

The Role of Artificial Intelligence in Fraud Detection: A Theoretical Examination of Financial Security, Risk Management, and Corporate Governance

Arowolo, Isiaka O (Ph.D)

Department of management and Accounting,
faculty of Management and Social Science,
Lead City University.

Email: arowologbenga@yahoo.com **Phone No:** +234 803 308 8681

Abstract

Artificial intelligence (AI) has become central to how organizations detect financial fraud, yet the theoretical literature addressing this shift remains split across three largely separate conversations: technical work on detection accuracy, governance work on organizational oversight, and an emerging and considerably more troubling literature on AI's use by fraud perpetrators themselves. This paper offers a theoretical examination that integrates these three conversations around a single argument: AI does not simply improve fraud detection as a technical capability; it restructures the relationship between financial security, enterprise risk management, and corporate governance in ways that classical fraud and agency theory did not anticipate and do not fully explain. Drawing on fraud triangle and fraud diamond theory, agency theory, and enterprise risk management frameworks, alongside recent empirical evidence on machine learning applications in financial statement fraud, anti-money laundering, and synthetic-identity crime, the paper develops an account of fraud detection as a genuinely adversarial and co-evolutionary process rather than a static technical problem to be solved once. It finds that AI simultaneously strengthens detection capability, in the form of demonstrably superior fraud- and misstatement-prediction models, and destabilizes the assumptions underlying that capability, because the same generative technologies improving detection are available to, and increasingly used by, the perpetrators detection systems are built to catch. The paper argues that this dual-use reality has direct implications for corporate governance, since boards and audit committees can no longer treat AI-enabled fraud detection as a static control to be implemented once and monitored passively, but must instead govern an ongoing arms race whose terms shift as the underlying technology develops on both sides.

Keywords: artificial intelligence; fraud detection; financial security; risk management; corporate governance; agency theory; anti-money laundering; deepfakes.

Introduction

Financial fraud has always been, at its theoretical core, an information problem: perpetrators exploit gaps between what an organization's records show and what actually occurred, and detection depends on an organization's capacity to notice that gap before the damage compounds. For most of the modern history of fraud examination, that capacity was bounded by the same constraints that bound all human cognition, auditors and investigators could review only a sample of transactions, and the patterns distinguishing fraudulent activity

from ordinary business noise were often too subtle or too dispersed across a large dataset for a human reviewer to reliably detect. Artificial intelligence changes this calculus directly, and the evidence that it does so is now substantial: machine-learning models trained on financial statement data have been shown to substantially outperform the ratio-based benchmarks fraud examiners have relied on for decades (Bao, Ke, Li, Yu, & Zhang, 2020), and deep-learning approaches applied to the unstructured text of corporate disclosures have demonstrated an ability to detect fraud signals embedded in managerial language that purely numerical models miss entirely (Craja, Kim, & Lessmann, 2020).

Yet treating this as a straightforward technical success story understates what is actually happening. The same generative and machine-learning technologies that have improved fraud detection are simultaneously available to the people fraud detection is designed to catch, and the resulting picture is not one of steadily improving organizational defenses but of a genuinely adversarial and co-evolving contest. Deepfake-enabled impersonation has already produced fraud losses in the tens of millions of dollars in individual incidents, synthetic identities constructed from a blend of real and fabricated personal data represent one of the fastest-growing categories of financial crime, and the technology underlying both capabilities is advancing on a timeline that is not obviously slower than the timeline on which detection technology is advancing (Dhingra, Chakraborty, & Sharma, 2026). A theory of AI and fraud detection that examines only the detection side of this contest, without examining the perpetration side, is at best incomplete and at worst actively misleading about the level of organizational security AI adoption actually provides.

This paper develops a more complete theoretical account by integrating three literatures that have generally been treated separately: classical fraud theory and its extensions, agency theory and the governance rationale for fraud control, and enterprise risk management as the organizational framework within which fraud detection technology is actually deployed. Its central argument is that **AI restructures fraud detection from a static technical control into a dynamic, co-evolutionary governance problem, and that corporate governance frameworks built around the assumption of a stable detection technology are increasingly poorly matched to the adversarial reality AI adoption has created.** The paper reviews the theoretical foundations relevant to this argument; examines AI's demonstrated contribution to fraud detection across structured, unstructured, and network-based financial data; turns to the adversarial dimension in which AI functions as a fraud-enabling rather than fraud-detecting technology; and considers the resulting implications for financial security, enterprise risk management, and the corporate governance structures responsible for overseeing all of the above.

The stakes of getting this theoretical account right are not merely academic. Financial statement fraud, money laundering, and identity-based financial crime collectively account for losses measured in the hundreds of billions of dollars annually across the global economy, and the organizations bearing primary responsibility for containing those losses — audit committees, risk officers, compliance functions — are making real-time decisions about how much to trust AI-generated fraud indicators, how much autonomy to grant automated detection

systems, and how much of their own limited oversight capacity to devote to understanding technology whose internal workings few board members are trained to evaluate. A theory that treats AI-enabled detection as a solved technical problem offers those decision-makers false reassurance; a theory that treats it as an unstable, adversarially contested capability offers a considerably more accurate, if less comfortable, basis for governance. This paper aims to supply the latter.

Conceptualizing AI-Enabled Fraud Detection

Fraud detection has traditionally been understood as a control activity situated within an organization's broader system of internal controls, positioned to catch what preventive controls fail to stop. AI intervenes in this activity primarily by expanding the volume, granularity, and type of evidence that can be examined for signs of fraud. Bertomeu, Cheynel, Floyd, and Pan (2021), using machine learning to detect ongoing financial misstatements, find that accounting variables alone predict misstatement poorly and that genuinely useful detection requires combining accounting information with audit and market variables — an early indication that AI's contribution to fraud detection is not simply processing existing financial data faster, but integrating data sources that human reviewers have historically examined in isolation from one another, if at all.

This paper organizes AI's role in fraud detection around three data domains that the literature has developed somewhat independently: structured financial and transactional data, of the kind Bao et al. (2020) and Bertomeu et al. (2021) examine; unstructured textual data, of the kind Craja et al. (2020) examine in corporate disclosures; and network and relational data, of the kind anti-money laundering research examines in tracing the movement of illicit funds across accounts and institutions (Chen, Van Khoa, Teoh, Nazir, Karuppiah, & Lam, 2018). Each domain has generated its own detection methodology and its own body of empirical evidence, but each is also increasingly exposed to the adversarial dynamic this paper treats as central: a detection model trained on one domain's historical fraud patterns is vulnerable to perpetrators who adapt their methods once they understand, or can infer, how that model works — a vulnerability classical fraud theory, developed before large-scale automated detection existed, did not need to theorize.

Theoretical Foundations

Fraud Theory: From the Fraud Triangle to the Fraud Diamond

Cressey's (1953) fraud triangle remains the foundational theoretical account of why individuals commit fraud, proposing that fraud occurs at the intersection of three conditions: perceived financial pressure, perceived opportunity to commit the act without detection, and a rationalization that allows the perpetrator to reconcile the act with their self-image as a fundamentally honest person. Wolfe and Hermanson (2004) extend this framework by adding a fourth element, capability, arguing that pressure, opportunity, and rationalization are insufficient without the perpetrator also possessing the specific skills and position necessary to recognize and exploit the opportunity in practice. This extension is theoretically significant for

the present paper because AI directly affects the capability element in ways the original triangle could not anticipate: AI-generated tools lower the skill threshold required to exploit a given fraud opportunity, meaning individuals who previously lacked the capability to execute sophisticated fraud — constructing a convincing synthetic identity, or producing a persuasive deepfake impersonation — may now possess it simply by accessing widely available generative tools, independent of their own underlying technical skill.

This capability shift cuts in both directions relevant to this paper's argument. On the detection side, AI expands the capability of the organization defending against fraud, in the sense developed throughout the sections below. On the perpetration side, AI expands the capability of the individual or organized group attempting fraud, a dynamic the fraud diamond's capability element anticipates conceptually but that Wolfe and Hermanson (2004), writing before generative AI existed in its current form, could not have examined empirically. A theoretically complete account of AI and fraud therefore has to treat capability as contested terrain rather than a fixed organizational asset — both sides of a fraud event now have access to a rapidly advancing capability-enhancing technology, and which side's capability advances faster in a given period is an empirical question rather than a theoretical certainty.

It is worth noting, too, that AI's effect on the fraud triangle's other elements is not confined to capability alone. Perceived opportunity, Cressey's (1953) second element, has historically depended on a perpetrator's assessment of the likelihood that a given act will be detected — an assessment that AI-enabled detection, all else equal, should reduce by raising the perceived probability of discovery. But the adversarial evidence discussed later in this paper complicates that expectation: where perpetrators are aware, correctly or not, that detection systems are trainable and therefore exploitable, perceived opportunity may not fall in step with actual detection capability, and could in some circumstances rise if perpetrators come to believe, whether accurately or not, that generative tools give them a durable edge over automated defenses. Rationalization, the third element, is likewise not obviously unaffected by AI's involvement: a perpetrator using a generative tool to construct a synthetic identity or fraudulent disclosure may find it easier, not harder, to rationalize the act as a technical exercise mediated by a tool rather than a direct personal deception — a psychological distancing effect the classical fraud triangle, developed to explain face-to-face and paper-based fraud, did not need to theorize.

Agency Theory and the Governance Rationale for Fraud Detection

Jensen and Meckling's (1976) agency theory supplies the governance rationale underlying fraud detection as an organizational function. Where ownership and control are separated, as in the modern public corporation, agents (managers) may pursue interests that diverge from those of principals (shareholders), and the resulting agency costs include the costs of monitoring mechanisms designed to constrain that divergence — among which fraud detection systems are a direct and consequential example. On this account, AI-enabled fraud detection is best understood not merely as a technical control but as a monitoring mechanism whose adoption is itself a governance decision, one that changes the cost structure of monitoring and, in principle, the level of agency cost a firm can economically justify reducing.

This governance framing carries an important theoretical implication that the purely technical fraud-detection literature tends to understate: because AI-enabled monitoring is cheaper and more comprehensive than the human-reviewer-based monitoring it replaces, agency theory would predict firms rationally investing in more monitoring than before, up to the point where the marginal cost of additional AI-enabled detection equals its marginal benefit in reduced agency cost. Munoko, Brown-Liburd, and Vasarhelyi (2020), examining the ethical implications of AI in auditing, identify accountability for AI-generated conclusions as a distinct governance risk at the firm level — a risk with direct relevance here, since a firm that expands AI-enabled monitoring without correspondingly clarifying who is accountable for that monitoring's failures or false conclusions may reduce one category of agency cost while inadvertently creating another.

Enterprise Risk Management as an Organizing Framework

The Committee of Sponsoring Organizations of the Treadway Commission's enterprise risk management framework (COSO, 2017) provides the organizational architecture within which fraud detection technology is actually deployed in practice, situating fraud risk as one component of a broader risk landscape that must be identified, assessed, responded to, and monitored in relation to the organization's strategy and performance objectives, rather than managed as an isolated technical function. This framework matters theoretically for the present paper because it resists treating fraud detection as a discrete technology decision separable from the organization's broader risk appetite and governance structure — a resistance the adversarial, co-evolutionary dynamic discussed later in this paper makes considerably more consequential than COSO's original framework, developed before AI-enabled fraud at scale existed, could have anticipated. An enterprise risk management approach to AI-enabled fraud detection has to treat the technology's own vulnerabilities and failure modes as a risk category in their own right, not merely as a solved input to the organization's broader risk picture.

AI's Contribution to Fraud Detection Capability

Structured-Data Fraud Detection: Financial Statement and Transactional Fraud

Bao et al. (2020) develop a machine-learning fraud-prediction model using raw financial statement numbers rather than the financial ratios earlier detection models relied upon, and find it substantially outperforms established benchmark models in identifying fraud among publicly traded U.S. firms — a methodological advance with direct implications for how effectively financial statement fraud can be identified before it produces material harm to investors and other stakeholders. Bertomeu et al. (2021) extend this evidence, finding that machine learning combining accounting, audit, and market variables improves misstatement detection meaningfully beyond what accounting variables alone can achieve. Taken together, these findings establish that AI's contribution to structured-data fraud detection is genuine and well-evidenced, but also that its value depends on the breadth of data supplied to the model — a model trained narrowly on the same accounting data fraud has always been detected from is unlikely to match the performance these studies document.

The theoretical significance of this data-breadth finding extends beyond a purely technical observation about model inputs. It implies that AI's fraud-detection advantage is, at root, an advantage in integration rather than an advantage in any single algorithm's sophistication: the models Bao et al. (2020) and Bertomeu et al. (2021) describe outperform their predecessors substantially because they combine data sources — accounting, audit, market — that organizational silos have historically kept separate, not primarily because the underlying statistical or machine-learning technique is dramatically more powerful in the abstract. This has a direct practical implication for the governance discussion later in this paper: an organization's fraud-detection capability is bounded not only by the sophistication of the AI tools it purchases, but by the degree to which its own internal data architecture allows those tools to see across functional boundaries that were historically maintained for reasons — departmental autonomy, data-privacy compliance, legacy system incompatibility — that have little to do with fraud risk as such.

Text and Unstructured-Data Fraud Detection

Craja et al. (2020) extend fraud detection into unstructured data specifically, developing a hierarchical attention network that analyzes the managerial discussion and analysis sections of corporate annual reports alongside conventional financial ratios, and demonstrating that the language management uses to describe company performance carries fraud-relevant signal that purely numerical models cannot access. This finding is theoretically significant because it demonstrates that AI's fraud-detection advantage over earlier statistical methods is not confined to processing more of the same kind of data faster; it extends to processing entirely different kinds of data — natural language — that human fraud examiners have always had informal access to but that earlier detection technology could not systematically analyze at scale. The theoretical implication is that fraud, as a phenomenon, leaves a wider evidentiary trail than classical fraud-detection methodology assumed, and AI's contribution has been partly to reveal the existence of that wider trail rather than simply to process a previously identified trail more efficiently.

Anti-Money Laundering and Financial-Crime Network Detection

Chen et al. (2018), reviewing machine-learning applications in anti-money laundering, identify link analysis, behavioral modeling, risk scoring, and anomaly detection as the principal technique categories through which AI has been applied to suspicious-transaction detection, and find that network-based approaches capable of tracing relationships across accounts and institutions represent a particular strength of machine-learning methods relative to the rule-based threshold systems that preceded them. Goecks, Korzenowski, Terra Neto, de Souza, and Mareth (2022), conducting a systematic literature review spanning both anti-money laundering and financial fraud detection more broadly, find that research treating these two problems jointly remains comparatively sparse relative to research treating them separately, despite substantial conceptual overlap in the underlying detection techniques — a gap with practical significance, since money laundering and financial statement fraud frequently co-occur within the same criminal schemes, and detection systems built to catch only one may systematically miss evidence relevant to the other.

The network-analytic strength Chen et al. (2018) identify is worth dwelling on theoretically, because it points to a property of AI-enabled detection that the structured- and text-based methods discussed above do not share in the same degree: network-based detection derives much of its power not from any single transaction's characteristics but from the pattern of relationships among many transactions and account holders considered jointly, a form of evidence that is, by construction, difficult for any individual perpetrator to fully anticipate or control, since it depends on the behavior of counterparties and intermediaries the perpetrator does not directly manage. This relational property offers a partial, though only partial, counterweight to the co-evolutionary vulnerability discussed later in this paper: a perpetrator using generative AI to construct a more convincing individual transaction or identity still leaves a network-level footprint that a sufficiently comprehensive relational model may detect, even where the individual-level deception itself succeeds. Whether this relational advantage proves durable as adversaries increasingly coordinate across networks of synthetic identities specifically designed to obscure relational patterns, rather than merely individual transactions, remains an open empirical question this paper returns to in its discussion of future research.

The Adversarial Dimension: AI as a Fraud-Enabling Technology

Generative AI, Deepfakes, and Synthetic Identity

The theoretical account developed so far treats AI exclusively as a detection technology, but a complete examination of AI and fraud cannot omit the reverse relationship: the same generative capabilities that improve detection are simultaneously available to fraud perpetrators, and the evidence that they are being used is now substantial rather than speculative. Dhingra, Chakraborty, and Sharma (2026), examining the structure and development of AI-driven financial deception, document the convergence of two previously distinct threats — deepfake technology capable of defeating biometric authentication and voice-based verification, and synthetic identity fraud that blends genuine and fabricated personal data to construct fictitious personas capable of passing conventional identity checks. Their analysis identifies weaknesses specifically in the know-your-customer and anti-money laundering infrastructure that structured-data detection methods, of the kind reviewed in the preceding section, were built to secure, and traces how generative adversarial networks and voice-cloning technologies are lowering the technical skill threshold required to exploit those weaknesses — precisely the capability-shift dynamic Wolfe and Hermanson's (2004) fraud diamond anticipates conceptually.

This adversarial reality has a direct theoretical consequence for how fraud detection capability should be evaluated. A detection model's accuracy against historical fraud patterns, of the kind Bao et al. (2020) and Craja et al. (2020) document, measures performance against fraud as it has been committed, not fraud as it can now be committed using the same class of generative technology the detection model itself may depend on. Detection capability and fraud capability are therefore not independent variables that happen to both be improving; they are drawing on an overlapping technological base, and an organization's net security position depends on the relative rate at which each side's capability advances — a relative rate that

classical fraud theory, developed for a world in which technical capability advanced slowly and asymmetrically in favor of institutional defenders, did not need to model explicitly.

The Co-Evolutionary Arms Race

The implication of the preceding section is that AI-enabled fraud detection is better modeled as a co-evolutionary contest than as a static technical achievement. This framing has several theoretical consequences this paper treats as central to its overall argument. First, detection systems trained on historical fraud patterns face a structural disadvantage against perpetrators using generative tools to construct genuinely novel patterns, since the statistical regularities a machine-learning model learns from past fraud are precisely what a sophisticated perpetrator using generative AI can now vary deliberately to evade detection. Second, the explainability problems documented elsewhere in the AI-and-accounting literature (Kokina, Blanchette, Davenport, & Pachamanova, 2025) take on additional significance in this adversarial context, because a detection system whose reasoning cannot be fully inspected by its own operators is also a system whose vulnerabilities cannot be fully anticipated or patched before an adversary discovers and exploits them. Third, and most consequential for the governance discussion that follows, a co-evolutionary framing implies that fraud detection cannot be governed as a capital investment made once and then monitored passively; it requires the kind of continuous, adaptive oversight enterprise risk management frameworks were designed to support in principle (COSO, 2017), but which few organizations have historically applied with the intensity an actively evolving adversarial technology now requires.

A further consequence follows from treating capability, in Wolfe and Hermanson's (2004) sense, as shared infrastructure rather than a resource either side possesses exclusively. Because both detection systems and fraud-generation tools frequently draw on comparable underlying model architectures and, in some documented cases, overlapping training approaches, advances published or commercialized for legitimate detection purposes are not straightforwardly quarantined from adaptation by perpetrators, and advances in generative fidelity developed for entirely unrelated commercial purposes — voice synthesis for entertainment, image generation for design — have repeatedly proven adaptable to fraud with comparatively little additional engineering effort, as the deepfake and synthetic-identity evidence discussed above illustrates. This spillover dynamic means that an organization's fraud-security posture is partly hostage to the pace of AI development in domains entirely outside its own control or its own industry, a dependency classical enterprise risk management, built around risks an organization could reasonably expect to monitor within its own operating environment, was not designed to capture.

Financial Security and Risk Management Implications

From Detection to Enterprise Risk Management

The co-evolutionary framing developed above suggests that financial security in an AI-enabled environment cannot be reduced to detection-model accuracy, however well-evidenced that accuracy is in the specific studies reviewed earlier in this paper. Financial security, on this

account, is better understood as a property of the broader risk management system within which detection technology operates — a system that has to account explicitly for the possibility that its own detection tools may be circumvented by adversaries using comparable technology, rather than treating detection accuracy measured against historical data as a reliable proxy for security against future attacks. Zhang, Zhu, Dai, and Chen (2023), examining the ethical dimensions of AI in accounting more broadly, identify data security and misuse as recurring concerns that organizations deploying AI tend to treat as secondary to functional performance — an observation with direct application here, since a fraud detection system's own data and model architecture are themselves potential targets, not merely defensive tools, in an adversarial environment.

Explainability and the Evidentiary Requirements of Fraud Investigation

A further risk-management implication concerns the evidentiary status of AI-generated fraud findings once a case moves from internal detection toward investigation, disclosure, or litigation. Kokina et al. (2025) identify explainability as an unresolved problem in AI-enabled assurance work generally, and this problem carries particular weight in fraud contexts specifically, where a detection system's output frequently needs to withstand scrutiny from regulators, auditors, and courts, not merely inform an internal risk dashboard. Eulerich, Waddoups, Wagener, and Wood (2024), examining governance principles for automated systems more broadly, argue that control frameworks designed for manual or rule-based processes do not transfer cleanly to AI-enabled ones — a gap that, applied to fraud detection specifically, suggests organizations may be accumulating AI-flagged fraud indicators they cannot fully explain or defend, which is a meaningfully different and in some ways more precarious position than accumulating no indicators at all, since an unexplained flag ignored can itself become evidence of governance failure after the fact.

Regulatory and RegTech Responses

The financial-security implications developed above are not unfolding in a regulatory vacuum. Anagnostopoulos (2018), reviewing the emergence of regulatory technology as a distinct response to the broader fintech transformation of financial services, documents how AI-based systems are themselves increasingly used to automate compliance monitoring — scanning regulatory texts, tracking institutional activity against evolving legal requirements in real time, and reducing the manpower and error costs associated with manual regulatory reporting. This regulatory-technology layer adds a further wrinkle to the co-evolutionary model this paper develops, because it means AI is not simply the technology fraud detection and fraud perpetration both draw upon, but also the technology regulators themselves increasingly rely on to verify that financial institutions' own detection and reporting obligations are being met. Anagnostopoulos's analysis identifies governance as a persistent challenge running through this entire regulatory-technology literature, even as the technical capability of RegTech systems continues to improve — a pattern directly consistent with this paper's broader argument that AI's technical advances in the financial-security domain have consistently outpaced the governance structures meant to oversee them. For fraud detection specifically, the practical implication is that organizations face pressure from two directions simultaneously:

internal governance bodies expected to oversee AI-enabled detection systems whose reasoning those bodies may not fully understand, and external regulators increasingly equipped with their own AI-based tools capable of independently verifying whether an institution's fraud controls are functioning as represented.

Corporate Governance Implications

Board Oversight of AI-Enabled Fraud Risk

The agency-theoretic framing introduced earlier in this paper positions fraud detection as a monitoring mechanism boards and audit committees are ultimately responsible for overseeing, and the co-evolutionary dynamic discussed above changes what that oversight responsibility actually requires. Where fraud detection could once be treated, at the governance level, as a control implemented and periodically reviewed, the adversarial reality documented by Dhingra et al. (2026) implies that boards overseeing AI-enabled fraud detection are effectively overseeing a moving target, one whose defensive adequacy today provides limited assurance about its adequacy in eighteen months, given the pace at which generative fraud techniques are documented to be advancing. This has a direct implication for board composition and competency: agency theory's monitoring rationale for board oversight presumes the overseers possess sufficient understanding of what they are overseeing to exercise that oversight meaningfully, a presumption considerably harder to satisfy for AI-enabled, adversarially contested detection systems than for the largely static internal controls corporate governance frameworks were originally built to oversee.

This oversight gap is unlikely to close through generic AI literacy alone. A board member who understands, in general terms, how a machine-learning model works is not thereby equipped to evaluate whether a specific fraud-detection system's training data remains representative of current fraud patterns, whether its false-negative rate has drifted as perpetrators adapt their methods, or whether the vendor supplying the system has itself adequately hardened it against adversarial manipulation. Munoko et al. (2020) identify exactly this kind of specific, task-level accountability gap as a firm-level ethical risk in their broader analysis of AI in auditing, and the argument extends naturally to the board level: oversight competency for fraud-detection technology is not a generic skill acquired once, but a domain-specific and time-sensitive one that has to be actively maintained as both the detection technology and the threats it defends against continue to change. Boards that treat a single briefing on AI-enabled fraud controls as sufficient discharge of their oversight duty are, on this account, systematically underestimating what genuine oversight of a co-evolving adversarial system actually requires.

Redefining Agency Costs in an AI-Monitored Organization

Jensen and Meckling's (1976) original agency-cost framework treats monitoring cost and residual loss as the two components principals weigh in deciding how much monitoring to fund. The co-evolutionary dynamic this paper has developed suggests a third component deserves explicit theoretical recognition: the cost of monitoring the monitor — that is, the cost

of maintaining sufficient oversight of the AI detection system itself to ensure its outputs remain trustworthy as both detection and evasion technology continue to advance. Bou Reslan and Jabbour Al Maalouf (2024), examining AI adoption's broader effect on accounting work, find that adoption changes required skills rather than simply reducing headcount — a finding that, applied to the governance context here, suggests the skills boards and audit committees require to discharge fraud-oversight responsibilities are themselves shifting, from financial-statement literacy alone toward some working understanding of the AI systems now mediating a substantial share of the fraud-relevant evidence a board actually sees.

This third agency-cost component has a further property the original two-component framework does not anticipate: unlike conventional monitoring cost, which an organization can generally treat as a known and relatively stable line item, the cost of monitoring an AI-enabled detection system is itself a function of how fast the adversarial environment is changing, and is therefore neither fixed nor fully within the organization's control. A firm operating in an industry or jurisdiction experiencing a sudden surge in generative-AI-enabled fraud, of the kind Dhingra et al. (2026) document, faces a corresponding and largely exogenous increase in the cost of adequately monitoring its own detection systems, independent of any change in the firm's own risk appetite or governance practice. Agency theory has historically treated monitoring cost as a variable the principal controls through the scale of monitoring investment chosen; the model developed in this paper suggests that, for AI-enabled fraud detection specifically, monitoring cost also has a component driven by forces entirely external to the firm, a property with direct implications for how boards should budget for, and how researchers should model, the true cost of fraud oversight going forward.

Toward an Integrated Theoretical Model

Bringing the preceding sections together, this paper proposes that AI-enabled fraud detection be understood through three integrated layers rather than as a single technical capability. The capability layer, developed through the fraud triangle and fraud diamond (Cressey, 1953; Wolfe & Hermanson, 2004), explains why AI matters at the level of the individual fraud event: it expands the capability available to both perpetrators and defenders simultaneously, and classical fraud theory's own logic implies that whichever side's capability advances faster in a given period will determine whether a specific fraud attempt succeeds. The governance layer, developed through agency theory (Jensen & Meckling, 1976) and enterprise risk management (COSO, 2017), explains why AI matters at the organizational level: it changes the cost structure of monitoring, but only reduces net agency cost where the organization also invests in monitoring the monitor — maintaining the oversight competency needed to trust the detection system's own outputs. The adversarial layer, developed through the evidence on generative fraud and deepfake-enabled deception (Dhingra et al., 2026), explains why the first two layers cannot be treated as settled once implemented: because the underlying technology is shared across both sides of the fraud contest, both organizational capability and the governance structures overseeing it require continuous rather than periodic reassessment.

This three-layer model has a direct implication the more narrowly technical fraud-detection

literature does not fully draw out: an organization can simultaneously improve its measured fraud-detection accuracy, in the sense the empirical studies reviewed in this paper document, and see its actual financial security deteriorate, if the rate of improvement on the adversarial side outpaces the rate of improvement on the detection side and the organization's governance structures are not designed to track that relative rate. Financial security, on this account, is not a property that follows automatically from adopting better detection technology; it is an emergent property of the interaction between detection capability, adversarial capability, and the governance capacity to monitor both — which is precisely the integrated view this paper's three literatures, examined separately, do not individually supply.

The model also clarifies why the regulatory-technology developments discussed above (Anagnostopoulos, 2018) do not, on their own, resolve the governance challenge this paper identifies. RegTech systems automate compliance verification against existing regulatory requirements, but the co-evolutionary layer of this model implies that the requirements themselves are chasing a moving target, since regulators face the same lag between observing a new fraud technique and codifying a rule against it that individual organizations face in updating their own detection systems. Regulatory automation therefore compresses the time between a rule's issuance and its enforcement, which is valuable, but it does not compress the arguably more consequential time between a new fraud technique's emergence and a rule against it being written in the first place — a gap that remains fundamentally a governance and institutional-design problem rather than a purely technical one, and one this paper's model suggests deserves at least as much attention as detection accuracy itself.

Directions for Future Research

Several gaps follow from the theoretical model developed above. First, almost none of the structured-data and text-based fraud-detection literature reviewed in this paper (Bao et al., 2020; Bertomeu et al., 2021; Craja et al., 2020) has been tested specifically against adversarially generated fraud designed to evade machine-learning detection; research examining detection-model robustness against generative-AI-assisted fraud specifically, rather than against historical fraud patterns alone, would directly test this paper's central co-evolutionary claim.

Second, Goecks et al.'s (2022) finding that anti-money laundering and financial statement fraud research remain largely separate literatures despite conceptual overlap suggests an integration opportunity future research should pursue directly, particularly given that synthetic-identity and deepfake-enabled schemes documented by Dhingra et al. (2026) frequently implicate both categories simultaneously.

Third, the governance implications developed in this paper — particularly the claim that boards require new competencies to oversee AI-enabled, adversarially contested detection systems — remain theoretically derived rather than empirically tested; survey or case-study research examining what fraud-oversight competencies boards and audit committees currently possess, relative to what this paper's model suggests they require, would connect this paper's governance argument to observable practice.

Finally, the “cost of monitoring the monitor” this paper proposes as an extension to Jensen and Meckling's (1976) agency-cost framework has not, to the authors' knowledge, been operationalized or measured in any existing study; developing a method for estimating this cost component, and examining how it varies with the sophistication of an organization's AI-enabled detection infrastructure, would give the theoretical extension proposed here an empirical foundation it currently lacks.

Conclusion

Artificial intelligence has demonstrably improved organizations' capacity to detect financial fraud, and the evidence reviewed in this paper — spanning structured financial data, unstructured disclosure text, and transactional network analysis — leaves little doubt that AI-enabled detection substantially outperforms the methods it has replaced. Yet this paper has argued that evaluating AI's contribution to fraud detection solely on this evidence significantly understates what has actually changed, because the same technological developments improving detection are simultaneously available to, and increasingly used by, the perpetrators detection systems exist to catch. Integrating classical fraud theory, agency theory, and enterprise risk management with the emerging evidence on generative and deepfake-enabled fraud, this paper has proposed that AI-enabled fraud detection is best understood as a co-evolutionary governance problem rather than a static technical achievement — one in which an organization's measured detection accuracy and its actual financial security can diverge, and in which corporate governance structures built around the assumption of a stable detection technology are increasingly poorly matched to the adversarial reality AI has created.

None of this argues against continued investment in AI-enabled detection; the evidence reviewed throughout this paper is unambiguous that such investment produces real and measurable gains against the fraud patterns AI-enabled systems are trained to recognize. The argument, rather, is that the value of that investment cannot be assessed in isolation from the governance capacity surrounding it, and that treating detection accuracy as a sufficient proxy for financial security risks leaving organizations confident in defenses that a changing adversarial environment has already begun to erode. Boards, risk officers, and researchers who take seriously the co-evolutionary framing this paper proposes will find that the more demanding, and more useful, question is not whether an organization has adopted AI-enabled fraud detection, but whether it has built the governance capacity to keep asking, continuously, whether that detection remains adequate against a threat that is not standing still.

The practical and theoretical implication is the same: boards, risk managers, and researchers alike should stop treating AI-enabled fraud detection as a technology investment to be evaluated once against historical accuracy benchmarks, and start treating it as **an ongoing governance responsibility requiring continuous reassessment of the relative pace of detection and evasion capability**. Financial security in an AI-enabled environment is not a state an organization achieves; it is a relationship an organization has to actively sustain, against an adversary whose own capability is advancing on largely the same technological foundation the organization's defenses depend on.

References

- Anagnostopoulos, I. (2018). Fintech and regtech: Impact on regulators and banks. *Journal of Economics and Business*, 100, 7–25. <https://doi.org/10.1016/j.jeconbus.2018.07.003>
- Bao, Y., Ke, B., Li, B., Yu, Y. J., & Zhang, J. (2020). Detecting accounting fraud in publicly traded U.S. firms using a machine learning approach. *Journal of Accounting Research*, 58(1), 199–235. <https://doi.org/10.1111/1475-679X.12292>
- Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. (2021). Using machine learning to detect misstatements. *Review of Accounting Studies*, 26(2), 468–519. <https://doi.org/10.1007/s11142-020-09563-8>
- Bou Reslan, F., & Jabbour Al Maalouf, N. (2024). Assessing the transformative impact of AI adoption on efficiency, fraud detection, and skill dynamics in accounting practices. *Journal of Risk and Financial Management*, 17(12), Article 577. <https://doi.org/10.3390/jrfm17120577>
- Chen, Z., Van Khoa, L. D., Teoh, E. N., Nazir, A., Karuppiah, E. K., & Lam, K. S. (2018). Machine learning techniques for anti-money laundering (AML) solutions in suspicious transaction detection: A review. *Knowledge and Information Systems*, 57(2), 245–285. <https://doi.org/10.1007/s10115-017-1144-z>
- Committee of Sponsoring Organizations of the Treadway Commission. (2017). *Enterprise risk management—Integrating with strategy and performance*. COSO.
- Craja, P., Kim, A., & Lessmann, S. (2020). Deep learning for detecting financial statement fraud. *Decision Support Systems*, 139, Article 113421. <https://doi.org/10.1016/j.dss.2020.113421>
- Cressey, D. R. (1953). *Other people's money: A study in the social psychology of embezzlement*. Free Press.
- Dhingra, D., Chakraborty, P., & Sharma, K. (2026). Digital deceit in finance: Deepfakes, synthetic identities, and the future of fraud prevention. In M. V. Achim (Ed.), *Financial crime in the digital era* (Contributions to Finance and Accounting). Springer. https://doi.org/10.1007/978-3-032-20198-0_2
- Eulerich, M., Waddoups, N., Wagener, M., & Wood, D. A. (2024). Development of a framework of key internal control and governance principles for robotic process automation (RPA). *Journal of Information Systems*, 38(2), 29–49. <https://doi.org/10.2308/ISYS-2023-067>
- Goecks, L. S., Korzenowski, A. L., Terra Neto, P. G., de Souza, D. L., & Mareth, T. (2022). Anti-money laundering and financial fraud detection: A systematic literature review. *Intelligent Systems in Accounting, Finance and Management*, 29(2), 71–85. <https://doi.org/10.1002/isaf.1509>
- Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, 3(4), 305–360. [https://doi.org/10.1016/0304-405X\(76\)90026-X](https://doi.org/10.1016/0304-405X(76)90026-X)
- Kokina, J., Blanchette, S., Davenport, T. H., & Pachamanova, D. (2025). Challenges and opportunities for artificial intelligence in auditing: Evidence from the field. *International Journal of Accounting Information Systems*, 56, Article 100734.

<https://doi.org/10.1016/j.accinf.2025.100734>

Munoko, I., Brown-Liburd, H. L., & Vasarhelyi, M. (2020). The ethical implications of using artificial intelligence in auditing. *Journal of Business Ethics*, 167(2), 209–234.

<https://doi.org/10.1007/s10551-019-04407-1>

Wolfe, D. T., & Hermanson, D. R. (2004). The fraud diamond: Considering the four elements of fraud. *The CPA Journal*, 74(12), 38–42.

Zhang, C., Zhu, W., Dai, J., & Chen, X. (2023). Ethical impact of artificial intelligence in managerial accounting. *International Journal of Accounting Information Systems*, 49, Article 100619. <https://doi.org/10.1016/j.accinf.2023.100619>

Zhang, C., Zhu, W., Dai, J., Wu, Y., & Chen, X. (2025). Drivers and concerns of adopting artificial intelligence in managerial accounting. *Accounting & Finance*. Advance online publication. <https://doi.org/10.1111/acfi.13404>