

Beyond Severity Scores: A Conceptual Framework for Quantifying Cybersecurity Risk in Artificial Intelligence Systems

Shubham Paikrao

spaikrao@terpmail.umd.edu

University of Maryland

Research Article

Keywords: AI cybersecurity risk, risk quantification, FAIR extension, prompt injection, model drift, EU AI Act

Posted Date: August 13th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-9612663/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: No competing interests reported.

Abstract

Artificial intelligence is being deployed at organizational scale faster than security postures can adapt. Existing risk quantification frameworks, including the Common Vulnerability Scoring System (CVSS), Factor Analysis of Information Risk (FAIR), and machine learning based risk models, were designed for conventional IT environments and share a fundamental structural gap: none adequately accounts for the threat landscape specific to AI systems. Prompt injection, model drift, training data poisoning, and adversarial manipulation are not addressed by these frameworks in any meaningful way. Neither is the regulatory exposure accumulating under the EU AI Act, CMMC, and emerging AI governance requirements worldwide. This paper introduces a conceptual framework for assessing, communicating, and managing cybersecurity risk specific to AI systems. Instead of proposing an additional scoring formula, it identifies five dimensions of AI risk that current tools consistently fail to capture: Attack Surface, Model Exploitability, Data Integrity, Regulatory Exposure, and Control Maturity. The framework explains how these dimensions interact to produce financial loss exposure and is designed to serve multiple audiences simultaneously. It provides senior executives and boards with clarity that requires no background in security engineering, gives GRC professionals a structure they can operationalize within existing risk programs, and offers regulators a reference point for developing proportionate AI governance standards. This framework is not a finished actuarial product. It is scaffolding, sturdy enough to support consistent and comparable conversations about AI risk that current frameworks simply cannot enable. The field's most pressing gap is not insufficient threat intelligence but the absence of shared vocabulary for translating that intelligence into organizational decisions.

1. Introduction

Three incidents illustrate the problem. A bank's AI-powered loan officer recommends denials based on patterns its own human auditors cannot explain. A hospital's diagnostic model quietly degrades following a routine software update, producing subtly incorrect risk scores for months before the cause is identified. A retailer's customer-facing chatbot is manipulated into disclosing confidential pricing logic through adversarial user inputs. Each represents a cybersecurity failure. Not one would be caught, predicted, or measured by any risk framework currently in widespread use.

This is not a gap at the edges of cybersecurity practice. It sits at its center. Organizations are deploying AI systems considerably faster than they are building the governance structures required to secure them. AI-related fraud incidents surged 1,210% year-over-year as of 2025 ^[1], yet only 11% of enterprises have deployed security tools designed specifically to protect AI systems ^[1]. The frameworks that security professionals currently rely on, CVSS for vulnerability severity, FAIR for financial risk quantification, and compliance-based maturity models for governance, were designed for an environment in which the primary threats originated from software vulnerabilities, misconfigured infrastructure, and human error. They were not designed for an environment in which the risk surface includes a model's reasoning process, the integrity of its training data, its susceptibility to language-based manipulation, and its behavioral stability under distribution shift.

This paper proposes the AI Cyber Risk Framework, a conceptual structure for understanding, organizing, and communicating the cybersecurity risks specific to AI systems. It does not replace existing tools. CVSS continues to serve its designed purpose. FAIR remains the most rigorous publicly available approach to financial risk quantification. What this framework provides is a layer above these tools, supplying the vocabulary and structure organizations need to ask the right questions about AI risk before attempting to measure it. The framework is

designed to be actionable for a CISO briefing a board, usable by a GRC team constructing a risk register, and relevant to a regulator drafting AI governance requirements.

1.1 Why This Problem Matters Now

The timing of this gap matters more than it might appear. Three forces are converging simultaneously, and none are decelerating.

AI deployment is expanding into domains where failure carries material consequences. Healthcare triage, credit decisioning, critical infrastructure monitoring, and legal analysis are no longer experimental applications. They are operational. The failure modes of AI systems in these domains are categorically different from those of the software they replace. A misconfigured firewall produces an immediate, identifiable signal. A clinical AI model producing subtly incorrect outputs may go undetected for months before the anomaly is traced to the model.

Simultaneously, the regulatory environment is hardening. The EU AI Act came into force in 2024 with mandatory risk classifications and conformity requirements for high-risk AI systems ^[2]. NIST published its AI Risk Management Framework in 2023 ^[3]. The SEC now requires public companies to disclose material cybersecurity incidents within four business days, and AI system failures are increasingly recognized as falling within that definition ^[2, 3]. The compliance timeline organizations anticipated is compressing.

The insurance market faces a related structural problem. Current cyber policies were not written with AI-specific loss scenarios in mind, and the gap between what they cover and what organizations actually face is widening. As ISACA has documented, AI risks including model drift, bias, privacy breaches, and regulatory gaps are becoming embedded across business functions, yet most organizations lack the tools and expertise to assess this exposure, a gap that critically undermines their ability to make successful claims when coverage is needed ^[5]. Underwriters are making judgment calls in the absence of a shared measurement framework that would give them more reliable inputs.

Organizations that develop a rigorous internal approach to AI risk quantification now will be better positioned to satisfy regulators, negotiate insurance terms, and make defensible investment decisions. Those that delay will find themselves reacting to requirements they had the opportunity to anticipate.

1.2 Scope and Contributions

This paper makes four contributions. First, it diagnoses precisely why CVSS, FAIR, and ANN-based models fall short when applied to AI-specific risk, with sufficient specificity that practitioners can identify exactly where each framework breaks down. Second, it introduces five risk dimensions that any adequate AI risk framework must address and that existing tools consistently fail to capture. Third, it proposes a conceptual structure for integrating those dimensions into existing organizational risk management programs rather than requiring practitioners to rebuild from first principles. Fourth, it maps implementation across organizational levels, from board reporting to technical control mapping, so the framework is immediately operational rather than aspirationally useful.

2. Why Existing Frameworks Fall Short for AI Systems

Before proposing a new structure, precision is required about what existing frameworks accomplish and where, specifically, they fail. The gaps that AI systems expose are structural rather than superficial, and they are not

visible to practitioners working exclusively within existing tools.

2.1 CVSS: Measuring Severity, Not Risk

CVSS is the most widely deployed vulnerability scoring system in use today. Its Base Score quantifies the technical severity of a vulnerability across eight dimensions: attack vector, complexity, privileges required, user interaction, and impacts on confidentiality, integrity, and availability. This serves a genuine purpose: it provides the security community with a shared language for describing the abstract severity of a flaw.

The limitation is fundamental. FIRST, the governing body behind CVSS, states explicitly in the v4.0 User Guide that CVSS-B scores are designed to measure the severity of a vulnerability and should not be used alone to assess risk^[4]. Despite this, most organizations use CVSS Base Scores as the primary, and often exclusive, input to remediation prioritization. The consequence is well documented: a vulnerability scored 9.8 Critical in a non-production system with no external exposure may legitimately warrant less remediation urgency than a 5.0 Medium affecting a customer-facing system that processes financial transactions.

For AI systems, this limitation is compounding. Prompt injection does not reside in code; it exists in the interaction between a model's architecture and adversarial language inputs. CVSS provides no encoding for this. It cannot score model drift, in which a deployed system's behavior degrades incrementally without any single exploitable vulnerability materializing. It cannot represent the risk introduced when an organization exposes a large language model through APIs connected to live databases, email systems, or enterprise tooling. CVSS does not oversimplify AI risk. It does not capture it at all.

2.2 FAIR: The Right Structure, Missing the Right Inputs

FAIR is the most rigorous publicly available framework for cyber risk quantification. It addresses the correct question: what is the probable frequency and financial magnitude of a loss event? It answers through a structured decomposition of threat frequency, vulnerability, and loss magnitude, doing so probabilistically. Rather than producing a point estimate, FAIR uses Monte Carlo simulation to generate a range of outcomes, which more accurately represents the distributional nature of risk^[6]. For mature cybersecurity risk domains, it is the appropriate tool.

The challenge for AI systems is not FAIR's structure but the inputs that structure requires. FAIR's probabilistic engine depends on calibrated estimates of loss event frequency and magnitude. For established threat categories, data breaches, ransomware, insider theft, a decade of incident records, insurance claims, and regulatory disclosures has produced usable frequency and magnitude estimates. For AI-specific threats, this empirical foundation does not yet exist. The annualized rate of prompt injection attacks resulting in material financial loss, or of model drift producing decisions sufficient to trigger regulatory action, cannot currently be estimated with actuarial confidence because the incidents are too recent and too underreported^[7].

The FAIR Institute addressed this gap in 2024 through FAIR-AIR, which applies FAIR's probabilistic structure to generative AI risk scenarios^[8]. This is a meaningful contribution. However, FAIR-AIR operates within FAIR's existing ontology. It does not introduce new risk dimensions for AI-specific attack surfaces, and it cannot be empirically calibrated in the current absence of AI-specific loss data. The framework proposed in this paper extends FAIR-AIR's direction by naming the specific dimensions that must be captured and providing a structure for estimation before actuarial data matures.

2.3 ANN-Based and Statistical Models: Accuracy Without Accountability

A substantial body of research applies machine learning to cyber risk prediction, drawing on historical incident data, threat intelligence feeds, and network telemetry to forecast breach likelihood or severity [9]. These models can achieve high predictive accuracy. Their limitations for organizational risk management are structural rather than technical.

The first limitation is opacity. A neural network that predicts a 73% probability of a significant AI-related incident within the next twelve months cannot identify which factors drove the prediction in terms actionable by a CISO or interpretable by a board. The second limitation is that predictions are not decisions. Risk management requires not only a probability but a cost estimate: what will a given incident cost the organization, and how does that compare to the cost of prevention? Statistical models rarely produce financial outputs in usable form. The third limitation is data dependency: these models require the same actuarial training data that FAIR lacks for AI-specific threats. A model calibrated on historical breach data from conventional IT environments does not carry meaningful predictive validity for LLM behavior under adversarial manipulation.

2.4 Compliance Frameworks: Measuring Effort, Not Risk

NIST AI RMF, CMMC, ISO 27001, and the EU AI Act's conformity assessment requirements all provide governance structure: specified processes, controls, and documentation requirements. This is valuable. Compliance, however, is not risk measurement. An organization can achieve full CMMC Level 2 compliance while operating an AI system with uncontrolled prompt injection exposure, because CMMC was not designed to address LLM attack surfaces. Compliance framework scores reflect process maturity. They carry no information about the financial exposure an organization's AI systems currently represent or which risk dimension is the primary driver [10]. These are distinct questions, and compliance frameworks are not designed to answer them.

Table 1
Framework gap analysis: what existing tools cannot capture for AI systems

Framework	What It Measures	Key Gap for AI Systems	Financial Output?
CVSS	Technical vulnerability severity	Cannot encode prompt injection, model drift, or AI-specific attack surfaces	No
FAIR	Probabilistic financial risk	Lacks AI-specific ontology; inputs cannot be calibrated without AI loss history	Yes, but inputs are missing
ANN / Statistical Models	Breach likelihood prediction	Opaque outputs; no financial output; trained on non-AI incident data	Rarely
NIST AI RMF / CMMC	Governance process maturity	Measures compliance effort, not actual risk or financial exposure	No
EU AI Act	Regulatory risk classification	Categorizes risk class but does not quantify financial exposure within that class	No

3. The Framework: Five Dimensions of AI Cyber Risk

The framework organizes the cybersecurity risk landscape for AI systems into five dimensions. These dimensions do not operate independently: they interact, amplify one another, and in some configurations serve as compensating controls. The framework's purpose is not to produce a definitive numerical output but to provide organizations with a consistent structure for asking the right questions about AI risk, questions that current tools do not support.

3.1 Attack Surface: How Exposed Is the Model?

Every AI system presents an attack surface that is structurally distinct from that of traditional software. For a large language model, every interface that accepts input constitutes part of that surface: API endpoints, chat interfaces, file upload mechanisms, tool-use integrations, and retrieval-augmented generation pipelines drawing on external data. Each represents a potential entry point for prompt injection, the manipulation of a model's behavior through crafted inputs, which OWASP identifies as the leading vulnerability in LLM applications ^[11].

The attack surface is not a binary property. A model accessible exclusively to authenticated internal users through a monitored interface occupies a fundamentally different exposure position than one exposed to arbitrary public inputs through an API connected to live databases. Agentic systems, those capable of taking actions, invoking external tools, or spawning sub-agents, carry substantially larger attack surfaces than conversational systems. Research demonstrates that GPT-4 class agents are susceptible to indirect prompt injection in approximately 24% of cases under standard conditions, with browser-based agents partially compromised in up to 86% of cases ^[12]. These are not theoretical risk projections.

Assessment indicators: external API accessibility; agentic tool-use scope; multimodal input support; retrieval pipeline integration with external data sources.

3.2 Model Exploitability: How Manipulable Is the Model?

AI models are not uniformly susceptible to adversarial manipulation. The factors that determine exploitability differ considerably from those governing exploitability in traditional software. Relevant characteristics include whether a model is a closed proprietary system or an open-weight model, the rigor of adversarial testing to which it has been subjected, whether its system prompts and operational constraints are publicly known, and the organization's response latency to newly discovered jailbreak techniques.

OpenAI acknowledged in December 2024 that prompt injection is unlikely to be fully resolved because it is an architectural property of how language models process instructions rather than a correctable software defect ^[13]. This reframing carries operational significance. Model exploitability is not a problem to be eliminated; it is a persistent condition to be managed. Organizations that maintain systematic adversarial testing programs, harden system prompts, and implement input sanitization demonstrate measurably lower exploitability profiles than those that do not ^[14]. The performance gap between these two postures is material and continues to widen.

Assessment indicators: model transparency (closed versus open-weight); documented red-teaming frequency; public availability of system prompts; patch response latency for known jailbreak techniques.

3.3 Data Integrity Risk: How Trustworthy Is What the Model Learned?

An AI system is only as reliable as the data on which it was trained and the data it continues to receive in production. Data integrity risk encompasses three distinct failure modes not addressed by traditional security frameworks.

Training data poisoning is a deliberate attack in which adversarial data is introduced into a model's training set to produce systematically biased or manipulated outputs. Organizations fine-tuning models on proprietary datasets with weak provenance controls are particularly exposed to this vector.

Model drift refers to the gradual degradation in model performance as the real-world distribution of inputs diverges from the training distribution. Unlike a conventional software failure, drift does not produce error signals; it produces subtly incorrect outputs that may persist undetected for extended periods. Drift is statistically detectable and measurable using established monitoring methods^[15], but only where such monitoring has been implemented.

Hallucination risk denotes the tendency of language models to generate confident, internally coherent, factually incorrect outputs. In low-consequence contexts, this represents an inconvenience. In legal analysis, clinical decision support, or financial advisory applications, it constitutes direct liability exposure. NIST AI 600-1 designates hallucination as a core AI trustworthiness risk^[3].

Assessment indicators: training data provenance controls; production drift monitoring implementation; hallucination rate auditing in high-stakes deployment domains.

3.4 Regulatory Exposure: What Does Non-Compliance Actually Cost?

The regulatory environment for AI has transitioned from voluntary guidance to enforceable obligation. The EU AI Act classifies AI systems by risk category and imposes mandatory conformity assessments, incident reporting requirements, and transparency obligations on systems in the high-risk tier, encompassing credit decisioning, employment screening, critical infrastructure, and healthcare applications^[2]. Non-compliance carries penalties of up to 35 million euros or 7% of global annual turnover, whichever is the higher figure.

In the United States, CMMC Level 2 and 3 requirements under DFARS, effective 2025, apply to contractors whose systems store, process, or transmit Controlled Unclassified Information^[17]. The SEC's mandatory cybersecurity disclosure rules require public companies to report material incidents within four business days^[16], and AI system failures are increasingly interpreted as falling within the scope of this obligation.

Existing frameworks address this dimension inadequately. CVSS provides no mechanism for regulatory exposure. FAIR can incorporate regulatory fines as a secondary loss category, but only where the practitioner explicitly includes this element, and current practice suggests most do not. Compliance frameworks confirm whether an organization is compliant; they do not quantify the financial exposure associated with current gaps or identify which gaps carry the highest exposure within a specific deployment context.

3.5 Control Maturity: How Well Are the Risks Being Managed?

Control maturity functions as the counterweight to the preceding four dimensions. An organization with a large attack surface operating under rigorous monitoring, active adversarial testing, tight access controls, and practiced

incident response occupies a fundamentally different risk position than one with the same surface area and no controls in place. The raw exposure appears identical; the actual risk is not.

Control maturity for AI systems encompasses both traditional security controls adapted for AI deployment contexts and a set of AI-specific controls with no direct equivalent in conventional IT security practice. The AI-specific controls include systematic adversarial testing of production systems, input sanitization and output filtering for LLM applications, model validation and drift monitoring pipelines, data lineage tracking for training datasets, human-in-the-loop requirements for high-stakes decision processes, and explainability tooling enabling auditors to reconstruct the reasoning behind model outputs. Organizations that have implemented these controls demonstrate measurably lower residual risk profiles even where their attack surfaces are substantial [8, 14].

Assessment indicators: adversarial testing frequency; output monitoring coverage; access control granularity for AI system interfaces; human oversight requirements for consequential decisions.

Table 2
AI cyber risk dimensions: definitions, indicators, and assessment questions

Dimension	Definition	High-Risk Indicators	Key Assessment Question
Attack Surface	Total accessible interface through which adversarial inputs can reach the AI system	External API access; agentic tool use; multimodal inputs; RAG pipelines with external data	What are the complete input channels to this system, and what actions can the system take on behalf of an adversary?
Model Exploitability	Susceptibility of the model architecture to adversarial manipulation and instruction override	Black-box model; no red-teaming history; publicly known system prompts; slow patch response	How resistant is the model to adversarial instruction override, and how quickly are new techniques addressed?
Data Integrity Risk	Risk of loss from compromised training data, production drift, or model hallucination	Unvalidated training data provenance; no drift monitoring; high-stakes deployment without hallucination auditing	Can the organization verify what the model learned, whether it continues to behave as designed, and at what rate its outputs are incorrect?
Regulatory Exposure	Financial and operational risk from non-compliance with applicable AI governance frameworks	EU AI Act high-risk classification; CMMC applicability; SEC disclosure obligations	Which regulatory obligations apply to this system, and what is the financial exposure associated with current compliance gaps?
Control Maturity	Strength of implemented security and governance controls that reduce residual AI risk	No adversarial testing; no output monitoring; no human oversight for high-stakes decisions	Are AI-specific security controls implemented, or only conventional IT controls that do not address AI attack vectors?

4. How the Framework Works in Practice

4.1 A Structured Assessment Process, Not a Calculation

The framework does not produce a precise numerical score. The actuarial foundation required to make such a score defensible does not yet exist: no sufficient historical record of AI-specific loss events, frequencies, or magnitudes has accumulated to support outputs comparable in rigor to FAIR estimates for mature threat domains.

What the framework provides is a structured assessment process for conversations that most organizations are currently conducting without adequate structure. The five dimensions supply a consistent agenda for any AI risk review: attack surface, exploitability, data integrity, regulatory exposure, and control maturity, evaluated in that sequence. Each dimension carries associated indicators and directional assessments, high, medium, or low, that enable comparison across AI systems and longitudinal tracking over time, without requiring empirical data that does not yet exist.

4.2 Implementation Across Organizational Levels

The framework operates at three organizational levels simultaneously. Its application differs by level.

4.2.1 Board and Executive Level

At the board level, the framework enables four categories of discussion that are currently difficult to conduct with analytical rigor. Portfolio prioritization: which AI deployments carry the highest combined risk across the five dimensions, and is security investment allocated proportionally? Regulatory liability: what is the organization's actual exposure under the EU AI Act, CMMC, or SEC disclosure requirements, and is it being actively managed? Insurance adequacy: does existing cyber coverage address AI-specific loss scenarios, including regulatory fines and AI-enabled fraud? Materiality assessment: would an AI incident occurring today meet the threshold for mandatory SEC disclosure? The framework provides a consistent analytical structure for each of these questions rather than leaving them to ad hoc judgment.

4.2.2 GRC and Risk Professionals

At the GRC level, the five dimensions map directly onto standard risk register structure. Each AI system in an organization's inventory can be assessed against all five dimensions, with inherent risk ratings assigned per dimension and residual scores revised as controls are implemented. The framework integrates naturally with FAIR: as AI-specific loss data accumulates through incident reporting obligations, insurance claims, and regulatory disclosures, FAIR's probabilistic engine can be applied to this dimensional structure to produce calibrated financial outputs. The framework is designed as the analytical bridge between the current state, in which most organizations have no coherent AI risk structure, and the target state of quantified AI risk management.

4.2.3 Security Practitioners

At the practitioner level, each dimension maps to defined technical work. Attack surface reduction corresponds to API security hardening, access restriction, and agentic scope limitation. Exploitability reduction corresponds to adversarial testing programs, system prompt hardening, and input sanitization implementation. Data integrity management corresponds to provenance controls, drift monitoring pipelines, and hallucination rate auditing. Regulatory gap closure corresponds to CMMC control mapping and EU AI Act conformity documentation. Control maturity improvement corresponds to the AI-specific security controls described in Section 3.5. The framework provides practitioners with a direct line of sight between technical work performed and the organizational risk reduction that work produces.

4.3 Three Illustrative Scenarios

The following scenarios demonstrate how the five-dimensional structure produces differentiated risk profiles across distinct deployment contexts, enabling prioritization decisions that no existing framework supports.

Table 3
Illustrative risk profiles across three AI deployment scenarios

AI System	Attack Surface	Model Exploitability	Data Integrity Risk	Regulatory Exposure	Control Maturity	Risk Priority
Public-facing LLM customer service agent (bank), API-integrated with account lookup	HIGH: External access, API integrations, agentic scope	HIGH: Black-box, limited red-teaming, known jailbreaks unpatched	MEDIUM: Curated data, moderate hallucination controls	HIGH: EU AI Act, SEC disclosure, potential CMMC	MEDIUM: Monitoring present, adversarial testing limited	CRITICAL: Immediate investment required
Internal code review assistant (software firm), authenticated developers only	LOW: Internal, authenticated, limited scope	MEDIUM: Black-box, active red-team program	LOW: No proprietary training data, outputs monitored	LOW: No high-risk classification, no CMMC applicability	HIGH: Quarterly adversarial testing, strong monitoring	MANAGED: Current controls appropriate
AI-assisted medical triage (hospital), integrated with EHR system	MEDIUM: Internal access, EHR integration, sensitive data	MEDIUM: Proprietary fine-tuned model, limited adversarial testing	HIGH: High-stakes outputs, documented hallucination risk in clinical LLMs	CRITICAL: HIPAA, EU AI Act high-risk, NIST AI RMF gaps	LOW: Basic monitoring only, no adversarial controls	CRITICAL: Regulatory and patient safety exposure

These profiles address the investment prioritization question that boards and CISOs most require answered. The bank's customer-facing system and the hospital's triage system both reach Critical priority, but through different dimensional drivers. The bank's profile is dominated by attack surface and exploitability. The hospital's is dominated by data integrity and regulatory exposure. The overall priority designation is identical; the remediation roadmaps are not. An organization with dimensional visibility allocates resources differently, and more effectively, than one operating from a single aggregate score.

5. Relationship to Existing Frameworks

The framework is designed to complement existing tools rather than displace them. The nature of the relationship with each is relevant for practitioners integrating it into established programs.

With CVSS

CVSS retains its function for vulnerability severity tracking. When a CVE is identified in an AI system's underlying infrastructure, CVSS scores that severity appropriately. What this framework supplies is the five-dimensional risk context that CVSS cannot provide. A Critical CVSS score in an AI system with a limited attack surface and high control maturity may legitimately require less urgent remediation than a Medium score in a system with high exploitability and no monitoring. CVSS cannot make this distinction. This framework can.

With FAIR

The five dimensions map directly onto FAIR's ontological structure. Attack surface and model exploitability inform FAIR's threat event frequency estimates. Data integrity risk informs loss magnitude through both primary losses, incorrect decisions and operational disruption, and secondary losses, regulatory fines and reputational damage. Regulatory exposure informs FAIR's secondary stakeholder reaction loss categories. Control maturity reduces FAIR's vulnerability probability. As AI incident data accumulates through regulatory disclosure obligations, insurance claims, and industry reporting, this framework's dimensional assessments will supply the structured inputs FAIR requires to produce calibrated financial outputs for AI-specific scenarios [6, 8].

With compliance frameworks

Rather than treating NIST AI RMF, CMMC, and the EU AI Act as independent tracks, this framework incorporates them as inputs to the regulatory exposure dimension. An organization's compliance posture against these frameworks directly informs its regulatory exposure assessment. The framework then situates that compliance assessment within the broader risk profile. A CMMC gap in an organization with a limited external AI footprint and high control maturity represents a different risk than the identical gap in an organization operating public-facing AI systems with no adversarial testing program. Compliance frameworks cannot surface this distinction. This framework does.

6. Implications for Policy and Regulation

The framework has implications that extend beyond organizational risk management. The five dimensions offer regulators a structure for developing proportionate, risk-based AI governance requirements that move beyond binary high-risk/low-risk classification.

The EU AI Act's high-risk designation is a meaningful regulatory foundation, but it treats all high-risk systems uniformly regardless of their actual dimensional risk profile. A high-risk system with a limited attack surface, comprehensive controls, and low exploitability faces identical requirements to one with the opposite profile across every dimension. These represent materially different risk situations. A dimensional framework enables regulators to apply proportionate treatment: greater scrutiny for systems with elevated scores across multiple dimensions, and lighter-touch oversight for high-risk-classified systems demonstrating strong control maturity and limited external exposure^[2, 10]. This form of proportionality characterizes mature regulatory frameworks in other domains, and AI governance will require it.

The insurance market faces a structurally related problem. As ISACA has documented, AI risks including model drift, bias, and regulatory gaps are becoming embedded across business functions, yet the tools to assess this exposure systematically do not yet exist^[5]. Prompt injection leading to financial fraud, regulatory fines under the EU AI Act, and model drift producing mass incorrect decisions are not consistently addressed within standard cyber insurance terms. Underwriters are exercising discretionary judgment in the absence of structured measurement inputs. A shared dimensional framework provides insurers with a common basis for AI-specific coverage design and supplies the market with the foundation required for coherent actuarial product development. FAIR performed this function for quantitative cyber insurance underwriting as it applied to conventional threats. This framework is positioned to perform the equivalent function for AI-specific risk.

The most consequential gap in AI risk management is not a deficit of threat intelligence. It is the absence of shared analytical vocabulary. Regulators, insurers, boards, and security professionals are each attempting to

manage the same underlying risk using incompatible terminology, tools, and assumptions. This framework proposes a common analytical language that functions across all four audiences.

7. Limitations and Future Work

This paper argues for the necessity of the framework's conceptual structure. It does not assert that this structure is currently sufficient for full quantitative risk management. The limitations are as follows.

The most significant limitation is the absence of calibration data. The five dimensions can be assessed directionally, high, medium, or low, with reasonable inter-rater consistency at present. Converting directional assessments into calibrated financial outputs requires actuarial data on AI-specific loss event frequency and magnitude. This data does not yet exist in sufficient volume. This is not a design limitation; it is an accurate characterization of the current state of the field. FAIR encountered the same constraint in its first decade and resolved it progressively as incident reporting, insurance loss data, and regulatory disclosures accumulated. An analogous trajectory is anticipated for this framework.

The second limitation is dimensional weighting. The five dimensions do not contribute equally to risk across all deployment contexts. Regulatory exposure is the dominant dimension for healthcare AI systems. Attack surface dominates for customer-facing LLM deployments. The framework does not currently specify relative dimensional weighting for different deployment categories. Developing domain-specific weighting guidance, validated through empirical analysis of AI incidents across sectors, is the primary future research direction.

Additional future work includes: empirical validation of the five-dimensional structure against documented AI incidents; development of domain-specific assessment rubrics for each dimension; integration with emerging AI incident databases, including the MIT AI Risk Repository and CISA AI advisories, as these resources mature; and pilot implementation across financial services, healthcare, and government organizations to assess practical usability and identify gaps not currently addressed by the framework.

8. Conclusion

The risk measurement tools organizations currently employ were not designed for AI systems. Adapting them without structural modification does not address the problem; it produces a false assurance of coverage. CVSS measures severity in code and is blind to prompt injection. FAIR quantifies risk probabilistically but lacks the input data required for AI-specific threats. Compliance frameworks document process maturity but provide no information about what an AI incident would cost. The gap between what organizations need to know about AI risk and what existing frameworks can tell them is structural, not a matter of parameter recalibration.

This framework addresses that gap at the level at which it exists: the conceptual structure through which AI risk is understood and communicated. Defining five dimensions that must be assessed for any AI deployment, attack surface, model exploitability, data integrity, regulatory exposure, and control maturity, provides organizations with a consistent, multi-level structure that is immediately usable by boards, GRC professionals, and security practitioners without depending on actuarial data that does not yet exist.

The field's most important open question is whether AI-specific risk dimensions can be calibrated to sufficient precision to produce defensible financial outputs before regulators and insurers require them. This framework

provides the structured answer to that question's prerequisite. Calibration requires definition first. Providing that definition is the contribution of this paper.

Declarations

Funding

Open access funding provided by the University of Maryland, College Park. The author did not receive support from any organization for the submitted work.

Competing Interests

The author has no relevant financial or non-financial interests to disclose.

Author Contributions

Shubham Paikrao is the sole author of this work. The author contributed to the conceptualization, formal analysis, and writing of the original draft. The author read and approved the final manuscript.

Data Availability

No datasets were generated or analyzed during the current study. The framework presented is conceptual in nature. All cited sources are publicly available and referenced in the reference list.

Ethics Approval

No ethics approval was required for this conceptual framework paper.

AI-assisted tools were used for copy editing and grammatical refinement of human-generated text.

References

1. CyberSecStats: AI cybersecurity statistics 2026 (Q1 + Q2). <https://www.cybersecstats.com/ai-cybersecurity-statistics-2026-q1-q2/>
2. European Parliament: Regulation (EU) 2024/1689 — The AI Act. Off J. Eur. Union (2024). <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689>
3. National Institute of Standards and Technology: Artificial Intelligence Risk Management Framework Generative AI Profile: NIST AI 600-1. U.S. Dept of Commerce: (2024). <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf>
4. FIRST: Common Vulnerability Scoring System v4.0: User Guide. Forum of Incident Response and Security Teams: (2023). <https://www.first.org/cvss/user-guide>
5. ISACA, ISACA Now Blog: Cyber insurance in crisis with AI blind spots. (2025). <https://www.isaca.org/resources/news-and-trends/isaca-now-blog/2025/cyber-insurance-in-crisis-with-ai-blind-spots>
6. Freund, J., Jones, J.: Measuring and Managing Information Risk: A FAIR Approach. Elsevier/Butterworth-Heinemann, Waltham, MA (2014). <https://www.fairinstitute.org/>

7. Stødle, K., Flage, R., Guikema, S., Aven, T.: Artificial intelligence for risk analysis: a risk characterization perspective. *Risk Anal.* **45**, 738–751 (2025). <https://pmc.ncbi.nlm.nih.gov/articles/PMC12032382/>
8. Copeland, J.B., Lebo, J.: A FAIR Artificial Intelligence (AI) Cyber Risk Playbook. FAIR Institute (2024). <https://www.fairinstitute.org/blog/fair-artificial-intelligence-ai-cyber-risk-playbook>
9. ScienceDirect: A data-driven risk assessment of cybersecurity challenges posed by generative AI, *Digital*, **4**(1), (2024). <https://www.sciencedirect.com/science/article/pii/S2772662225000360>
10. National Institute of Standards and Technology: Cybersecurity Framework 2.0. U.S. Dept of Commerce: (2024). <https://doi.org/10.6028/NIST.CSWP.29>
11. OWASP: OWASP Top 10 for Large Language Model Applications, Security Project: : LLM01 Prompt Injection. OWASP Gen AI (2025). (2025). <https://genai.owasp.org/llmrisk/llm01-prompt-injection/>
12. Zhan, Q., Liang, Z., Ying, Z., Kang, D.: InjecAgent: benchmarking indirect prompt injections in tool-integrated LLM agents. In: *Findings of ACL 2024*, pp. 10471–10506 (2024). <https://doi.org/10.18653/v1/2024.findings-acl.624>
13. OpenAI: Statement on prompt injection: Reported in: Premai Blog, Prompt Injection Attacks in 2025 (2026). (2024). <https://blog.premai.io/prompt-injection-attacks-in-2025-vulnerabilities-exploits-and-how-to-defend/>
14. Chen, S., et al.: Struq: defending against prompt injection with structured queries. In: *Proc. 34th USENIX Security Symposium* (2025). <https://www.usenix.org/system/files/usenixsecurity25-chen-sizhe.pdf>
15. Hinder, F., Vaquet, V., Hammer, B.: One or two things we know about concept drift: a survey on monitoring in evolving environments. *Front. Artif. Intell.* **7**, 1330257 (2024). <https://doi.org/10.3389/frai.2024.1330257>
16. U.S. Securities and Exchange Commission: Cybersecurity Risk Management, Strategy, Governance, and Incident Disclosure. Final Rule, Release Nos. 33-11216; 34-97989 (July 26, 2023, effective December 18, 2023). <https://www.sec.gov/newsroom/speeches-statements/gerding-cybersecurity-disclosure-20231214>
17. U.S. Department of Defense, Defense Acquisition Regulations System: Defense Federal Acquisition Regulation Supplement: Assessing Contractor Implementation of Cybersecurity Requirements (DFARS Case 2019-D041). *Federal Register*, vol. 90, no. 173, pp. 43560–43577 (September 10, 2025). <https://www.federalregister.gov/documents/2025/09/10/2025-17359/defense-federal-acquisition-regulation-supplement-assessing-contractor-implementation-of>