Assessing the Effectiveness of Artificial Intelligence Cybersecurity Controls: Practitioner Perspectives on Risk Mitigation in AI Systems

Richard B Antwi

1 School of Computer and Information Sciences, University of the Cumberlands, Williamsburg, Kentucky.

Abstract — The rapid integration of artificial intelligence (AI) into organizational operations has intensified the need for effective cybersecurity controls. However, the extent to which existing security measures adequately protect AI systems against both conventional and AI-specific threats remains poorly understood from a practitioner's perspective. This qualitative study explores how AI professionals perceive the effectiveness of current cybersecurity strategies in mitigating security and privacy risks in AI systems. Semi-structured interviews were conducted with 12 AI and cybersecurity professionals across financial services, healthcare, cloud computing, and technology sectors. Thematic analysis of the data identified four principal dimensions of effectiveness perception: (1) foundational controls are viewed as effective for conventional threats but insufficient for AI-specific risks; (2) AI-specific defenses are recognized as necessary but remain immature in most organizational settings; (3) monitoring capabilities exhibit a persistent gap between infrastructure-level and model behavioral detection; and (4) organizational and governance factors significantly moderate technical control effectiveness. Participants consistently described a maturity gap between the pace of AI deployment and the sophistication of protective measures. These findings have important implications for practitioners, policymakers, and researchers working to strengthen AI security frameworks and governance structures.

Keywords — AI cybersecurity effectiveness, risk mitigation, AI security controls, practitioner perceptions, qualitative research, machine learning security, AI governance, data protection, adversarial threats, information security

I. INTRODUCTION

Artificial intelligence (AI) systems have become deeply embedded in critical organizational processes, from fraud detection and credit risk assessment in financial institutions to clinical decision support in healthcare. This rapid and widespread adoption has been accompanied by an expanding and increasingly sophisticated cybersecurity threat landscape. AI systems face not only the conventional security risks encountered by all digital systems but also a distinct set of AI-specific vulnerabilities rooted in the nature of machine learning architectures, training data dependencies, and inference mechanisms [1, 2].

Significant scholarly and practitioner attention has been devoted to identifying and cataloging these threats. Data poisoning, adversarial manipulation, model inversion, model extraction, and membership inference attacks represent well-documented categories of AI-specific threats with demonstrated real-world applicability [3, 4]. Frameworks such as the National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework (AI RMF) [5] and the MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS) [6] have been developed to guide organizations in managing these risks. Yet, despite this body of guidance, a fundamental question remains substantially underexplored: how effective are current cybersecurity controls in practice at mitigating security and privacy risks in AI systems?

This question matters for several reasons. The technical feasibility of a control does not guarantee operational effectiveness under real-world organizational constraints. Practitioners must navigate trade-offs among security stringency, model performance, deployment velocity, resource availability, and regulatory compliance. The perceived and actual effectiveness of controls shapes resource allocation decisions, governance priorities, and risk tolerance thresholds across organizations [7, 8].

This paper addresses this gap by reporting findings from a qualitative study guided by the Research Question (RQ): How do AI professionals perceive the effectiveness of these strategies in mitigating security and privacy risks? Drawing on semi-structured interviews with 12 AI and cybersecurity professionals, the study provides a rich, practitioner-grounded account of how current AI cybersecurity measures are evaluated in terms of their practical effectiveness, where they fall short, and what factors moderate their impact.

The findings contribute to both the scholarly understanding of AI security effectiveness and to practical guidance for organizations seeking to strengthen their AI risk mitigation postures.

II. LITERATURE REVIEW

A. The Threat Landscape for AI Systems

The vulnerability profile of AI systems extends substantially beyond that of traditional software systems. Whereas conventional cybersecurity threats primarily target the confidentiality, integrity, and availability of information assets through established attack vectors, AI systems introduce additional attack surfaces that target the logic and learning mechanisms of machine learning models [9, 10].

Adversarial attacks, first systematically described by Szegedy et al. [11] and subsequently operationalized for practical settings by Goodfellow et al. [12], demonstrate that carefully crafted perturbations to model inputs can cause dramatically incorrect predictions without triggering conventional security alerts. Data poisoning attacks corrupt training data to degrade model performance or introduce hidden backdoor behaviors [13, 14]. Model inversion attacks reconstruct sensitive attributes of training data from model outputs [15], while model extraction attacks approximate proprietary model functionality through systematic API querying [16]. Membership inference attacks determine whether specific records were included in a model's training dataset, posing privacy risks in regulated industries [17].

These threats have been demonstrated across diverse model types and application domains, including healthcare, financial services, and autonomous systems. The challenge they pose is compounded by the opacity of many machine learning models, which complicates both threat detection and post-incident forensic analysis [18].

B. Effectiveness of Current AI Security Controls

Research on the effectiveness of AI security controls has primarily taken a technical focus, evaluating specific defenses under controlled experimental conditions. Adversarial training, defensive distillation, and input preprocessing have been proposed as defenses against adversarial attacks, though their robustness under adaptive attackers has been repeatedly challenged [19, 20]. Certified defenses offering formal robustness guarantees exist for constrained settings but incur substantial computational costs that limit practical scalability [21].

Data sanitization and provenance tracking mechanisms have been proposed as primary defenses against data poisoning, with demonstrated effectiveness in controlled scenarios [22]. Differential privacy provides formal privacy guarantees against inference attacks, though its application reduces model utility through privacy-utility trade-offs that remain challenging to optimize [23]. Rate limiting, query monitoring, and access controls have been proposed as defenses against model extraction and are commonly deployed in practice [24].

While these technical evaluations are valuable, they rarely account for the organizational, operational, and cultural factors that determine how effectively controls perform in real deployment environments. Studies examining practitioner perceptions of AI security effectiveness are sparse. Surveys of cybersecurity professionals have documented widespread concerns about AI-specific threats alongside acknowledgment of inadequate preparedness [25, 26]. However, these studies typically rely on self-reported readiness measures rather than deep qualitative exploration of how professionals evaluate specific controls in context.

C. Organizational Factors Moderating Security Effectiveness

Security control effectiveness is recognized across the information security literature as substantially moderated by organizational factors including governance structures, resource availability, cross-functional collaboration, regulatory context, and security culture [27, 28]. Controls that are technically sound may underperform due to inconsistent implementation, inadequate monitoring, insufficient expertise, or poor alignment between security objectives and business priorities [29].

In the AI security context, these organizational moderators are amplified by the novelty of the domain and the persistence of silos between AI development and cybersecurity teams [30]. The rapid pace of AI deployment frequently outstrips the development of appropriate security controls and governance structures, creating environments in which organizations are perpetually catching up to their own risk exposure [31]. Understanding how practitioners perceive this dynamic is essential for developing actionable recommendations that account for real-world organizational constraints.

III. RESEARCH METHODOLOGY

A. Research Design

A qualitative research design was employed, using semi-structured interviews as the primary data collection method. This approach was selected as appropriate for the study's purpose of exploring how practitioners perceive and evaluate the effectiveness of AI cybersecurity controls. This is an inherently interpretive phenomenon shaped by professional experience, organizational context, and subjective judgment [32]. Qualitative inquiry enabled the capture of nuanced,

experience-based accounts that survey methods cannot adequately access.

B. Participant Profiles

Twelve AI and cybersecurity professionals participated in the study, recruited through purposive sampling to ensure representation of diverse functional roles and industry sectors relevant to AI system security. Participants held roles spanning AI security engineering, machine learning engineering, AI governance, data science, cloud security architecture, AI privacy, cybersecurity consulting, DevSecOps engineering, IT auditing, and executive security leadership. Table I summarizes participant profiles.

TABLE I. PARTICIPANT PROFILES

ID	Role	Experience	Industry
P1	Senior AI Security & Risk Analyst	7+ years	Financial Services (Banking)
P2	Lead ML Engineer (AI Risk & Security)	10+ years	FinTech / Digital Banking
P3	Senior AI Security Engineer / MLOps Specialist	9+ years	Cloud AI Platforms / SaaS
P4	Director of AI Governance & Compliance	12+ years	Banking / Financial Services
P5	Senior Data Scientist	8+ years	Healthcare Analytics / HealthTech
P6	AI Product Lead / Co-Founder	6+ years	AI SaaS Startup (FinTech)
P7	Senior IT Auditor / AI Risk Auditor	11+ years	Banking / Financial Services
P8	Cloud Security Architect (AI/ML Platforms)	10+ years	Enterprise Cloud / Big Tech
P9	AI Privacy & Responsible AI Lead	9+ years	Global Technology / Digital Platforms
P10	Senior AI Security Consultant	13+ years	Consulting (Multi-sector)
P11	Senior DevSecOps Engineer (AI/ML)	9+ years	Enterprise Technology / SaaS
P12	Chief Information Security Officer (CISO)	18+ years	Financial Services / Banking

C. Data Collection

Semi-structured interviews were conducted using a protocol organized around the study's main research question. Interview questions pertinent to the research question explored participant assessments of how well current cybersecurity measures mitigate AI-specific and general security risks, how effectiveness is measured within their organizations, where they perceive the most significant gaps in protection, and how organizational and governance factors influence security outcomes. Interviews were conducted virtually, audio-recorded with participant consent, and professionally transcribed. Duration ranged from approximately 45 to 75 minutes per interview.

D. Data Analysis

Thematic analysis was conducted following the six-phase process described by Braun and Clarke [33]: familiarization with data, generating initial codes, constructing themes, reviewing themes, defining and naming themes, and producing the report. Coding was conducted inductively from the data. Emerging codes were iteratively refined and clustered into higher-order themes. Analytical rigor was supported through member checking with a subset of participants, reflexive journaling, and systematic peer debriefing. Thematic saturation was assessed iteratively; no substantively new codes emerged after the tenth interview.

IV. FINDINGS

Thematic analysis produced four major themes describing how participants perceive the effectiveness of AI cybersecurity controls: (1) foundational controls are effective but insufficient; (2) AI-specific defenses are necessary but immature; (3) monitoring effectiveness is uneven; and (4) organizational and governance factors critically moderate effectiveness. Each theme is described below with illustrative participant quotations. Participant identifiers (P1–P12) are used throughout to protect individual confidentiality.

A. Theme 1: Foundational Controls Are Effective but Insufficient

The most consistent finding across participant accounts was that foundational cybersecurity controls including encryption, role-based access control (RBAC), multi-factor authentication (MFA), network segmentation, and Security Information and Event Management (SIEM)-based monitoring are perceived as effective for addressing conventional cybersecurity threats but materially insufficient when confronted with AI-specific risks.

P1 articulated this assessment: “They are effective for known risks but less effective for emerging threats. There are gaps, especially in AI-specific threat detection.” P2 echoed this: “I would describe current security measures as evolving but not fully mature. They are effective against known threats, but AI introduces new types of risks that existing controls are not always designed to handle.” P7, operating from an independent audit perspective, observed that compliance with established security standards does not translate to comprehensive AI security: “From an audit standpoint, one of the key observations is that organizations are often compliant with existing standards, but compliance does not necessarily equate to security especially in the context of AI systems.”

Participants consistently described conventional controls as addressing the 'envelope' around AI systems, protecting infrastructure and data transmission, while leaving the model lifecycle substantially unprotected. P12 framed this at the executive level: "Current measures are effective in addressing traditional cybersecurity risks, but less so when it comes to AI-specific threats. There is still a gap between the pace of AI innovation and the maturity of security practices." P9 identified a particularly consequential blind spot: "Organizations often assume that if data is encrypted and access is controlled, the system is secure but AI introduces new pathways for risk that are not fully addressed by these controls."

This perception was moderated by industry context. Participants from highly regulated financial services environments generally reported higher perceived effectiveness of foundational controls than participants from startup or early-stage technology environments, where P6 acknowledged: "I would describe them as 'good enough for now,' but not where we want them to be long-term." Table II summarizes the perceived effectiveness ratings across control types as reported by participants.

TABLE II. PARTICIPANT-REPORTED PERCEIVED EFFECTIVENESS OF SECURITY CONTROL CATEGORIES

Control Category	Effectiveness (Conventional)	Effectiveness (AI-Specific)	Representative Participant View
Encryption & Data Masking	High	Moderate	Effective for data-at-rest/transit protection; insufficient against model inference attacks (P9)
Access Control (RBAC/MFA)	High	Moderate	Strong baseline control; gaps in granularity for model/pipeline access (P8)
Network Segmentation	High	Moderate	Effective for infrastructure isolation; limited relevance to model-level threats (P3)
SIEM / Infrastructure Monitoring	High	Low–Moderate	Robust for system events; poor visibility into model behavioral anomalies (P7, P10)
Data Validation & Provenance	Moderate	Moderate	Meaningful protection against poisoning; implementation inconsistent (P2, P4)
Adversarial Testing	N/A	Low–Moderate	Recognized as necessary but tools and practices still maturing (P2, P3)
Model Behavioral Monitoring	N/A	Low–Moderate	Emerging practice; gaps in most organizations (P7, P10, P11)
Privacy-Preserving Techniques	N/A	Moderate	Effective when implemented; adoption limited by complexity and utility trade-offs (P9, P10)

B. Theme 2: AI-Specific Defenses Are Necessary but Immature

A second major theme in participant accounts concerned the state of AI-specific defensive capabilities. Participants universally acknowledged the necessity of defenses targeting AI-specific threats such as data poisoning, adversarial manipulation, model extraction, and privacy inference while consistently characterizing these defenses as underdeveloped relative to the sophistication of current threats.

1) Data Integrity Controls

Data validation and provenance tracking were generally regarded as moderately effective when rigorously implemented, but participants identified significant inconsistencies in practice. P2 observed: "For data poisoning, we implement strict validation checks and data provenance tracking to ensure that training data is trustworthy." However, P10 noted a persistent organizational gap: "A recurring issue I encounter is that data used for model training is not always subject to the same level of scrutiny as production data, which creates a hidden risk." P4 further observed that training data governance often receives insufficient regulatory and internal policy attention: "One recurring challenge is ensuring that data used for model training complies with privacy requirements."

2) Adversarial Defenses

Adversarial testing and model hardening were characterized as the least mature area of AI security practice. While participants described implementing adversarial testing in varying degrees, the perceived effectiveness of these defenses was consistently described as partial and context-dependent. P3 described the implementation approach: "We train models with adversarial examples to improve their robustness. However, I would say this is still an evolving area, and there is no one-size-fits-all solution." P2 noted the gap between testing capability and defense comprehensiveness: "We conduct testing using adversarial examples to evaluate how resilient our models are, but I would say this area is still developing. There are tools available, but they are not yet as mature as traditional cybersecurity tools."

P10, drawing on consulting experience across multiple organizations, identified a systemic gap between awareness and implementation: "In many cases, I find that organizations are aware of these risks conceptually, but they have not yet translated that awareness into concrete controls." This gap between conceptual understanding and operational implementation was a recurring theme across participant accounts, particularly for adversarial defenses.

3) Model Access and Anti-Extraction Controls

Controls designed to prevent model extraction and unauthorized model access, including API authentication, rate

limiting, query monitoring, and model watermarking, were perceived as effective against opportunistic attacks but vulnerable to determined adversaries with sufficient query budgets. P3 described implemented controls: "We restrict access to model endpoints and implement rate limiting, authentication, and monitoring to detect suspicious usage patterns. In some cases, we also use techniques like watermarking models to detect unauthorized use." P8 offered a candid assessment of their limitations: "Organizations tend to be strong in areas like network security and access control, but weaker in areas like model robustness and adversarial defense. There is a maturity gap between securing the infrastructure and securing the model itself."

4) Privacy-Preserving Techniques

Differential privacy, federated learning, and related privacy-preserving machine learning techniques were acknowledged by a subset of participants as meaningful protections against inference attacks. However, their adoption was reported as limited by implementation complexity and the privacy-utility trade-off. P9 described the tension: "A challenge we often face is ensuring that privacy protections do not compromise the effectiveness of the model, which requires careful design and testing." P10 confirmed that widespread adoption of these advanced techniques remains a future state: "Organizations are starting to adopt more advanced techniques such as differential privacy and federated learning, but these are still not widely implemented."

C. Theme 3: Monitoring Effectiveness Is Uneven

A third major theme concerned the perceived uneven effectiveness of monitoring capabilities across the two dimensions required for comprehensive AI system security: infrastructure-level monitoring and model behavioral monitoring.

Participants universally reported high perceived effectiveness of infrastructure and system-level monitoring, supported by mature SIEM platforms, behavioral analytics, intrusion detection systems, and cloud-native monitoring solutions. P1 described the monitoring stack: "We use a combination of SIEM systems, behavioral analytics platforms, and model monitoring tools." P8 noted the maturity of infrastructure monitoring relative to model monitoring: "Organizations are very good at detecting when a server goes down, but not necessarily when a model is being subtly manipulated or producing unreliable outputs."

Model behavioral monitoring encompassing detection of prediction drift, concept drift, anomalous input distributions, and potential adversarial probing was consistently characterized as significantly less mature and less effective. P7 identified this gap as a systematic finding across audited organizations: "A gap I frequently observe is that organizations monitor system uptime and performance very well, but they do not always have sufficient controls in place to monitor the integrity and behavior of the AI models themselves." P10 described the nature of the gap: "Organizations are very good at detecting when a server goes down, but not necessarily when a model is being subtly manipulated or producing unreliable outputs."

P2 recounted a concrete example that illustrated why model behavioral monitoring matters for security, not just performance: "One instance I recall was when a model began showing unexpected bias in its predictions due to changes in input data distribution. While it wasn't a direct security breach, it highlighted how monitoring is not just about detecting attacks but also about ensuring model integrity." P3 described a detected adversarial probing attempt: "In one case, we detected unusual input patterns that were consistent with adversarial probing; someone was systematically testing the model to understand its behavior. We were able to block those requests and adjust our defenses accordingly."

The effectiveness of monitoring was also described as heavily dependent on resource availability and organizational maturity. P11 observed that automated, integrated monitoring provides substantially more reliable protection than manual or ad hoc approaches: "One of the advantages of DevSecOps is that monitoring is continuous and integrated, rather than being an afterthought. Security that depends on manual enforcement is inherently fragile. Automation is what makes security scalable."

D. Theme 4: Organizational and Governance Factors Critically Moderate Effectiveness

The fourth and perhaps most overarching theme in participant accounts was that technical control effectiveness is substantially moderated and in many cases constrained by organizational and governance factors. Participants described a set of structural and cultural dynamics that determine whether technically sound controls translate into operational security outcomes.

1) The AI Development-Cybersecurity Team Gap

All 12 participants described a persistent gap between AI development teams focused on model performance and innovation and cybersecurity teams focused on risk mitigation. This misalignment was identified as a primary moderator of control effectiveness, creating conditions in which security measures are applied inconsistently or too late in the development lifecycle. P1 noted: "AI teams focus on performance, while cybersecurity focuses on risk. Misalignment often occurs during development and deployment stages." P4 described the governance-level

manifestation of this gap: "These teams often operate with different priorities and perspectives. Without proper alignment, this can lead to gaps in security coverage."

P7, from an audit perspective, observed that this structural misalignment produces systematic security coverage gaps: "AI development teams prioritize performance and innovation, while cybersecurity teams focus on risk mitigation. This misalignment can result in security controls being applied after the fact, rather than being integrated into the development process from the beginning." P10 identified the organizational implications: "Security is introduced too late in the development process, which makes it harder and more expensive to implement effectively. Without strong governance, these priorities can conflict."

2) Governance Maturity as an Effectiveness Multiplier

Participants described governance maturity as a multiplier of technical control effectiveness. Organizations with well-developed AI governance structures, including formal pre-deployment review processes, documented risk management procedures, and clear accountability frameworks, were consistently described as achieving substantially higher security effectiveness than those relying primarily on technical controls without structured governance. P4 articulated this principle: "Controls must be documented, enforced, and auditable. Policies alone are not enough; they must be supported by strong oversight and accountability." P12 framed governance as an enabling condition for technical effectiveness: "Governance is essential because it provides the structure and accountability needed to manage AI risks effectively."

P7 identified inconsistency in governance enforcement as a primary source of security gap across audited organizations: "Having a policy is one thing, but ensuring consistent implementation and enforcement across the organization is where many challenges arise." P9 connected governance maturity directly to trust outcomes: "Organizations that fail to prioritize privacy and security in their AI systems risk not only regulatory penalties, but also loss of user trust — which can be far more damaging in the long term."

3) Resource Constraints and Organizational Size

Resource availability was identified as a significant moderator of security control effectiveness, particularly for smaller organizations and startups. P6, representing a startup context, described the resource-constrained reality: "The biggest challenge is balancing speed and security. In a startup, there is constant pressure to move fast, release features, and stay competitive. Security can sometimes feel like a constraint, especially when resources are limited. We don't have large dedicated security teams, so we have to prioritize carefully." P6 acknowledged that current controls in that organizational context were "'good enough for now,' but not where we want them to be long-term."

P10 observed a systematic correlation between organizational maturity, which is often a function of resource availability and regulatory exposure, and security effectiveness across consulting clients: "Effectiveness is highly dependent on organizational maturity. In more mature environments, security measures are integrated into the AI lifecycle. In less mature environments, security is often reactive." P12 articulated the executive perspective on resource investment: "AI security must become a strategic priority, not just a technical concern."

4) Framework Limitations as a Structural Constraint

Participants identified limitations in current AI risk management frameworks as a structural constraint on security effectiveness. While the NIST AI RMF was widely referenced as a valuable organizing framework, its practical translation into technical controls was consistently described as challenging. P4 observed: "Many frameworks provide high-level guidance but lack practical implementation details. This can make it difficult for organizations to translate governance requirements into actionable controls." P3 noted that engineering teams are often left to develop custom implementations because frameworks lack technical specificity: "Frameworks provide good guidance, but they can be difficult to implement in practice. This forces engineering teams to rely on experimentation and custom implementations."

P11 argued that framework effectiveness depends on operationalization through automation: "Frameworks are useful, but they need to be implemented through automation to be effective. A policy that is not enforced through automation is difficult to sustain in a high-speed environment." P10 identified the resulting gap as one of the most consequential structural weaknesses in AI security practice: "There is a clear need for tools and frameworks that bridge the gap between high-level guidance and hands-on implementation."

V. DISCUSSION

A. The Effectiveness Paradox of Conventional Controls

The consistent finding that conventional cybersecurity controls are viewed as effective for conventional threats but insufficient for AI-specific risks points to what may be characterized as an effectiveness paradox in current AI security practice. Organizations have invested substantially in mature, well-understood security controls such as encryption, access management, network security, and SIEM-based monitoring and these controls do provide genuine protection

against the broad spectrum of conventional cyber threats. However, they leave AI systems exposed at the model lifecycle level, where the most distinctive and consequential AI-specific vulnerabilities reside.

This finding is consistent with technical literature documenting that adversarial attacks, data poisoning, and inference attacks can succeed against systems with strong conventional security perimeters because these threats bypass infrastructure-level defenses entirely [9, 19]. The practical implication is that organizations cannot assess their AI security posture on the basis of conventional security maturity alone. A separate, AI-specific security evaluation framework is required, a need that current frameworks partially address but do not fully operationalize [5, 6].

B. The Maturity Gap in AI-Specific Defenses

The characterization of AI-specific defenses as necessary but immature reflects a documented gap in the AI security field between the sophistication of proposed technical defenses and their practical implementation at scale. Adversarial training, certified robustness, differential privacy, and model watermarking are all technically available, but practitioners consistently describe implementation barriers including computational cost, reduced model accuracy, tool immaturity, and lack of standardized guidance [19, 21, 23].

The observation by P10 that organizations are often "aware of these risks conceptually, but they have not yet translated that awareness into concrete controls" identifies a specific type of organizational failure that is distinct from general lack of awareness. It suggests that awareness-raising initiatives, while necessary, are insufficient without corresponding investment in practical tooling, implementation guides, and professional training that bridges the gap between conceptual understanding and operational deployment.

C. Implications of the Monitoring Effectiveness Gap

The finding that infrastructure monitoring is perceived as effective while model behavioral monitoring is consistently identified as a gap has significant implications for how AI security incidents are detected and responded to. Many AI-specific threats, including gradual data poisoning, adversarial probing, and subtle model drift induced by distribution shift, may operate undetected for extended periods if organizations rely solely on infrastructure-level monitoring [34, 35].

The incident recounted by P3, involving the detection of adversarial probing through unusual input pattern analysis, illustrates the value of model behavioral monitoring for security purposes. The observation by P7 that this capability is consistently absent in audited organizations suggests that model behavioral monitoring represents a critical capability gap that current security operations tooling does not adequately support. Developing and deploying production-ready model behavioral monitoring that integrates with existing security operations centers represents an important research and engineering priority.

D. Governance as the Multiplier of Technical Effectiveness

Perhaps the most policy-relevant finding of this study is the consistent characterization of governance maturity as a multiplier of technical control effectiveness. Technical controls implemented within organizations with strong governance structures, clear accountability, documented processes, formal review cycles, and cross-functional collaboration mechanisms were described as substantially more effective than identical controls implemented in governance-deficient environments.

This finding is consistent with the broader information security governance literature, which documents that technical controls are necessary but not sufficient conditions for effective security outcomes [27, 36]. In the AI security context, where the novelty of threats and the complexity of systems create significant organizational uncertainty, governance structures that provide clear roles, responsibilities, and enforcement mechanisms are particularly critical. The persistent gap between AI development and cybersecurity teams identified across all participant accounts represents a structural governance failure that technical controls alone cannot remedy.

These findings suggest that AI security improvement programs that focus exclusively on technical controls, deploying new tools, implementing new algorithms, or adopting new frameworks, are likely to achieve only partial gains in the absence of corresponding governance improvements. Effective AI security requires simultaneous investment in technical capabilities, organizational structures, and governance frameworks.

E. Sector and Size Moderation of Effectiveness

The variation in perceived effectiveness across industry sectors and organizational sizes documented in this study suggests that AI security effectiveness is not a uniform organizational property but rather a context-dependent outcome shaped by regulatory environment, resource availability, and organizational maturity. Financial services organizations, subject to the most demanding regulatory scrutiny, generally reported higher perceived effectiveness of foundational controls than organizations in less regulated sectors or startup contexts.

This variation has important implications for policy and framework development. Generic AI security frameworks that do not account for organizational size, resource constraints, and sector-specific regulatory contexts may be of limited

practical utility for the majority of organizations deploying AI systems. Tiered or scaled guidance that provides appropriately calibrated recommendations for organizations at different maturity levels represents an important gap in current AI security framework development.

VI. CONCLUSION

This study examined how AI professionals perceive the effectiveness of cybersecurity controls in mitigating security and privacy risks in AI systems. Drawing on semi-structured interviews with 12 practitioners across diverse organizational contexts, four principal themes were identified. Foundational cybersecurity controls are widely regarded as effective against conventional threats but are materially insufficient to address AI-specific risks. AI-specific defenses are recognized as necessary but remain immature in most organizational settings, with adversarial defense being particularly underdeveloped. Monitoring effectiveness is uneven, with strong infrastructure-level capabilities and persistent gaps in model behavioral monitoring. Organizational and governance factors, including team alignment, governance maturity, resource availability, and framework quality, critically moderate the effectiveness of technical controls.

These findings make several contributions to research and practice. Empirically, the study provides practitioner-grounded evidence of how AI cybersecurity control effectiveness is perceived across diverse organizational contexts, contributing nuance to a debate that has been dominated by technical evaluation studies. Conceptually, the study advances the argument that AI security effectiveness cannot be reduced to technical control deployment but must be understood as a complex function of technical, organizational, and governance factors operating in interaction.

For practitioners, the findings underscore the importance of developing AI-specific security evaluation frameworks alongside investment in model behavioral monitoring capabilities and cross-functional governance mechanisms. For policymakers and framework developers, the consistent reports of implementation guidance deficits in current frameworks and the resulting gap between framework adoption and security effectiveness point to a critical need for more actionable, tiered, and technically specific guidance.

Limitations of the study include its qualitative design, which prioritizes depth over generalizability, and the concentration of participants in financial services and technology sectors. Future research should include quantitative assessments of control effectiveness across larger samples, longitudinal studies tracking effectiveness changes as AI security practices mature, and comparative analyses examining how effectiveness perceptions vary across regulatory environments and national contexts.

VII. CONFLICT OF INTEREST

The author declares no conflict of interest in relation to this research.

VIII. ACKNOWLEDGMENT

The author sincerely thanks all 12 research participants for their time and the depth of insight they contributed to this study. Their practitioner perspectives are the foundation of this work.

IX. AUTHOR'S BIOGRAPHY

Richard Antwi is a doctoral student at the University of the Cumberlands. His research interests include AI cybersecurity, risk management, and information security governance.

X. REFERENCES

[1] S. Russell, P. Norvig, Artificial Intelligence: A Modern Approach, 4th ed., Pearson, Upper Saddle River, NJ, 2020.

[2] M. Taddeo, T. McCutcheon, L. Floridi, "Trusting artificial intelligence in cybersecurity is a double-edged sword", Nature Machine Intelligence, 2019, 1, 557–560.

[3] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Madry, B. Li, T. Goldstein, "Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses", IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 45, 1563–1580.

[4] B. Biggio, F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning", Pattern Recognition, 2018, 84, 317–331.

[5] National Institute of Standards and Technology, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST, Gaithersburg, MD, 2023. https://airc.nist.gov/RMF

[6] MITRE, MITRE ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems, 2023. https://atlas.mitre.org

[7] Z.A. Soomro, M.H. Shah, J. Ahmed, "Information security management needs more holistic approach: A literature review", International Journal of Information Management, 2016, 36, 215–225.

[8] R.N. Rajapakse, M. Zahedi, M.A. Babar, H. Shen, "Challenges and solutions when adopting DevSecOps: A systematic review", Information and Software Technology, 2022, 141, 106700.

[9] L. Huang, A.D. Joseph, B. Nelson, B.I. Rubinstein, J.D. Tygar, "Adversarial machine learning", Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence, Chicago, IL, USA, October 2011, pp. 43–58.

[10] S.A. Seshia, D. Sadigh, S.S. Sastry, "Toward verified artificial intelligence", Communications of the ACM, 2022, 65, 46–55.

[11] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, "Intriguing properties of neural networks", Proceedings of the International Conference on Learning Representations, Banff, Canada, April 2014.

[12] I.J. Goodfellow, J. Shlens, C. Szegedy, "Explaining and harnessing adversarial examples", Proceedings of the International Conference on Learning Representations, San Diego, CA, May 2015.

[13] X. Chen, C. Liu, B. Li, K. Lu, D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning", arXiv:1712.05526, 2017.

[14] A. Shafahi, W.R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, T. Goldstein, "Poison frogs! Targeted clean-label poisoning attacks on neural networks", Advances in Neural Information Processing Systems, 2018, 31.

[15] M. Fredrikson, S. Jha, T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures", Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security, Denver, CO, October 2015, pp. 1322–1333.

[16] F. Tramer, F. Zhang, A. Juels, M.K. Reiter, T. Ristenpart, "Stealing machine learning models via prediction APIs", Proceedings of the USENIX Security Symposium, Austin, TX, August 2016, pp. 601–618.

[17] R. Shokri, M. Stronati, C. Song, V. Shmatikov, "Membership inference attacks against machine learning models", Proceedings of the IEEE Symposium on Security and Privacy, San Jose, CA, May 2017, pp. 3–18.

[18] A.B. Arrieta, N. Diaz-Rodriguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", Information Fusion, 2020, 58, 82–115.

[19] A. Athalye, N. Carlini, D. Wagner, "Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples", Proceedings of the International Conference on Machine Learning, Stockholm, Sweden, July 2018, pp. 274–283.

[20] N. Carlini, D. Wagner, "Towards evaluating the robustness of neural networks", Proceedings of the IEEE Symposium on Security and Privacy, San Jose, CA, May 2017, pp. 39–57.

[21] J. Cohen, E. Rosenfeld, Z. Kolter, "Certified adversarial robustness via randomized smoothing", Proceedings of the International Conference on Machine Learning, Long Beach, CA, June 2019, pp. 1310–1320.

[22] J. Steinhardt, P.W. Koh, P.S. Liang, "Certified defenses for data poisoning attacks", Advances in Neural Information Processing Systems, 2017, 30.

[23] C. Dwork, A. Roth, "The algorithmic foundations of differential privacy", Foundations and Trends in Theoretical Computer Science, 2014, 9, 211–407.

[24] P. Juuti, S. Szyller, S. Marchal, N. Asokan, "PRADA: Protecting against DNN model stealing attacks", Proceedings of the IEEE European Symposium on Security and Privacy, Stockholm, Sweden, June 2019, pp. 512–527.

[25] Verizon, Data Breach Investigations Report 2024, Verizon Communications Inc., New York, 2024.

[26] IBM Security, Cost of a Data Breach Report 2024, IBM, Armonk, NY, 2024.

[27] H.A. Kruger, W.D. Kearney, "A prototype for assessing information security awareness", Computers & Security, 2006, 25, 289–296.

[28] R. Von Solms, J. Van Niekerk, "From information security to cyber security", Computers & Security, 2013, 38, 97–102.

[29] K.J. Knapp, T.E. Marshall, R.K. Rainer, F.N. Ford, "Information security: Management's effect on culture and policy", Information Management & Computer Security, 2006, 14, 24–36.

[30] S. Ransbotham, D. Kiron, "Analytics as a source of business innovation", MIT Sloan Management Review, 2017, 58, 1–6.

[31] M. Coeckelbergh, AI Ethics, MIT Press, Cambridge, MA, 2020.

[32] J.W. Creswell, J.D. Creswell, Research Design: Qualitative, Quantitative, and Mixed Methods Approaches, 5th ed., SAGE Publications, Thousand Oaks, CA, 2018.

[33] V. Braun, V. Clarke, "Using thematic analysis in psychology", Qualitative Research in Psychology, 2006, 3, 77–101.

[34] T. Dietterich, E. Kong, "Machine learning bias, statistical bias, and statistical variance of decision tree algorithms", Technical Report, Oregon State University, 1995.

[35] J. Lu, A. Liu, F. Dong, F. Gu, J. Gama, G. Zhang, "Learning under concept drift: A review", IEEE Transactions on Knowledge and Data Engineering, 2019, 31, 2346–2363.

[36] I.D. Raji, A. Smart, R.N. White, M. Mitchell, T. Gebru, B. Hutchinson, J. Smith-Loud, D. Theron, P. Barnes, "Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing", Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, Barcelona, January 2020, pp. 33–44.