence does not detect or prevent threats directly but informs the functions that do, so that its effect is realized only through the fraud detection, the security operations, and the resilience that it informs, and is correspondingly difficult to attribute to the threat intelligence as distinct from the functions it serves. The third is the prevalence of vendor evidence in the domain, for much of the evidence for the effect of artificial intelligence on threat intelligence is produced by the providers of the technology, whose incentives the appraisal must discount. The evidence is therefore assessed as favorable but confounded, supporting the conclusion that artificial intelligence improves the threat-intelligence function while cautioning that the magnitude of the improvement is harder to establish than in the domains where the effect is more directly measurable.

### 4.3 Case Evidence and Synthesis

The case evidence in threat intelligence centers on the documented experience of the major payment networks and the largest banks, which have made the most substantial and visible investments in capability. The acquisition by a major payment network of a specialized artificial-intelligence threat-intelligence provider, at a cost of several billion dollars, is the most prominent documented instance and illustrates both the value the institution attached to the capability and the strategy of acquiring rather than building it. The payment network has reported that the acquired capability, integrated with its existing fraud and security functions, has enhanced its ability to identify and mitigate threats across its network by analyzing transactional and threat data to surface anomalies and emerging threats that inform its defenses. The largest banks have made comparable, if less publicly detailed, investments, with the documented evidence indicating that the institutions that invest the most heavily in cybersecurity, measured in hundreds of millions of dollars annually, direct a material part of that investment to the artificial-intelligence-enabled threat intelligence that informs their defenses.

The synthesis of the threat-intelligence domain arrives at a judgment of a favorable but confounded effect, accorded to moderate confidence. The direction of the effect is clear and is supported across the sources, for artificial intelligence demonstrably increases the volume, the speed, and the discrimination of the threat-intelligence function, and the investments considered by the most sophisticated institutions reflect their judgment of its value. The magnitude of the effect is harder to establish, for the mediation of the effect through the functions that threat intelligence informs, the difficulty of the counterfactual, and the prevalence of vendor evidence all complicate its quantification. The conditions of the effect are visible in the case evidence, which indicates that the value of the artificial-intelligence-enabled threat intelligence is realized only when it is integrated with the functions it informs and only when the institution possesses the data and

the capability to act upon the intelligence it produces. The threat-intelligence domain thus exemplifies a pattern that recurs across the synthesis, in which the direction of the effect of artificial intelligence is clear and favorable but its magnitude is conditioned by the integration, the data, and the capability that surround the technology and is harder to establish than the enthusiasm surrounding the technology would suggest.

### 4.4 The Generative Turn in Threat Intelligence

The arrival of the generative and the large language model systems has begun to transform the threat-intelligence function in ways that the synthesis of the preceding subsections, focused on the established machine-learning applications, has only begun to capture, and the generative turn warrants the separate treatment that its prominence and its recency justify. The generative systems are applied in threat intelligence to the synthesis and the communication of the intelligence, generating the summaries, the analyses, and the briefings that the human analysts and the decision-makers require from the underlying data, and compressing the time and the labor that the synthesis of the intelligence formerly required. The systems are applied, further, to the interrogation of intelligence, permitting the analysts to query the vast body of threat data in the natural language that the generative systems understand, and surfacing the patterns and connections that the manual analysis would miss. And the systems are applied to the generation of threat scenarios and the simulations that inform the institution's preparation, modelling the attacks that the institution might face and the defenses that would counter them.

The generative turn introduces, alongside its benefits, the hazards specific to the generative systems that the synthesis must weigh, and that bear on the assessment of its effect. The generative systems are prone to the confabulation that produces the confident but false output, a hazard particularly consequential in the threat-intelligence context, where the false intelligence could misdirect the institution's defenses, and the deployment of the generative systems in threat intelligence requires the human oversight that detects and corrects the confabulation. The generative systems are vulnerable, further, to the manipulation that the adversary can mount through the crafted input, and the threat-intelligence systems that ingest the data from many sources, some of which the adversary can influence, are exposed to the manipulation that the adversary could mount through the poisoning of the sources. The generative turn in threat intelligence thus extends the favorable effects that the synthesis documents while introducing the generative-specific hazards that the deployment must manage, and it illustrates, in the newest application of the technology to the threat-intelligence function, the pattern that recurs across the synthesis, in which the favorable effects of the technology are accompanied by the new hazards that the technology introduces and that the responsible deployment must address.

## 5. Fraud Detection

The second focus domain is fraud detection, the function by which an institution identifies the fraudulent transactions and activities among the vast volume of the legitimate ones, and it is the domain in which the evidence for a substantial and favorable effect of artificial intelligence is strongest and most consistent. This section examines the transition from the rule-based to the adaptive-learning methods of fraud detection, the evidence for the effect of artificial intelligence on the accuracy of detection and the reduction of false positives, and the case evidence that grounds the synthesis in documented institutional experience. The domain is central to the cyber risk management of banking, for fraud is among the most direct and the most measurable harm that the institution's cyber defenses address, and it is the domain in which the longest experience with artificial intelligence has accumulated.

### 5.1From Rules to Adaptive Learning

Fraud detection in its traditional form relies on the rule-based methods examined in the preceding section, flagging the transactions that match the predefined patterns that the rules encode as indicative of fraud. These methods, long the mainstay of fraud detection, suffer from the limits common to the rule-based approach, namely their blindness to the novel fraud patterns that their rules do not encode, their rigidity in the face of the evolving tactics of the fraudsters, and their tendency to generate false positives that burden the institution and inconvenience its customers. Fraudsters, moreover, learn the rules through their interactions with the system and adapt their tactics to evade them, so that the rule-based methods are perpetually behind the adversary they confront. The limits of the rule-based methods are particularly consequential in fraud detection, for the volume of the transactions is immense, the fraud patterns evolve continually, and the false positives that the methods generate impose a direct cost in the legitimate transactions they impede and the customers they inconvenience.

Artificial intelligence transforms fraud detection by replacing the application of predefined rules with the learning of fraud patterns from the data. The machine-learning methods of fraud detection analyze the historical transaction data to establish patterns of legitimate and fraudulent activity, and they apply the learned patterns to the identification of fraud among the new transactions, detecting the anomalies and suspicious patterns that indicate fraud. The methods are adaptive, learning from the new data as it arrives and adjusting to the evolving tactics of the fraudsters, and they are capable of detecting the novel fraud patterns that no predefined rule encodes, identifying the fraud by its deviation from the learned patterns of the legitimate activity rather than by its match to a predefined pattern of fraud. The methods operate in real time, analyzing the transactions as they occur and detecting the fraud quickly enough to prevent or to limit the harm, and they are capable of incorporating a far larger and more granular set of features than the rule-based methods could, drawing on the transactional, the behavioral, and the contextual data to improve the discrimination of the fraud from the legitimate activity. This transition, from the application of the predefined rules to the learning of the patterns from the data, is the central development in the domain and the source of the effect that the evidence documents.

### 5.2 Evidence of Effect on Detection and False Positives

The evidence for the effect of artificial intelligence on fraud detection is the strongest and the most consistent of the four domains, supported across the peer-reviewed, industry, and institutional sources, and converging on the conclusion that the adaptive-learning methods materially outperform the rule-based methods they displace on the dimensions of detection accuracy, false-positive reduction, and real-time operation. The peer-reviewed evidence indicates that the machine-learning methods detect fraud more accurately than the rule-based methods, identifying a higher proportion of the actual fraud while flagging a lower proportion of the legitimate activity, and that the methods improve continually as they learn from the new data. The industry evidence indicates that a large and rising majority of financial institutions have adopted the technology for detection for fraud, that the adoption reflects the institutions' experience of its effect, and that the institutions that have adopted it report the reductions in fraud losses and in false positives that motivated the adoption.

The quantified evidence, where it is available and of sufficient quality to credit, indicates effects of substantial magnitude. The industry evidence reports that a substantial fraction of the issuers and the acquirers in a major payment network have saved sums measured in the millions of dollars in fraud over a period of two years through the application of artificial intelligence, that the consumers themselves increasingly expect

and trust the technology to protect them, and that the institutions that have deployed it have improved the detection of the fraud while reducing the false positives that burden their operations and inconvenience their customers. The reduction of the false positives is itself a substantial benefit, for the false positives of the rule-based methods impose a direct cost in the legitimate transactions they impede, the customers they inconvenience, and the analyst capacity they consume, and the evidence indicates that the artificial-intelligence methods reduce the false positives materially even as they improve the detection of the actual fraud, achieving the improvement on both dimensions simultaneously that the rule-based methods, which traded the one against the other, could not. The evidence on fraud detection is accordingly assessed as supporting a substantial and favorable effect, according to high confidence, the strongest finding of the synthesis.

The confidence of this finding is tempered, but not overturned, by several considerations. The quantified figures derive in part from the institutions and the vendors with an interest in the favorable result, and the appraisal discounts them accordingly, but the convergence of the evidence across the independent peer-reviewed sources and the institutional experience supports the finding even after the discount. The effect is conditioned by the quality and the breadth of the data on which the methods depend, for the methods learn the fraud patterns from the data, and the evidence indicates that their effectiveness is directly proportional to the quality and the breadth of the data they can access, a condition that the institutions with the richest data are best positioned to meet. And the effect is subject to the adversarial co-evolution examined in Section 8, for the fraudsters adapt to the artificial-intelligence methods as they adapted to the rule-based methods, and the fraudsters have themselves begun to deploy artificial intelligence, so that the favorable effect documented in the evidence is realized in a contest that continues to evolve. These considerations condition the finding but do not overturn it, for the evidence for the substantial and favorable effect of artificial intelligence on fraud detection is the strongest and the most consistent of the synthesis.

### 5.3 Case Evidence and Synthesis

The case evidence in fraud detection is rich and well documented, centering on the experience of the largest banks and the major payment networks, which have deployed the technology at the greatest scale and over the longest period. The largest bank in the United States has reported that its deployment of artificial intelligence across its fraud and cyber functions has held its fraud costs approximately flat even as the volume of the attacks against it has risen, an outcome that the institution attributes to its artificial-intelligence-enabled detection and that, if credited, represents a substantial effect, for the holding of the fraud costs flat against a rising volume of attacks implies a material improvement in the rate at which the attacks are detected and prevented. The institution has integrated the technology into a comprehensive risk-management framework, anchored in the established cybersecurity frameworks, and has invested in the responsible governance of the technology, illustrating the conditions under which the effect is realized. The major payment networks have reported comparable effects, with the documented evidence indicating that the application of artificial intelligence to the analysis of the transactional data across their networks has improved the detection of fraud and reduced the account-takeover and synthetic-identity fraud that the networks confront.

The synthesis of the fraud-detection domain arrives at a judgment of substantial and favorable effect, accorded high confidence, qualified by the conditions and the counter currents that the synthesis identifies. The direction and the magnitude of the effect are supported across peer-reviewed, industry, and institutional evidence, converging on the conclusion that the adaptive-learning methods materially outperform the rule-

based methods on detection accuracy, false-positive reduction, and real-time operation, and the case evidence of the largest institutions grounds the conclusion in documented experience. The conditions of the effect are visible in the case evidence, which indicates that the effect is realized when the institution possesses the rich data on which the methods depend, integrates the technology into a comprehensive risk-management framework, and governs it responsibly, and that the largest institutions, best positioned to meet these conditions, realize the effect most fully. The counter currents are visible in the adversarial co-evolution, the dependence on data quality, and the new vulnerabilities that the technology introduces, examined in Section 8. The fraud-detection domain thus provides the clearest evidence of the synthesis that artificial intelligence improves cyber risk management substantially, while illustrating that the improvement is conditioned by the data, the integration, and the governance that surround the technology and is realized in a contest that continues to evolve.

### 5.4 The Techniques of Artificial-Intelligence Fraud Detection

The favorable effect of artificial intelligence on fraud detection, established in the preceding synthesis, is realized through a range of techniques whose characterization deepens the understanding of the effect and its conditions. The techniques span the principal families of machine learning, and the evidence indicates that the institutions deploy them in combination rather than in isolation, assembling the complementary techniques into the systems that achieve the effect. The supervised-learning techniques, trained on the labelled data of the known fraud and the known legitimate activity, learn to classify the new activity by its resemblance to the labelled examples, and they are effective against the fraud patterns that resemble the known examples on which they were trained. The unsupervised-learning techniques, which do not require labelled data, detect the fraud by its deviation from the learned patterns of the normal activity, and they are effective against the novel fraud patterns that the supervised techniques, trained only on the known examples, would miss. The combination of the supervised and the unsupervised techniques, which the evidence indicates the sophisticated institutions employ, achieves the detection of both the known and the novel fraud that neither technique alone would achieve.

**Table 3:** Principal Techniques of Artificial-Intelligence Fraud Detection

| Technique Family | Mechanism | Principal Strength and Limit |
|---|---|---|
| **Supervised learning** | Learns to classify activity from labelled examples of fraud and legitimate activity | Strong on known patterns; limited against novel fraud absent from training data |
| **Unsupervised learning** | Detects fraud as a deviation from learned patterns of normal activity | Detects novel fraud; higher false-positive tendency requiring tuning |
| **Deep learning** | Learns complex, high-dimensional patterns through layered neural networks | High accuracy on complex patterns; greater opacity and data demand |
| **Ensemble methods** | Combines multiple models to improve accuracy and robustness | Improved accuracy and stability; added complexity in validation |
| **Behavioral and graph analytics** | Models' normal behavior and relationships to detect anomalies and coordinated fraud | Strong on account takeover and coordinated rings; depends on rich relational data |

Table 3: The principal technique families of artificial-intelligence fraud detection, their mechanisms, and their characteristic strengths and limits. The evidence indicates that institutions combine the families, assembling complementary techniques into systems that achieve the documented effect.

The characterization of the techniques deepens the understanding of the conditions on which the effect depends. The dependence on the data, identified in the synthesis, is visible in the techniques, for the supervised techniques require the labelled data of the known fraud, the unsupervised techniques require the representative data of the normal activity, the deep-learning techniques require the large volume of the data that their complexity demands, and the behavioral and graph techniques require the rich relational data of the behavior and the relationships, so that the effect of each technique is conditioned by the availability and the quality of the data it requires. The combination of the techniques, which the sophisticated institutions employ, itself requires the data and the capability that the combination demands, and it is the institutions with the richest data and the greatest capability, the largest banks and the major payment networks, that deploy the most sophisticated combinations and realize the effect most fully, consistent with the synthesis finding that the effect is conditioned by the data and the capability that surround the technology.

### 5.5 The Explainability Requirement in Fraud Detection

The deployment of artificial intelligence in fraud detection in the regulated banking context confronts a requirement that bears on both the effect of the technology and the governance burden it imposes, namely the requirement that the institution be able to explain the decisions of the fraud-detection systems. The requirement arises from several sources. It arises from the regulatory expectations, which require the institution to understand and to justify the decisions of the systems it deploys, particularly where the decisions affect the consumers, as the decision to decline a transaction or to flag an account does. It arises from the operational necessity that the human analysts who adjudicate the flags that the systems raise require the explanation of the flags to investigate them efficiently. And it arises from the customer relationship, for the institution that declines a transaction or flags an account must be able to explain the decision to the affected customer. The explainability requirement is, in consequence, not an optional enhancement but a condition of the deployment of the technology in the regulated context.

The explainability requirement bears on the effect of the technology in a manner that the synthesis must weigh. On the one hand, the requirement constrains the deployment of the most opaque and the most powerful techniques, for the institution that cannot explain the decisions of a technique may be unable to deploy it in the regulated context, regardless of its accuracy, so that the explainability requirement may limit the institution to the less opaque and the less powerful techniques and thereby constrain the effect. On the other hand, the explainability methods that render the techniques interpretable, examined in the literature on explainable artificial intelligence in cybersecurity, mitigate this constraint, permitting the institution to deploy the powerful techniques while satisfying the explainability requirement, though, as Section 8 examines, the explainability methods introduce their own vulnerability to the adversarial exploitation of the explanations they produce. The explainability requirement thus figures in the assessment of the effect of the technology, both as a constraint that may limit the deployment of the most powerful techniques and as a governance burden that the deployment of the technology imposes, and it illustrates the synthesis finding that the favorable effects of the technology are realized only through the discharge of the governance burden that the regulated context requires.

### 5.6 The Financial Technology Dimension

The synthesis of the fraud-detection domain has centered on the largest banks and the major payment networks, which deploy the technology at the greatest scale, but the financial-technology firms that operate within and alongside the regulated banking system warrant separate attention, both because they have been among the most aggressive adopters of the technology and because their experience illustrates the conditions and the limits of the effect in a different institutional context. The financial-technology firms, frequently built from the outset around the data and the technology that the established institutions have had to adopt, have deployed artificial intelligence in fraud detection as a core capability rather than an addition to the prior methods, and the evidence indicates that they have achieved the favorable effects that the synthesis documents, detecting the fraud more accurately and reducing the false positives relative to the rule-based methods. The financial-technology experience illustrates that the effect of the technology is not confined to the largest established institutions but extends to the institutions that deploy it as a core capability, provided they possess the data and the capability on which the effect depends.

The financial-technology dimension also illustrates the conditions and the limits of the effect in a manner that complements the experience of the largest institutions. The financial-technology firms, frequently possessing the rich data and the technical capability that the effect requires but lacking the scale and the established governance of the largest banks, illustrate both the realization of the effect where the data and the capability are present and the importance of the governance that the effect requires, for the evidence indicates that the financial-technology firms that have deployed the technology without the commensurate governance have encountered the limits and the failures that the ungoverned technology can produce, including the false positives and the bias that the synthesis identifies. The financial-technology dimension thus extends the synthesis beyond the largest institutions, illustrating that the effect of the technology depends on the data, the capability, and the governance that surround it rather than on the scale or the type of the institution, and that the institution of any type that possesses the data and the capability and discharges the governance burden realizes the effect, while the institution that lacks them does not.

## 6. Security Operations Automation

The third focus domain is security operations automation, the application of artificial intelligence to the work of the security operations center, the function in which the institution monitors its systems for signs of compromise, triages the alerts that the monitoring generates, investigates the alerts that warrant investigation, and responds to the incidents that the investigation confirms. This section examines the automated and augmented security operations center, the evidence for the effect of artificial intelligence on the speed of detection and response, and the case evidence that grounds the synthesis. The domain is where the operational pressures of alert overload and the skills gap, examined in Section 3, bear most directly, and it is where the effect of artificial intelligence on the speed of detection and response, as distinct from its accuracy, is most clearly documented.

### 6.1 The Automated and Augmented Security Operations Centre

The security operations center in its traditional form is a labor-intensive function in which human analysts monitor the flow of alerts that the institution's security systems generate, triage the alerts to distinguish the consequential from the inconsequential, investigate the alerts that warrant investigation, and coordinate the response to the incidents that the investigation confirms. The function is constrained by the operational

pressures examined in Section 3, namely the alert overload that inundates the analysts with a volume of alerts that exceeds their capacity to investigate, and the skills gap that leaves the function understaffed, and these constraints degrade both the speed and the completeness of the detection and the response that the function provides. The security operations center is, in consequence, the domain in which the augmentation of the scarce human capacity by artificial intelligence offers the most direct operational benefit, for the technology can absorb the routine and high-volume work that consumes the analysts' capacity and can compress the time that the detection and the response require.

Artificial intelligence is applied in the security operations center across a spectrum from the augmentation of human analysts to the automation of the functions they perform. At the augmentation end of the spectrum, technology assists the human analysts, triggering the alerts to surface the consequential ones, enriching the alerts with the context that the investigation requires, and generating the summaries and the analyses that accelerate the human work, while leaving the consequential decisions to the human analysts. At the automation end of the spectrum, the technology performs the functions of the security operations center with limited human intervention, triaging and investigating the alerts, and, in the most advanced deployments, the agentic systems that have begun to appear, reasoning about the alerts and orchestrating the response with a degree of autonomy that approaches that of the human analyst. The spectrum reflects both the maturation of the technology, which has progressed from the augmentation toward automation as its capability has grown, and the considered choices of the institutions, which calibrate the degree of automation to their assessment of the technology's reliability and the consequences of its errors. Across the spectrum, the role of technology is to relieve the operational pressures of alert overload and the skills gap, absorbing the routine work, triggering the flood of alerts, and multiplying the effective capacity of the scarce human analysts.

### 6.2 Evidence of Effect on Detection and Response Times

The evidence for the effect of artificial intelligence on security operations is strong, particularly on the dimension of the speed of detection and response, and it is supported across the industry surveys, the institutional reports, and the vendor evidence, though the last is discounted in the appraisal. The most substantial and credible evidence concerns the compression of the time that the detection and containment of an incident require. The established industry research indicates that the organizations that deploy artificial intelligence and automation extensively in their security operations achieve a materially shorter detection-and-containment lifecycle than the organizations that do not, with the reported difference amounting to a substantial reduction in the period between the compromise and its containment, a reduction whose value is considerable given the evidence that the harm of an incident grows with the time it remains uncontained.

**Table 1:** Reported Effects of Artificial Intelligence on Security Operations Metrics

| Metric | Reported Effect of Artificial Intelligence | Source Type and Appraisal |
|---|---|---|
| **Breach detection-and-containment lifecycle** | Organizations using extensive artificial intelligence and automation reported a substantially shorter lifecycle than those not doing so (on the order of an 80-day reduction against a multi-month average) | Established industry research; high credibility; self-reported organizational data |

| | | |
|---|---|---|
| **Mean time to respond** | Adopters of agentic security operations tooling reported approximately 50 percent faster meantime to respond | Vendor evidence; discounted; directionally consistent with other sources |
| **Detection accuracy and false positives** | Reported improvements in detection accuracy and reductions in false-positive alerts | Vendor evidence; discounted; consistent with peer-reviewed direction |
| **Analyst time on routine analysis** | Reported large reductions in the share of analyst time consumed by routine threat analysis | Vendor and industry evidence; directionally credible |
| **24/7 coverage versus manual reporting** | A large majority of security operations centers operate continuously, yet a majority still rely on manual reporting, indicating incomplete automation | Independent security operations survey; high credibility |

Table 1: Reported effects of artificial intelligence on security operations metrics, with the type and the appraisal of the source. The effects on the speed of detection and response are supported by credible independent evidence; the more specific quantified claims derive substantially from vendor sources and are discounted accordingly, though they are directionally consistent with the independent evidence.

The independent survey evidence complicates and enriches the picture that the headline figures present. The survey of security operations indicates that, while a large majority of the security operations centers now operate continuously, a majority still rely on the manual reporting that the automation is meant to displace, indicating that the automation of the security operations function, though substantial, remains incomplete, and that the gap between the coverage that the institutions have achieved and the operational efficiency that full automation would provide remains open. This evidence is valuable precisely because it complicates the narrative of comprehensive transformation that the vendor evidence presents, indicating that the effect of artificial intelligence on security operations, though real and substantial in the dimension of the speed of detection and response, is uneven in its realization and incomplete in its penetration of the function. The evidence on security operations is accordingly assessed as supporting a strong but qualified effect, accorded moderate-to-high confidence, with the effect on the speed of detection and response well supported and the completeness of the transformation more limited than the headline claims suggest.

The qualification of the effect extends to several conditions and limits that the appraisal identifies. The effect on the speed of detection and response is established more firmly than the effect on the accuracy of detection, for the time-based metrics are more readily measured and more consistently reported than the accuracy metrics, and the appraisal accords correspondingly higher confidence to the finding of compressed detection and response times than to the finding of improved detection accuracy. The effect is conditioned by the degree and the maturity of the automation that the institution has achieved, for the survey evidence indicates that the automation remains incomplete at many institutions and that the realization of the effect depends on the institution's progression along the spectrum from augmentation to automation. The effect introduces new dependencies and failure modes, examined in Section 8, for the automation of the security operations function creates a reliance on the artificial-intelligence systems whose own failures, whether through error or through adversarial manipulation, become a new source of risk. The security-operations domain thus provides strong evidence of a favorable effect on the speed of detection and response, qualified by the incompleteness of the transformation, the firmer support for the speed than the accuracy effect, and the new dependencies that the automation introduces.

### 6.3 Case Evidence and Synthesis

The case evidence in security operations centers on the documented experience of the institutions and the providers that have deployed the automated and augmented security operations center, and it illustrates both the effect and its conditions. The documented deployments of the agentic security operations tooling, in which the artificial-intelligence systems reason about the alerts and orchestrate the response with a degree of autonomy, report substantial compression of the mean time to respond and substantial relief of the analyst burnout that the alert overload and the skills gap produce, though the evidence derives substantially from the providers of the tooling and is discounted accordingly. The documented experience of the institutions that have deployed the technology indicates that the automation of the routine work of the security operations function has relieved the operational pressures that degraded the function, absorbing the high-volume triage that consumed the analyst capacity and freeing the scarce human analysts for the consequential work that requires their judgment. The independent evidence of the incomplete penetration of automation, in the persistence of the manual reporting at most of the security operations centers, grounds the synthesis in the realistic recognition that the transformation, though substantial, remains in progress.

The synthesis of the security-operations domain arrives at a judgment of strong but qualified effect, according to moderate-to-high confidence. The effect on the speed of detection and response is well supported by the credible independent evidence of the compressed detection-and-containment lifecycle and the consistent, if discounted, evidence of the compressed response times, and it is substantial in its magnitude and considerable in its value, given the growth of the harm with the time an incident remains uncontained. The qualification arises from the firmer support for the speed than the accuracy effect, the incompleteness of the transformation that the independent survey evidence reveals, and the new dependencies and failure modes that the automation introduces. The conditions of the effect are visible in the dependence on the maturity of the automation, and the integration of the technology with the human analysts, for the evidence indicates that the effect is realized most fully where the institution has progressed furthest along the spectrum from augmentation to automation and has integrated the technology effectively with the human judgment that the consequential decisions require. The security-operations domain thus provides strong evidence of a favorable effect on the speed of detection and response, the dimension on which the operational pressures bear most directly, qualified by the incompleteness of the transformation and the new risks that automation introduces.

### 6.4 The Agentic Turn and the Question of Autonomy

The most recent development in the automation of security operations, and the one whose assessment is the most provisional, is the emergence of the agentic systems that reason about the alerts and orchestrate the response with a degree of autonomy that approaches that of the human analyst. The agentic systems represent a development beyond the augmentation and the rule-based automation that preceded them, for they do not merely assist the human analyst or execute the predefined playbooks but reason about the alerts, connect the patterns across the disconnected data, and determine the response dynamically, adapting to the context as the human analyst would. The evidence on the agentic systems is the most recent and the most provisional of the security-operations domain, for the systems have been deployed only recently and the evidence for their effect is still forming, but the early evidence indicates that they compress the response times and relieve the analyst burden substantially, and the trajectory of the development suggests a future in which the agentic systems assume an increasing share of the security-operations function.

The agentic turn raises the question of autonomy that the deployment of artificial intelligence in security operations has always implied but that the agentic systems sharpen, namely the question of how much of the consequential decision-making the institution should delegate to the systems and how much it should reserve to the human analysts. The question is consequential, for the delegation of the decision-making to the systems realizes the efficiency that the automation promises but introduces the dependence on the systems whose failures, whether through error or through adversarial manipulation, become a source of the risk, while the reservation of the decision-making to the human analysts preserves the human judgment and the oversight that the consequential decisions require but forgoes the efficiency that the automation offers. The evidence indicates that the institutions navigate the question by calibrating the autonomy to the consequence of the decision and the reliability of the systems, reserving the most consequential decisions to the human analysts while delegating the routine and the high-volume decisions to the systems, and that the prudent deployment preserves the human oversight that the regulated context requires even as it realizes the efficiency that the automation offers. The agentic turn thus sharpens the question of autonomy without resolving it, and the assessment of its effect, the most provisional of the security-operations domain, awaits the evidence that the recency of the development has not yet permitted to accumulate.

### 6.5 The Human-in-the-Loop and the Integration Question

## 7. Cyber Resilience

The fourth focus domain is cyber resilience, the capacity of the institution to anticipate, withstand, recover from, and adapt to the cyber incidents that its defenses do not prevent, and it is the domain in which the effect of artificial intelligence is the least mature and the least directly evidenced, for resilience is a system-level outcome to which the detection and the response examined in the preceding domains contribute but which they do not exhaust. This section examines the relationship between detection and the response of the preceding domains and the resilience that is the subject of this one, the evidence for the effect of artificial intelligence on the resilience outcomes, and the case evidence that grounds the synthesis. The domain is treated last because resilience is, in a sense, the cumulative outcome to which the other domains contribute, and because the regulatory attention to resilience, which has intensified in recent years, frames the institutional pursuit of it.

### 7.1 From Detection to Resilience

Cyber resilience is a broader and more systemic concept than the detection and the response examined in the preceding domains, encompassing not only the identification and the containment of the incidents but the capacity of the institution to continue its critical operations through an incident, to recover its systems and its data after an incident, and to adapt its defenses in the light of the incidents it experiences. The regulatory usage, which has framed the institution's pursuit of resilience, emphasizes this systemic and operational character, directing the institutions to establish the realistic recovery objectives, the continuity arrangements, and the adaptive capacity that resilience requires, and shifting the focus of the supervisory attention from the prevention of incidents, which cannot be guaranteed, to the resilience that limits their consequences. The detection and the response examined in the preceding domains contribute to the resilience, for the faster and more accurate detection and the faster and more effective response limit the consequences of the incidents and thereby contribute to the resilience, but they do not exhaust it, for the resilience encompasses the continuity, the recovery, and the adaptation that extend beyond the detection and the response.

The contribution of artificial intelligence to cyber resilience is, in consequence, in part the cumulative effect of its contributions to the detection and the response examined in the preceding domains, and in part a more direct contribution to the continuity, the recovery, and the adaptation that resilience additionally requires. Cumulative contribution is better evidenced, for it follows from the contributions to the detection and the response that the preceding domains documented, and the faster and more accurate detection and the faster and more effective response that artificial intelligence provides limit the consequences of the incidents and thereby contribute to the resilience. The more direct contribution, to the continuity, the recovery, and the adaptation, is the less evidenced, for the application of artificial intelligence to these functions is more recent and the evidence for its effect is less developed, though the technology is increasingly applied to the prediction of the failures that threaten continuity, the automation of the recovery that follows an incident, and the analysis of the incidents that informs the adaptation of the defenses. The effect of artificial intelligence on resilience is thus assessed in part through its documented effects on the detection and the response that contribute to resilience and in part through the less developed evidence for its more direct contributions to the continuity, recovery, and adaptation.

### 7.2 Evidence of Effect on Resilience Outcomes

The evidence for the effect of artificial intelligence on the resilience outcomes, as distinct from the detection and the response that contribute to them, is the least mature of the four domains, reflecting both the recency of the application of the resilience functions and the difficulty of measuring the resilience outcomes directly. The resilience outcomes, namely the continuity maintained through the incidents, the recovery achieved after them, and the adaptation accomplished in their light, are difficult to observe and to measure, for they manifest in the incidents that the institution withstands and recovers from, which are themselves difficult to observe, and in the counterfactual harms that the resilience averts, which are not observable at all. The evidence that does bear on the question is, in consequence, more indirect and more limited than the evidence in the other domains, consisting in part of the documented contributions of artificial intelligence to the detection and the response that contribute to resilience, in part of the emerging evidence for its more direct contributions to the continuity, the recovery, and the adaptation, and in part of the regulatory and the institutional attention to resilience that frames the pursuit of it.

The regulatory evidence is particularly relevant to the resilience domain, for the regulators have intensified their attention to the cyber resilience of financial institutions, issuing the guidance that directs the institutions to establish the recovery objectives, the continuity arrangements, and the adaptive capacity that resilience requires, and conducting the examinations that assess the institutions' resilience. This regulatory attention frames the institution's pursuit of resilience and establishes the standards against which it is assessed, and it indicates the importance that the supervisors attach to the resilience outcomes, though it does not by itself establish the effect of artificial intelligence on those outcomes. The regulatory reports indicate that the institutions are increasingly applying artificial intelligence to the resilience functions, and that the supervisors are attentive both to the benefits that technology offers and to the risks that it introduces, including the new dependencies that the reliance on technology creates. The evidence on resilience is accordingly assessed as supporting a favorable but immature and indirectly evidenced effect, accorded lower confidence than the findings in the other domains, reflecting the recency of the application and the difficulty of the measurement.

### 7.3 Case Evidence and Synthesis

The case evidence in cyber resilience centers is on the documented experience of the largest institutions and on the regulatory framing of the resilience pursuit. The largest institutions have reported that their deployment of artificial intelligence contributes to their operational resilience, integrating the technology into the comprehensive risk-management frameworks that anchor their resilience and applying it to the prediction, the recovery, and the adaptation that resilience requires, though the documented evidence for the specific effect of the technology on the resilience outcomes, as distinct from the detection and the response, is more limited than in the other domains. The participation of a major bank in a cross-industry initiative directed specifically at the artificial-intelligence-enabled cyber risks and the resilience to them illustrates both the institution's attention to the intersection of artificial intelligence and resilience and the recognition that the technology is at once a contributor to the resilience and a source of the new risks that the resilience must encompass. The regulatory reports and the supervisory guidance ground the synthesis in the framework of standards and expectations within which the institutions pursue resilience.

The synthesis of the cyber-resilience domain arrives at a judgment of favorable but immature effect, accorded to lower confidence than the other domains. The direction of the effect is favorable, for the contributions of artificial intelligence to the detection and the response examined in the preceding domains contribute to the resilience, and the emerging applications of the technology to the continuity, the recovery, and the adaptation extend the contribution, but the magnitude of the effect on the resilience outcomes, as distinct from the detection and the response, is the least well-established of the synthesis, reflecting the recency of the application and the difficulty of the measurement. The conditions of the effect are visible in the integration of the technology into the comprehensive risk-management and resilience frameworks of the institutions and in the regulatory standards that frame the resilience pursuit. The countercurrent is particularly salient in the resilience domain, for the reliance on the artificial-intelligence systems that contribute to the resilience itself introduces a dependence whose failure would undermine the resilience, so that the technology is at once a contributor to the resilience and a source of the risk that the resilience must encompass. The cyber-resilience domain thus provides the least mature evidence of synthesis, supporting a favorable direction of effect while indicating that the magnitude and the conditions of the effect on the resilience outcomes remain the least well established and the most in need of the primary research that Section 10 proposes.

The synthesis of the security-operations domain has identified human-machine integration as a condition of the effect, and the integration warrants the deeper treatment that its importance justifies, for the evidence indicates that the effect of the technology depends as much on the integration of the technology with the human analysts as on the capability of the technology itself. The integration question concerns how the technology and the human analysts are combined into the security operations function, and the evidence indicates that the combination that realizes the effect is neither the replacement of the human analysts by the technology nor the mere addition of the technology to the unchanged human function, but the redesign of the function that combines the complementary strengths of the technology and the human analysts. Technology contributes the capacity to process the high volume of data and the alerts at a speed that human analysts cannot match, and the human analysts contribute the judgment, the context, and the oversight that the technology cannot supply, and the function that combines them effectively realizes the effect that neither alone would achieve.

The integration question bears on the assessment of the effect in a manner that the synthesis must weigh, for it establishes that the effect of the technology is not a property of the technology alone but of the technology as it is integrated with the human analysts, and that the same technology may realize the effect

in the institution that integrates it effectively and fail to realize it in the institution that does not. The evidence indicates that the institutions that have realized the effect most fully have redesigned their security operations functions to combine the technology and the human analysts effectively, while the institutions that have deployed the technology without the redesign, expecting the technology to realize the effect regardless of the integration, have realized the effect incompletely or not at all. The integration question thus reinforces the synthesis finding that the effect of the technology is conditioned by the human organization that surrounds it, and it cautions against the expectation that the deployment of the technology will realize the effect regardless of the integration, an expectation that the evidence indicates the institutions that hold it do not realize.

### 7.4 The Regulatory Framing of Resilience and the Role of Artificial Intelligence

The synthesis of the cyber-resilience domain has noted the regulatory framing of the resilience pursuit, and the framing warrants the deeper treatment that its importance to the domain justifies, for the regulatory attention to resilience has intensified in recent years and frames the institutional pursuit of resilience to which artificial intelligence contributes. The regulators have shifted the focus of their attention from the prevention of incidents, which cannot be guaranteed, to the resilience that limits their consequences, directing the institutions to establish the realistic recovery objectives, the continuity arrangements, and the adaptive capacity that resilience requires, and conducting the examinations that assess the institutions' resilience against these standards. The regulatory framing establishes the standards against which the institution's resilience is assessed and the expectations that the institution's pursuit of resilience must meet, and it frames the contribution of artificial intelligence to the resilience as a contribution to the meeting of the standards and the expectations that the regulators have established.

The regulatory framing also bears on the assessment of the effect of artificial intelligence on resilience in a manner that the synthesis must weigh, for the regulators are attentive both to the benefits that the technology offers to resilience and to the risks that it introduces, including the new dependence that the reliance on the technology creates. The regulatory reports indicate that the institutions are increasingly applying the technology to the resilience functions, and that the supervisors assess both the contribution of the technology to the resilience and the risk that the reliance on the technology introduces, recognizing that the technology is at once a contributor to the resilience and a source of the new risk that the resilience must encompass. Regulatory framing thus reinforces the synthesis finding that the effect of artificial intelligence on resilience is favorable but bounded by the new dependence that the reliance on technology creates, and it establishes the standards and expectations within which the institution pursues resilience and against which the contribution of the technology is assessed. The deepening regulatory attention to resilience and to the role of artificial intelligence within it, ensures that the assessment of the technology's effect on resilience will remain a matter of supervisory as well as scholarly concern and that the institution's deployment of the technology in the resilience functions will be conducted under the supervisory attention that the regulatory framing establishes.

## 8. The Adversarial Countercurrent and the Limits of Artificial Intelligence

The domain syntheses of the preceding sections have identified, alongside the favorable effects of artificial intelligence on the detection and the response across the four domains, a consistent countercurrent that qualifies the favorable effects and that must be weighed in any honest assessment of the technology's net effect. This section examines the countercurrent directly, treating the offensive use of artificial intelligence

by the adversary, the attacks on the defensive systems themselves, and the opacity, the false positives, and the governance burden that technology introduces. The examination is essential to the balance of the synthesis, for an assessment that attended only to the favorable effects and ignored the countercurrent would overstate the net effect of the technology and would mislead the institutions, the regulators, and the scholars who rely on the assessment. The countercurrent does not negate the favorable effects, which the evidence supports, but it conditions them and bounds them, and it establishes that the net effect of the technology on cyber resilience depends on a co-evolution whose trajectory is not yet settled.

### 8.1 Offensive Artificial Intelligence and the Weaponization of Technology

The first element of the countercurrent is the use of artificial intelligence by the adversary, for the same technology that improves the defense is available to the attacker, who deploys it to improve the offense. The evidence indicates that the adversaries have adopted artificial intelligence across the range of their operations, employing it to generate the convincing fraudulent communications that the social-engineering attacks require, to produce the synthetic media that the impersonation and the fraud increasingly employ, to discover the vulnerabilities that the attacks exploit, and to conduct the attacks at a scale and a speed that the manual methods could not achieve. The emergence of the dedicated offensive artificial-intelligence tools, marketed to adversaries for the express purpose of fraud and attack, illustrates the weaponization of the technology, and the evidence indicates that a substantial and rising fraction of the detected fraud now involves artificial intelligence on the part of the attacker.

The offensive use of artificial intelligence has several implications for the assessment of technology's net effect. The first is that it raises the baseline of the threat against which the defensive use of the technology must be measured, for the defensive improvements that the preceding sections documented are realized against an adversary that is itself improving through its own use of the technology, so that the net effect of the technology on the contest depends on the relative rates of the improvement of the offense and the defense. The second is that it accelerates adversarial co-evolution, for the adversary that deploys artificial intelligence adapts more quickly to the defenses, compressing the period over which a defensive improvement remains effective before the adversary adapts to it. The third is that it raises the stakes of the contest, for the artificial-intelligence-enabled attacks are more convincing, more scalable, and more difficult to detect than the manual attacks they augment or replace, so that the consequences of a defensive failure grow as the offense improves. The offensive use of artificial intelligence thus establishes that the favorable defensive effects documented in the preceding sections are realized in a contest in which the adversary is also improving, and that the net effect of the technology on the contest depends on the relative trajectories of the offense and the defense, which the evidence does not permit one to project with confidence.

### 8.2 Attacks on the Defensive Systems Themselves

The second element of the countercurrent is the vulnerability of the defensive artificial-intelligence systems themselves to attack, for the systems that detect the threats and the frauds are themselves a target, and the adversary that compromises or evades them undermines the defense they provide. The peer-reviewed literature on adversarial machine learning has documented a range of attacks on the defensive systems, including the evasion attacks that craft the inputs to escape detection, the data-poisoning attacks that corrupt the data on which the systems learn, and the model-extraction and inference attacks that compromise the confidentiality of the systems and the data. These attacks exploit the fundamental properties of the machine-learning systems, the dependence on the data, and the susceptibility to the crafted input, and they are not

incidental vulnerabilities that better engineering would eliminate but structural features of the technology that the defense must manage.

Of particular concern is the documented vulnerability of the systems to adversarial manipulation that exploits the very explainability that the regulated context requires. The research has demonstrated that the explanations which render the systems interpretable, and which the governance of the systems requires, can themselves be exploited by the adversary, who uses the explanation of the system to identify the features on which the detection depends and to craft the inputs that evade it, an attack that the literature has termed adversarial explainability and that has been demonstrated to bypass the detection systems with a high rate of success. This vulnerability illustrates a tension at the heart of the deployment of artificial intelligence in the regulated banking context, for the explainability that governance requires is at once a benefit, enabling the validation and the oversight of the systems, and a vulnerability, providing the adversary with the information to evade them. The attacks on the defensive systems establish that the deployment of artificial intelligence introduces a new attack surface, the systems themselves, and that the defense must encompass not only the threats that the systems detect but the attacks on the systems that detect them, a burden that the prior methods, which lacked the learned and data-dependent character of the artificial-intelligence systems, did not impose to the same degree.

### 8.3 Opacity, False Positives, and the Governance Burden

The third element of the countercurrent is the burden that the deployment of artificial intelligence imposes through its opacity, its false positives, and the governance that the regulated context requires. The opacity of the artificial-intelligence systems, the difficulty of articulating the basis of their decisions, complicates the validation, the explanation, and the oversight that the regulated banking context requires, for the institution must be able to understand and to justify the decisions of the systems it deploys, and the opacity of the systems renders this understanding and justification difficult. The literature on explainable artificial intelligence in cybersecurity has developed methods that render the systems interpretable, and these methods mitigate the opacity, but they do not eliminate it, and they introduce, as the preceding subsection noted, their own vulnerability to the adversarial exploitation of the explanations they produce. Opacity thus imposes a governance burden, the work of validating, explaining, and overseeing the opaque systems, that the deployment of the technology requires and that the prior, transparent methods did not impose to the same degree.

The false positives of the artificial-intelligence systems, though reduced relative to the rule-based methods as the fraud-detection synthesis noted, remain a substantial burden, for the systems, however accurate, flag some benign activity as malicious, and the false positives impose the costs of the legitimate activity they impede, the customers they inconvenience, and the analyst capacity they consume. The evidence indicates that the false positives remain a substantial challenge even in the artificial-intelligence systems, and that the systems, when poorly tuned or trained on unrepresentative data, can generate the false positives at a rate that degrades their value, and can do so unevenly across the populations they affect, raising the concern of the bias that the systems can introduce. The governance burden encompasses the management of these false positives and the bias they can reflect, the work of tuning and validating the systems to control the false positives, and ensuring their fairness, which the deployment of technology requires. The countercurrent of the opacity, the false positives, and the governance burden thus establishes that the favorable effects of the technology are realized only through the discharge of a substantial governance burden, and that the institution that deploys the technology without discharging the burden realizes the favorable effects

incompletely or not at all, and may introduce the new harms, of the unjustified decisions and the uneven treatment, that the ungoverned technology can produce.

### 8.4 The Dual-Use Character of Technology and the Governance Response

The elements of the countercurrent examined in the preceding subsections converge on a feature of the technology that warrants explicit articulation, namely its dual-use character, the property by which the same technology serves both the defender and the attacker, and that property carries implications for the governance of the technology that the synthesis must weigh. The dual-use character is not incidental but fundamental, for the artificial intelligence that learns the patterns of the threats to detect them is the same artificial intelligence that learns the patterns of the defenses to evade them, and the technology that improves the defense is, by its nature, available to improve the offense. The dual-use character establishes that the deployment of the technology by the defender does not confer a durable advantage, for the attacker deploys the same technology, and that the contest between the defender and the attacker is a contest between the parties wielding the same technology rather than a contest in which the technology favors one party over the other.

The dual character carries implications for the governance of technology that extends beyond the individual institution to the policy that frames the technology development and diffusion. The governance response at the level of the institution, examined in the discussion of Section 9, encompasses the data, the validation, the security, the human oversight, and the fairness of the systems, but the dual-use character raises governance questions that exceed the individual institution, including the questions of how the diffusion of the offensive capability might be constrained, how the development of the technology might be directed toward the defensive rather than the offensive applications, and how the institutions and the authorities might cooperate to maintain the defensive advantage against the adversaries that wield the same technology. These questions exceed the scope of the present synthesis, which assesses the effect of the technology rather than the policy that should govern it, but the dual-use character that the synthesis identifies establishes that the governance of the technology cannot be confined to the individual institution and must extend to the policy that frames the technology development and diffusion, a dimension that the synthesis flags for the attention of the policymakers and the future research that the assessment of the technology requires.

## 9. Cross-Domain Synthesis and Discussion

The domain syntheses of Sections 4 through 7 and the examination of the countercurrent in Section 8 permit a cross-domain synthesis that integrates the findings into an overall answer to the research question. This section presents that synthesis, characterizing the extent of the effect of artificial intelligence on cyber risk detection and incident response across the domains, the variation of the effect across them, the conditions on which the effect depends, and the countercurrent that bounds it. The synthesis arrives at an overall judgment that is differentiated rather than uniform, conditional rather than absolute, and bounded by the countercurrent rather than unqualified, and it grounds that judgment in the appraised evidence of the preceding sections.

**Table 2:** Summary of the Cross-Domain Synthesis

| Domain | Direction and Magnitude of Effect | Confidence | Principal Conditions and Limits |
|---|---|---|---|

| | | | |
|---|---|---|---|
| **Threat intelligence** | Favorable; magnitude harder to isolate | Moderate | Mediated through functions informed, confounded, vendor-heavy evidence |
| **Fraud detection** | Substantial and favorable; the strongest finding | High | Depends on data quality and breadth; adversarial co-evolution |
| **Security operations automation** | Strong on speed of detection and response; qualified | Moderate to high | Incomplete transformation; speed better evidenced than accuracy; new dependencies |
| **Cyber resilience** | Favorable but immature and indirectly evidenced | Lower | System-level outcome hard to measure; recency of application; new dependence |
| **Adversarial countercurrent** | Offsets and bounds the favorable effects across all domains | Established in direction | Offensive artificial intelligence; attacks on systems; opacity and governance burden |

Table 2: Summary of the cross-domain synthesis, reporting for each domain the direction and the magnitude of the effect, the confidence the appraised evidence permits, and the principal conditions and limits, together with the adversarial countercurrent that offsets and bounds the favorable effects across the domains.

The cross-domain synthesis supports an overall answer to the research question that is differentiated across the domains. The effect of artificial intelligence on cyber risk detection and incident response is substantial and favorable, and most firmly established, in fraud detection, where the adaptive-learning methods materially outperform the rule-based methods on detection accuracy, false-positive reduction, and real-time operation, and where the case evidence of the largest institutions grounds the conclusion in documented experience. The effect is strong but qualified in security operations automation, where the technology has materially compressed the time required to detect and contain incidents and has relieved the operational pressures that degraded the function, while the transformation remains incomplete, and the speed effect is better evidenced than the accuracy effect. The effect is favorable but confounded in threat intelligence, where the direction is clear, but the magnitude is harder to isolate from the functions the threat intelligence informs. And the effect is favorable but immature in cyber resilience, where the contributions of the other domains carry through to the resilience, but the direct effect on the resilience outcomes is the least well-established. The overall effect is thus real, meaningful, and domain-dependent, neither the uniform transformation that the enthusiasm surrounding the technology suggests nor the negligible effect that the skepticism toward it might imply.

The synthesis further establishes that the effect, when it is obtained, is conditional rather than absolute. The conditions recur across the domains, and there are three. The first is the data, for the artificial-intelligence systems learn from the data, and their effect is conditioned by the quality and the breadth of the data they can access, so that the institutions with the richest data realize the effect most fully and the effect erodes where the data is poor or unrepresentative. The second is governance, for the favorable effects are realized only through the discharge of the governance burden that the opacity, the false positives, and the new vulnerabilities of the technology impose, and the institution that deploys the technology without the

governance realizes the effects incompletely and may introduce new harms. The third is the human-machine integration, for the effect is realized most fully where the technology is integrated effectively with the human judgment that the consequential decisions require, augmenting rather than replacing the human analysts and preserving the human oversight that the regulated context demands. These conditions establish that the effect of artificial intelligence is not a property of the technology alone but of the technology as it is embedded in the data, the governance, and the human organization of the institution, and that the same technology may produce a substantial effect in the institution that meets the conditions and a negligible or a negative effect in the institution that does not.

The synthesis establishes, finally, that the favorable effects are bound by the adversarial counter current, and that the net effect of the technology on cyber resilience depends on a co-evolution whose trajectory is not settled. The favorable effects are real, but they are realized against an adversary that is itself improving through its own use of technology; they are achieved by systems that are themselves a target of the attacks, and they are accompanied by the opacity, the false positives, and the governance burden that the technology introduces. The net effect of the technology on the contest between the defender and the attacker depends on the relative rates at which the offense and the defense improve, which the evidence does not permit one to project, and the question of whether artificial intelligence will, on balance and over time, favor the defender or the attacker remains open. What the synthesis establishes is that the technology improves the defensive functions to the degree and under the conditions the domain syntheses identified, that the improvement is real and meaningful, and that it is realized in a contest in which the adversary also wields the technology, so that the improvement of the defensive functions does not translate straightforwardly into a net improvement in the security of the institution, which depends on the co-evolution whose trajectory the evidence leaves unsettled.

### 9.1 The Dynamics of the Adversarial Co-Evolution

The cross-domain synthesis has established that the favorable effects of artificial intelligence are bound by the adversarial countercurrent and that the net effect of technology depends on a co-evolution whose trajectory is not settled, and the dynamics of that co-evolution warrant the deeper examination that their importance to the net effect justifies. The co-evolution is the process by which the attacker and the defender adapt to each other, each responding to the advances of the other, so that an advance by either is met by an adaptation of the other that erodes the advantage the advance conferred. The deployment of artificial intelligence by the defender confers an advantage, the improved detection and response that the synthesis documents, but the advantage is met by the adaptation of the attacker, who adapts the attacks to evade the improved detection, deploys the attacker's own artificial intelligence to improve the attacks, and attacks the defensive systems themselves, eroding the advantage that the defensive deployment conferred. The net effect of the technology on the contest depends on the relative rates at which the attacker and the defender adapt, and the evidence does not permit one to project these rates with confidence, leaving the net effect uncertain.

Several features of co-evolution bear on the assessment of the net effect, and the synthesis weighs them without resolving the uncertainty they introduce. The first is the symmetry of the technology, for the artificial intelligence that improves the defense is available to the attacker, who deploys it to improve the offense, so that the technology does not confer a durable advantage on the defender but raises the capability of both the attacker and the defender, and the net effect depends on which of them the technology Favors. The second is the speed of the co-evolution, for the deployment of artificial intelligence by both the attacker and the

defender accelerates the co-evolution, compressing the period over which an advance remains effective before the adaptation erodes it, and raising the rate at which both the attacker and the defender must advance to maintain their position. The third is the asymmetry of the stakes, for the attacker needs to succeed only once while the defender must succeed always, so that co-evolution favors the attacker to the degree that technology raises the rate of the attacks more than it raises the rate of the successful defenses. These features establish that the net effect of the technology on the contest is uncertain, that it depends on the dynamics of a co-evolution whose trajectory the evidence does not permit one to project, and that the favorable effects on the defensive functions that the synthesis documents do not translate straightforwardly into a net improvement in the security of the institution.

### 9.2 Implications for the Governance of Artificial Intelligence in Cyber Risk Management

The synthesis carries implications for the governance of artificial intelligence in cyber risk management that warrant articulation, for the conditions on which the effect depends, and the counter current that bounds it together establishes the requirements of the governance that the responsible deployment of the technology requires. The first implication is that the governance must encompass the data on which the technology depends, for the effect is conditioned by the quality and the breadth of the data, and the governance must ensure that the institution possesses and maintains the data that the effect requires, managing the quality, the representativeness, and the currency of the data that the technology learns from. The second implication is that the governance must encompass the validation and the oversight of the systems, for the opacity of the systems complicates the understanding and the justification of their decisions, and the governance must ensure that the institution can validate, explain, and oversee the systems it deploys, discharging the governance burden that the opacity imposes. The third implication is that the governance must encompass the security of the systems themselves, for the systems are a target of the attacks, and the governance must ensure that the institution defends the systems against adversarial manipulation, data poisoning, and other attacks that the systems invite.

The fourth implication is that the governance must preserve the human oversight that the consequential decisions require, for the effect depends on the integration of the technology with the human analysts, and the governance must ensure that the institution preserves the human judgment and the oversight that the consequential decisions require even as it realizes the efficiency that the automation offers. The fifth implication is that the governance must attend to the fairness of the systems, for the systems can introduce the bias that the synthesis identifies, treating the populations they affect unevenly, and the governance must ensure that the institution monitors and controls the bias that the systems can reflect. These implications establish that the responsible deployment of artificial intelligence in cyber risk management requires a governance that encompasses the data, the validation, the security, the human oversight, and the fairness of the systems, and that the institution that deploys the technology without the governance realizes the favorable effects incompletely and may introduce the new harms that the ungoverned technology can produce. The synthesis thus connects the assessment of the effect of the technology to the governance that the effect requires, establishing that the effect is realized only through the discharge of the governance burden that the deployment of the technology imposes.

### 9.3 Implications for Practitioners, Institutions, and Policymakers

The synthesis carries implications for several audiences whose decisions bear on the deployment of artificial intelligence in cyber risk management, and the articulation of these implications translates the findings of the synthesis into the guidance that the audiences require. For the practitioners who deploy and operate the technology, the synthesis establishes that the effect of the technology depends on the data, the governance, and the human-machine integration that surround it, and that the realization of the effect requires the attention to these conditions as much as the deployment of the technology itself, cautioning against the expectation that the deployment of the technology will realize the effect regardless of the conditions. For the institutions that decide how much to invest in the technology, the synthesis establishes that the effect is real, meaningful, and domain-dependent, that it is most firmly established in fraud detection and least mature in cyber resilience, and that the investment should be calibrated to the domain-dependent effect and conditioned on the data, the governance, and the integration that the effect requires.

For the regulators and the policymakers who supervise the institutions and frame the environment within which they operate, the synthesis establishes that the technology improves the defensive functions to the degree and under the conditions the synthesis identifies, that the improvement is bounded by the adversarial countercurrent, and that the net effect on the security of the institution depends on a co-evolution whose trajectory is not settled, cautioning against both the complacency that would follow from an overestimate of the technology effect and the alarm that would follow from an underestimate of it. The synthesis establishes, further, that the responsible deployment of the technology requires the governance that the supervisory attention should encompass, and that the supervisory attention should extend to the data, the validation, the security, the human oversight, and the fairness of the systems, as well as to the new dependence and the systemic risk that the institutions reliance on the third-party providers of the technology introduces. The synthesis thus translates its findings into the guidance that the practitioners, the institutions, and the policymakers require, establishing that the assessment of the effect of the technology bears not only on the scholarly understanding of the technology but on the decisions of the audiences whose choices determine how the technology is deployed and governed in the cyber risk management of the financial system.

### 9.4 Threats to the Validity of the Synthesis

The credibility of a systematic review depends on the candid acknowledgment of the threats to the validity of its synthesis, and this subsection sets out the principal threats and the way the review has sought to address them. The first threat is publication and reporting bias that shapes the evidence base, for the favorable results of the deployment of the technology are more likely to be reported by the institutions and the vendors with an interest in them than the unfavorable results, so that the evidence base may overrepresent the favorable effects and underrepresent the failures and the limits. The review has sought to address this threat through the appraisal that discounts the evidence shaped by the interest in the favorable result, through the attention to the independent evidence that the bias does not shape, and through the explicit search for the evidence of the failures and the limits that the bias would suppress, but the threat cannot be eliminated, and the synthesis acknowledges that the evidence base it reviews may be biased toward the favorable result.

The second threat is confounding that complicates the attribution of the outcomes to the technology, for the outcomes improve over time for many reasons, and the attribution of the improvement to the technology, rather than to the other factors that change alongside it, is difficult and uncertain. The review has sought to address this threat through the attention to the baseline against which the effect is measured, through the

caution against the attribution to the technology of the improvements that other factors produce, and through the proposed quantitative study of Section 10, designed to identify the effect of the technology against the confounders, but the threat conditions the confidence of the synthesis, and the magnitude of the effect it reports is attended by the uncertainty that the confounding introduces. The third threat is the dynamism of the subject, for the synthesis captures the state of the evidence at a moment and may be overtaken by the developments that the rapid evolution of the technology and the threat environment produce, a threat the review acknowledges and that conditions the durability of its findings. These threats to validity do not vitiate the synthesis, which remains the most rigorous assessment that the available evidence supports, but they condition its findings, and the synthesis reports its findings with the hedging that the threats to validity require, distinguishing throughout the findings that the evidence establishes firmly from those that the threats to validity render more uncertain.

### 9.5 The Boundaries of Generalization

The synthesis has been confined to the United States banking context, a confinement that permits the coherent treatment of a single regulatory and market environment but that bounds the generalization of the findings, and the boundaries of the generalization warrant the explicit statement that their importance to the use of the findings justifies. The findings concern the effect of artificial intelligence on the cyber risk management of the United States financial institutions, and their extension to the institutions of other sectors and other jurisdictions, which operate under different regulatory regimes, confront different threat environments, and possess different data and capabilities, is not warranted without the examination of the differences that the extension would require. The effect of the technology, conditioned as the synthesis has found on the data, the governance, and the human-machine integration that surround it, may differ in contexts that differ in these conditions, and the findings of the synthesis, grounded in the United States banking context, characterize the effect in that context rather than in the contexts that differ from it.

The boundaries of the generalization extend further to the temporal dimension, for the synthesis captures the state of the evidence and the technology at a moment, and the rapid evolution of both the technology and the threat environment means that the findings characterize the effect at that moment rather than the effect that the subsequent developments will produce. The generative and the agentic systems whose deployment is recent, and whose evidence base is still forming, may alter the effect that the synthesis has characterized, and the adversarial co-evolution whose trajectory the synthesis has found unsettled may shift the net effect in the directions that the present evidence does not permit one to project. The synthesis is, in consequence, a characterization of the effect of the technology on the cyber risk management of United States banking as the evidence at the present moment permits one to assess it, valid for that context and that moment, and subject to the revision that the extension to the other contexts and the evolution of the technology and the threat environment will require. The acknowledgment of these boundaries is not a weakness of the synthesis but a condition of its honest use, ensuring that the findings are applied within the context and the moment for which the evidence supports them rather than extended to the contexts and the moments for which it does not.

## 10.A Proposed Agenda for Primary Empirical Research

The systematic review has revealed, alongside the findings it supports, the limitations of the existing evidence base, and in particular the scarcity of the independent, operationally realistic, and rigorously designed primary evidence that would permit the effect of artificial intelligence on cyber risk management

to be established with greater confidence than the existing base allows. This section converts those limitations into a forward program, proposing the agenda for the primary empirical research that the field requires. The agenda is proposed, not executed, within the present paper, which is a systematic review and synthesis rather than a primary study, and the proposals are set out clearly as designs for the research that the field should undertake rather than as research the present paper has conducted. The clarity of this distinction is a matter of research integrity, for the present paper does not manufacture the primary evidence it identifies as necessary but specifies the research that would generate it.

### 10.1 A Proposed Interview Study of Cybersecurity Professionals

The first element of the proposed agenda is a structured interview study of the cybersecurity professionals who deploy and operate the artificial-intelligence systems in the financial institutions, designed to generate the operationally grounded qualitative evidence that the existing base lacks. The study would recruit a sample of cybersecurity professionals across a range of institutional types, from the largest banks to the smaller and the financial technology institutions, and conduct structured interviews directed at the questions the systematic review has identified as underdetermined by the existing evidence. The interviews would explore the professionals' experience of the effect of technology on detection and the response conditions on which they find the effect to depend, the limits and the failures they have encountered, the governance burden they have borne, and the adversarial co-evolution they have observed. The study would be designed according to the established conventions of the qualitative research, with the sampling, the interview protocol, the consent, and the analysis specified in advance and conducted with the rigor that the credibility of the evidence requires, and it would generate the operationally grounded qualitative evidence, from the professionals who possess the direct experience, that the existing base, dominated by the vendor claims and the controlled studies, does not supply.

The proposed interview study would address several of the specific gaps the systematic review has identified. It would address the gap in the operationally realistic evidence, for the professionals possess the direct experience of the technology in the operational conditions that the controlled studies do not reflect. It would address the gap in the independent evidence, for the professionals suitably sampled and assured of the confidentiality that Candor requires, can provide the assessment that the vendors, with their interest in the favorable result, cannot. It would address the gap in the evidence regarding the conditions and the limits of the effect, because the professionals possess direct experience of the conditions on which the effect depends and the limits at which it fails. And it would address the gap in the evidence on the adversarial co-evolution, for the professionals to observe the co-evolution directly in their daily work. The study would thus generate qualitative evidence that the systematic review has identified as necessary and that the present synthesis, confined to the existing evidence, could not supply.

### 10.2 A Proposed Quantitative Study of Effect on Outcomes

The second element of the proposed agenda is a quantitative study of the effect of artificial intelligence on the cyber risk outcomes of the financial institutions, designed to generate rigorous and independent quantitative evidence that the existing base, dominated by vendor figures and controlled benchmark studies, does not supply. The study would assemble the data on the cyber risk outcomes, namely the threats detected, the frauds prevented, the incidents contained, and the resilience achieved, across a sample of institutions and over a period spanning the adoption of the technology, and it would estimate the effect of the technology on the outcomes through the methods of causal inference that the observational data require. The central

challenge of the study, which its design must address, is the problem of the counterfactual identified in the introduction, the difficulty of distinguishing the effect of the technology from the many other factors that change alongside its adoption, and the study would address this challenge through the established methods of the causal inference from the observational data, including the difference-in-differences designs that compare the institutions that adopt the technology with those that do not over the period of the adoption, the instrumental-variable methods that exploit the variation in the adoption that is unrelated to the outcomes, and the controls for the confounding factors that the richness of the data permits.

The proposed quantitative study would address the gaps in the quantitative evidence that the systematic review has identified. It would address the gap in the independent quantitative evidence, for the study would generate the estimates of the effect from the data through the transparent methods that the reader can examine, rather than relying on the figures of the vendors and the institutions with an interest in the favorable result. It would address the gap in the operationally realistic quantitative evidence, for the study would estimate the effect on the actual outcomes of the actual institutions in the operational conditions, rather than on the benchmark datasets under the controlled conditions of the existing studies. And it would address the gap in the evidence on the magnitude of the effect, for the study would generate the quantified estimates of the effect that the synthesis, confined to the heterogeneous and the incomparable figures of the existing base, could not. The study would be demanding, for the assembly of the data and the identification of the effect against the confounders are difficult, and the access to the data of the institutions, reluctant to disclose the details of their cyber postures, would require the cooperation that the sensitivity of the data complicates, but the study would generate the rigorous quantitative evidence that the field requires and that the existing base does not supply.

### 10.3 A Proposed Program of Documented Case Studies

The third element of the proposed agenda is a program of documented case studies of the deployment of artificial intelligence in financial institutions, designed to generate the detailed and operationally grounded case evidence that complements the interview and the quantitative studies. The program would conduct in-depth case studies of the deployment of the technology in a sample of institutions, documenting the deployment, the conditions, the effects, and the limits in the detail that the case-study method permits, and it would assemble the cases into a comparative analysis that surfaces the patterns across them. The case studies would rely, like the case evidence of the present synthesis, on the documented and the attributable record, supplemented, where the cooperation of the institutions permits, by the direct study of the deployment that the institutions disclose, and they would observe the strict discipline with respect to the evidence, relying only on the documented and the attributable information and refraining from the manufacture of the evidence, that the present paper has observed. The program of the case studies would generate the detailed and operationally grounded case evidence that grounds the general findings in the specific experience, complementing the qualitative evidence of the interview study and the quantitative evidence of the outcome study, and completing the triangulation of the qualitative, the quantitative, and the case evidence that the rigorous assessment of the technology effect requires.

### 10.4 The Triangulation of the Proposed Methods

The three elements of the proposed research agenda, the interview study of the cybersecurity professionals, the quantitative study of the effect on the outcomes, and the program of the documented case studies, are proposed not as alternatives among which the field should choose but as complementary methods whose

combination would provide the triangulation that the rigorous assessment of the technology effect requires. The triangulation of the methods addresses the limitations of each, for the qualitative evidence of the interview study, strong on the operationally grounded understanding of the conditions and the limits of the effect, is weak on the quantification that the assessment of the magnitude requires; the quantitative evidence of the outcome study, strong on the quantification of the effect, is weak on the understanding of the conditions and the mechanisms that the qualitative evidence supplies; and the case evidence, strong on the detailed and the contextual understanding of the particular deployments, is weak on the generalization that the larger samples of the other methods permit. The combination of the methods, each addressing the limitations of the others, would provide triangulated evidence that no single method alone could supply.

The triangulation would also address the threats to the validity that the synthesis has identified, for the combination of the methods would permit the cross-checking of the findings across the methods, the corroboration of the quantitative estimates by the qualitative understanding and the case evidence, and the detection of the biases and the confounders that any single method might miss. The interview evidence would surface the failures and the limits that the publication bias suppresses in the reported evidence; the quantitative study would identify the effect against the confounders that the other methods cannot control; and the case studies would ground the general findings in the specific experience that reveals the conditions and the mechanisms. The triangulation of the methods would thus provide not only the complementary strengths that each method contributes but the cross-validation that the combination permits, and it would generate the rigorous, independent, and operationally grounded evidence that the existing base lacks and that the assessment of the effect of artificial intelligence on the cyber risk management of the United States banking ultimately requires. The proposed agenda is offered, in consequence, not as a menu of the separate studies but as an integrated program whose combination would supply the triangulated evidence that the present synthesis, confined to the existing base, could not, and that the field requires to establish the effect of the technology with the confidence that the consequential decisions of the institutions, the regulators, and the policymakers demand.

## 11.Conclusion

This paper has addressed, through a systematic review and a case-based synthesis, the question of the extent to which artificial intelligence improves cyber risk detection and incident response in United States financial institutions. The question is consequential, for the institutions, the regulators, and the scholars who must assess the technology require an account of its effect that distinguishes the well-supported claims from the poorly supported ones and that characterizes the conditions and the limits of the effect, and the existing discussion, dominated by the enthusiasm of the institutions and the vendors and the skepticisms of the critics, has lacked the rigorous and the differentiated synthesis that the question demands. The paper has supplied synthesis, assembling and appraising the heterogeneous evidence according to a transparent protocol, synthesizing the appraised evidence within the four focus domains, grounding the synthesis in the documented institutional case evidence, and identifying the agenda for the primary research that the existing base lacks.

The synthesis has found that artificial intelligence improves cyber risk detection and incident response in the United States banking to a degree that is real, meaningful, and domain dependent. The improvement is most firmly established and most substantial in fraud detection, where the adaptive-learning methods materially outperform the rule-based methods they displace; it is strong but qualified in security operations automation, where the technology has compressed the time to detect and to contain the incidents while the

transformation remains incomplete; it is favorable but confounded in threat intelligence, where the direction is clear but the magnitude is harder to isolate; and it is favorable but immature in cyber resilience, where the contributions of the other domains carry through but the direct effect on the resilience outcomes is the least well established. The improvement, where it obtains, is conditional rather than absolute, depending on the data, the governance, and the human-machine integration that surround the technology, and it is bounded by the adversarial counter current, the offensive use of the technology by the adversary, the attacks on the defensive systems themselves, and the opacity, the false positives, and the governance burden that the technology introduces. The net effect of the technology on the contest between the defender and the attacker depends on a co-evolution whose trajectory the evidence does not permit one to project.

The paper has observed throughout a strict discipline with respect to the evidence, relying only on the documented and the attributable sources, appraising the quality of the evidence rather than accepting its claims at face value, distinguishing the findings the evidence establishes from those it merely suggests, and refraining from the manufacture of the primary data that the field lacks. This discipline has constrained the synthesis to the limits of the existing evidence, and the paper has been candid about those limits, but it has also rendered the synthesis credible, for the findings it reports are grounded in the appraised evidence and hedged according to the confidence the evidence permits. The proposed agenda for the primary empirical research, the interview study of the cybersecurity professionals, the quantitative study of the effect on the outcomes, and the program of the documented case studies, is directed at the generation of the evidence that would permit the effect of the technology to be established with the confidence that the existing base does not allow, and it converts the limitations of the present synthesis into the forward program that the field requires.

Artificial intelligence has become central to the cyber risk management of United States banking, and its centrality will only grow as the technology advances, the threat environment evolves, and the regulatory attention intensifies. The question of the extent to which the technology improves the detection and the response that the cyber risk management requires is, in consequence, of growing importance, and the rigorous and differentiated answer to it, grounded in the appraised evidence and attentive to the conditions and the limits and the counter current, is of growing value. The present paper has supplied the answer that the existing evidence permits, finding the effect real, meaningful, domain-dependent, conditional, and bounded, and identified the research that would refine the answer as the evidence base matures. It is to the rigorous, honest, and evidence-grounded assessment of one of the most consequential technologies of the contemporary financial system, and to the protection of the institutions, the consumers, and the financial system that cyber risk management ultimately serves, that this work is offered.

.

## References

Abdullahi, M., Baashar, Y., Alhussian, H., Alwadain, A., Aziz, N., Capretz, L. F., & Abdulkadir,

S. J. (2022). Detecting cybersecurity attacks in the Internet of Things using artificial intelligence methods: A systematic literature review. Electronics, 11(2), 198.

Aimeur, E., & Castelfranchi, C. (2024). Artificial intelligence in financial fraud detection: Challenges and opportunities. In S. K. Gupta et al. (Eds.), Emerging Technologies in Banking and Finance. Springer.

Ali, A., & Shah, M. (2024). What hinders adoption of artificial intelligence for cybersecurity in the banking sector? Information, 15(12), 760.

Aljunaid, S. K., Almheiri, S. J., Dawood, H., & Khan, M. A. (2025). Secure and transparent banking: Explainable AI-driven federated learning model for financial fraud detection. Journal of Risk and Financial Management, 18(4), 179.

Apruzzese, G., Laskov, P., Montes de Oca, E., Mallouli, W., Brdalo Rapa, L., Grammatopoulos,

A. V., & Di Franco, F. (2023). The role of machine learning in cybersecurity. Digital Threats: Research and Practice, 4(1), 1-38.

Bahnsen, A. C., Aouada, D., Stojanovic, A., & Ottersten, B. (2016). Feature engineering strategies for credit card fraud detection. Expert Systems with Applications, 51, 134-142.

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317-331.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Srndic, N., Laskov, P., Giacinto, G., & Roli, F. (2013). Evasion attacks against machine learning at test time. In Machine Learning and Knowledge Discovery in Databases (pp. 387-402). Springer.

Bouchama, F., & Kamal, M. (2021). Enhancing cyber threat detection through machine learning-based behavioral modeling of network traffic patterns. International Journal of Business Intelligence and Big Data Analytics, 4(9), 1-9.

Carcillo, F., Le Borgne, Y. A., Caelen, O., Kessaci, Y., Oble, F., & Bontempi, G. (2021). Combining unsupervised and supervised learning in credit card fraud detection. Information Sciences, 557, 317-331.

Dal Pozzolo, A., Boracchi, G., Caelen, O., Alippi, C., & Bontempi, G. (2018). Credit card fraud detection: A realistic modelling and a novel learning strategy. IEEE Transactions on Neural Networks and Learning Systems, 29(8), 3784-3797.

Federal Deposit Insurance Corporation. (2025). 2025 report on cybersecurity and resilience. FDIC.

Federal Financial Institutions Examination Council. (2019). Business continuity management booklet, FFIEC information technology examination handbook. FFIEC.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations.

Guembe, B., Azeta, A., Misra, S., Osamor, V. C., Fernandez-Sanz, L., & Pospelova, V. (2022). The emerging threat of artificial intelligence on competition in cybersecurity. Information, 13(1), 39.

Hassan, M., Aziz, L. A.-R., & Andriansyah, Y. (2023). The role artificial intelligence in modern banking: An exploration of AI-driven approaches for enhanced fraud prevention, risk management, and regulatory compliance. Reviews of Contemporary Business Analytics, 6(1), 110-132.

IBM Security & Ponemon Institute. (2025). Cost of a data breach report 2025. IBM Corporation.

International Monetary Fund. (2024). Artificial intelligence and its impact on financial markets and financial stability (Global Financial Stability Report, Chapter 3). IMF.

Kaur, R., Gabrijelcic, D., & Klobucar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. Information Fusion, 97, 101804.

Kovacevic, A., Radenkovic, S. D., & Nikolic, D. (2024). Artificial intelligence and cybersecurity in banking sector: Opportunities and risks. arXiv preprint arXiv:2412.04495.

Kuzlu, M., Fair, C., & Guler, O. (2021). Role of artificial intelligence in the internet of things (IoT) cybersecurity. Discover Internet of Things, 1(1), 7.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations.

Mienye, I. D., & Sun, Y. (2023). A deep learning ensemble with data resampling for credit card fraud detection. IEEE Access, 11, 30628-30638.

Mohammed, M. A., Boujelben, M., & Abid, M. (2023). A novel approach for fraud detection in blockchain-based healthcare networks using machine learning. Future Internet, 15(8), 250.

Najem, R., Bahnasse, A., & Talea, M. (2024). Toward an enhanced fraud detection system in banking using artificial intelligence. Procedia Computer Science, 233, 880-889.

Nassar, A., & Kamal, M. (2021). Machine learning and big data analytics for cybersecurity threat detection: A holistic review of techniques and case studies. Journal of Artificial Intelligence and Machine Learning in Management, 5(1), 51-63.

National Institute of Standards and Technology. (2024). Artificial intelligence risk management framework: Generative artificial intelligence profile (NIST AI 600-1). U.S. Department of Commerce.

Nguyen, T. T., & Reddi, V. J. (2023). Deep reinforcement learning for cyber security. IEEE Transactions on Neural Networks and Learning Systems, 34(8), 3779-3795.

Office of the Comptroller of the Currency. (2024). Semiannual risk perspective. OCC.

Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. P. (2018). SoK: Security and privacy in machine learning. IEEE European Symposium on Security and Privacy, 399-414.

PYMNTS Intelligence & Block. (2025). The 2025 state of fraud and financial crime in the United States. PYMNTS.

Rajhans, M., & Khawarey, V. (2026). Empirical analysis of adversarial robustness and explainability drift in cybersecurity classifiers. arXiv preprint arXiv:2602.06395.

Rane, N., Choudhary, S., & Rane, J. (2024). Artificial intelligence and machine learning in business and management: Applications, opportunities, and challenges. Studies in Economics and Business Relations, 5(1), 23-41.

Renaud, K., Warkentin, M., & Westerman, G. (2023). From ChatGPT to HackGPT: Meeting the cybersecurity threat of generative AI. MIT Sloan Management Review, 64(4).

Sarker, I. H. (2023). Machine learning for intelligent data analysis and automation in cybersecurity: Current and future prospects. Annals of Data Science, 10(6), 1473-1498.

Sarker, I. H., Furhad, M. H., & Nowrozy, R. (2021). AI-driven cybersecurity: An overview, security intelligence modeling and research directions. SN Computer Science, 2(3), 173.

Sarker, I. H., Kayes, A. S. M., Badsha, S., Alqahtani, H., Watters, P., & Ng, A. (2020). Cybersecurity data science: An overview from machine learning perspective. Journal of Big Data, 7(1), 41.

Shaukat, K., Luo, S., Varadharajan, V., Hameed, I. A., & Xu, M. (2020). A survey on machine learning techniques for cyber security in the last decade. IEEE Access, 8, 222310-222354.

Sun, N., Ding, M., Jiang, J., Xu, W., Mo, X., Tai, Y., & Zhang, J. (2023). Cyber threat intelligence mining for proactive cybersecurity defence: A survey and new perspectives. IEEE Communications Surveys & Tutorials, 25(3), 1748-1774.

Vassakis, K., Petrakis, E., & Kopanakis, I. (2023). Explainable artificial intelligence in cybersecurity: A comprehensive review. ScienceDirect (Computers & Security review series).

Verizon. (2025). 2025 data breach investigations report (DBIR). Verizon Business.

Yaseen, A. (2023). AI-driven threat detection and response: A paradigm shift in cybersecurity.

International Journal of Information and Cybersecurity, 7(12), 25-43.

Zaheer, N., Mahmood, T., & Mehmood, R. (2025). Intrinsic bias in cybersecurity-focused machine learning models: Reliability and trust implications. Computers & Security, 148, 103-121.

Zhang, Z., Al Hamadi, H., Damiani, E., Yeun, C. Y., & Taher, F. (2022). Explainable artificial intelligence applications in cyber security: State-of-the-art in research. IEEE Access, 10, 93104-93139.

Zografopoulos, I., Ospina, J., Liu, X., & Konstantinou, C. (2021). Cyber-physical energy systems security: Threat modeling, risk assessment, resources, metrics, and case studies. IEEE Access, 9, 29775-29818.

Zou, Q., Sun, X., Liu, P., & Singhal, A. (2024). An empirical study of attack-related events in cyber threat intelligence. Computers & Security, 139, 103-119.

 Affiliation: Lamar University, Beaumont, Texas, USA