IADIS International Journal on Computer Science and Information Systems
Vol. 21, No. 1, pp. 53-68
ISSN: 1646-3692

STRENGTHENING ORGANISATIONAL DEFENCE: A GOVERNANCE FRAMEWORK AGAINST AI-DRIVEN CYBER THREATS

Petra Maria Asprion[1], Bettina Schneider[1] and Luca Knecht[2]
[1]*University of Applied Sciences and Arts Northwestern Switzerland FHNW, Basel, Switzerland*
[2]*ACE project services ag, Bern, Switzerland*

ABSTRACT

Artificial Intelligence (AI) is reshaping both economic activity and organisational operations, but its misuse has helped to amplify the scale and sophistication of cybercrime. Criminal actors increasingly employ AI to automate attacks, craft convincing social engineering campaigns, and exploit system vulnerabilities, creating risks that many existing cybersecurity frameworks are not designed to address. This research investigates how AI enables such malicious activities by identifying and categorizing the key criminal elements associated with AI-driven threats. The study used a qualitative design combining a structured literature review with semi-structured expert interviews to ensure methodological rigor and practical relevance. Using the findings from these methods, a governance-oriented framework for strengthening organizational defence has been developed. The framework presents clear, actionable measures that help organisations evaluate vulnerabilities, improve preparedness, and establish governance mechanisms suited to AI-related risk. Expert validation hast demonstrated its practical relevance for cybersecurity teams, Information Technology (IT) managers, and risk officers. By offering structured guidance tailored to medium- and large-sized organisations, these results contribute to ongoing policy discussions by offering a systematic, practice-aligned approach to strengthening organisational defence against emerging AI-enabled threats.

KEYWORDS

AI Governance, AI-Driven Cyber Threats, Cybersecurity Frameworks, Organisational Defence, Governance Framework, Risk Management

1. INTRODUCTION

Artificial Intelligence (AI) has rapidly evolved into a transformative technology, frequently compared to the industrial revolution for its far-reaching societal and economic impact (Adewale & Segun, 2024; Makridakis, 2017; Mohamed, 2023). Advances in core AI techniques,

including machine learning (ML), deep learning (DL), symbolic reasoning, evolutionary computation, and probabilistic modelling have generated substantial benefits across multiple sectors; however, these same capabilities increasingly facilitate opportunities for criminal misuse (Blauth, Gstrein, & Zwitter, 2022; Caldwell et al., 2020; Malatji, 2023; Rawat et al., 2023).

The dual-use nature of AI and its ability to support legitimate innovation while also facilitating malicious activity has contributed to a marked escalation in the scale, speed, and sophistication of cyberattacks (Blauth et al., 2022; Obioha-Val et al., 2025). Cybercriminals employ AI for e.g., automated phishing, deepfake-based impersonation, autonomous exploitation, and data manipulation, significantly amplifying offensive capabilities (Bueermann & Rohrs, 2024; Brundage et al., 2018; Loh et al., 2024).

The World Economic Forum's Global Risk Report 2024 identifies AI-generated misinformation and disinformation as the second-highest global risk, citing implications for political stability and public safety (Bueermann & Rohrs, 2024). Cybercrime overall increased by an estimated 600% between 2020 and 2023 (Rao et al., 2023). As AI enables more scalable and adaptive attacks, organisations face growing difficulty in detecting and mitigating AI-enhanced threats (Malatji & Tolah, 2024). Traditional and established cybersecurity frameworks are excellent tools, but they have shortcomings when it comes to offering timely mitigation controls for AI-driven attack automation and the rapid evolution of threats (ENISA, 2023a; Melaku, 2023, Mohamed, 2023).

Human vulnerabilities further exacerbate exposure: AI-powered social engineering exploits cognitive biases, while persistent workforce shortage limits organisations' ability to respond effectively (Melaku, 2023). In parallel, privacy risks intensify as AI systems rely on extensive datasets, with anonymised data increasingly susceptible to inference, reconstruction, and membership attacks (Admass, Munaye & Diro, 2024; He et al., 2025; Liu et al., 2025). Regulatory and governance gaps compound these risks (Feretzakis et al., 2024; Murdoch, 2021; Radanliev, 2025). The lack of coordinated global standards allows cross-border misuse, while the speed of AI innovation outpaces existing legal, ethical, and security frameworks (Evang, 2022; Yeung, 2024).

Major standards such as ISO/IEC 27001 (2023a) and the NIST CSF (2024) provide essential baselines but lack granularity for AI-specific threat vectors because they are lagging rapidly emerging phenomena, as the update cycle is always conservatively slow but accurate (Evang, 2022; Malatji, 2023).

At the same time, the fast-evolving tactics of threat actors hinder the development of a stable taxonomy of cybercriminal behaviour, limiting coordinated responses across jurisdictions (ENISA, 2023a). These developments underscore the urgent need for organisations to deepen their understanding of AI-enabled risks (Admass et al., 2024; Maurya, 2023).

This research addresses these challenges by examining the criminal elements associated with malicious uses of AI and developing a conceptual governance framework to support organisational defence. The study is guided by the following research questions (RQs):

RQ1: Which criminal elements characterise the malicious use of AI?
RQ2: How can these elements be systematically categorised?
RQ3: What requirements are necessary for developing a conceptual governance framework for AI-driven threats?
RQ4: How can these requirements be modelled into a coherent framework?
RQ5: To what extent does the resulting framework demonstrate practical utility?

2. LITERATURE REVIEW

2.1 Foundations of AI

AI has become a pivotal technological force, reshaping the global cyber-threat landscape through rapid advances in ML, DL, NLP and other emerging technologies and features and encompasses a broad spectrum of computational techniques (e.g., Khan et al., 2025; Xu et al., 2021). Contemporary literature identifies AI as a general-purpose technology that penetrates nearly all sectors of society, amplifying both opportunities and risks (Adewale & Segun, 2024; Mohamed, 2023). These developments have profound cybersecurity implications, as AI systems introduce new attack surfaces, accelerate adversarial capabilities, and enable large-scale exploitation of digital infrastructures (Brundage et al., 2018; King et al., 2020). In addition, the dual-use characteristics of AI—its potential for beneficial and malicious applications—make understanding AI-driven (cyber)crime essential for organisational defence (Malatji, 2023). The rapid increase of AI tools, techniques and methods increase exposure to strategic vulnerabilities, including model extraction, adversarial inputs, and data dependency risks (Liu et al., 2025).

2.2 Malicious Applications and Criminal Opportunities

Literature identifies core elements that characterize malicious and high-risk AI applications. First, AI enhances deception techniques—such as deepfakes, synthetic speech, and manipulated images produced through generative adversarial networks (GANs)—commonly used in fraud, harassment, and disinformation (Zeng, 2022). Second, AI enables autonomous exploitation, including automated phishing, vulnerability scanning, credential harvesting, and malware generation that surpass human operational speed (Bueermann & Rohrs, 2024; Brundage et al., 2018). Third, AI models themselves are vulnerable to inversion, data poisoning, and adversarial perturbations, undermining system integrity (ENISA, 2023b).

Privacy invasion forms another major risk element: AI-driven analytics support large-scale surveillance, re-identification of anonymized data, and inference of sensitive attributes (Admass et al., 2024). Identity-related crimes—such as biometric spoofing, face manipulation, and synthetic identity creation—are increasingly facilitated by AI (Liu et al., 2025). From an organisational perspective, AI-powered manipulation of trust, authority, and social cues has further escalated the sophistication of social engineering attacks (Hadnagy, 2018).

2.3 Types of Malicious Actors in Cybercrime

Cybercrime literature describes a wide array of motivations, modalities, and actor types involved in malicious activity (Golop and Săvulescu, 2024; Martineau et al., 2023; Pawlicka et al., 2021). Motivations range from financial gain and political influence on espionage, activism, and strategic destabilization (Rao et al., 2023). Technological affordances have enabled criminals to conduct attacks with unprecedented scale, precision, and anonymity. AI-powered malware uses for example polymorphism, automated obfuscation, and adaptive evasion to bypass traditional detection mechanisms (Brundage et al., 2018).

Threat actors operate across a spectrum of capability and intent, including so-called script kiddies, organized crime groups, insider threats, and state-sponsored adversaries (Bueermann & Rohrs, 2024). State-level actors increasingly deploy AI for automated reconnaissance, advanced persistent threats, disinformation operations, and critical infrastructure disruption (ENISA, 2023b; OECD, 2019). AI-driven threats also exploit vulnerabilities in interconnected systems, digital supply chains, and cloud infrastructures (MITRE ATT&CK, 2024). The literature consistently highlights that AI amplifies existing cybercrime vectors while introducing wholly new attack modalities that exploit model-specific weaknesses (Blauth, et al., 2022; Kaloudi & Li, 2020; Kazimierczak, et al., 2024; Liu et al., 2025).

2.4 Defensive Applications of AI in Cybersecurity

Defensive applications of AI span technical, operational, and strategic domains (Afghani, 2025; Fard, Selmic & Khorasani, 2023). On the strategic level, AI governance must incorporate principles such as transparency, accountability, fairness, privacy, and explainability, as emphasized by OECD (2019) or by NIST AI RMF (2023). Organisational resilience frameworks call for integration of AI risk management into enterprise security strategies, incorporating regulatory requirements such as the General Data Protection Regulation (GDPR), the ISO/IEC 27001 (2023a) security standard or the forthcoming European Union (EU) AI Act (European Parliament & Council of the European Union, 2024).

Operational defences emphasize awareness training, behavioural monitoring, cyber hygiene, and interdisciplinary collaboration across IT, legal, and compliance teams (Al-Hawamleh, 2024; Concepcion & Palaoag, 2024; Obi et al., 2024; Shareef, 2024). Threat intelligence sharing remains essential in addressing fast-evolving attack techniques, particularly AI-generated phishing and synthetic content exploitation (ENISA, 2023b).

Technical mitigation strategies include adversarial robustness, model hardening, anomaly detection, secure data pipelines, encryption, red-teaming, and AI-specific penetration testing (MITRE ATLAS, 2023). Privacy-preserving ML approaches—such as federated learning and differential privacy—reduce risks associated with centralized data storage and model inversion attacks (Dwork & Roth, 2014; Kairouz et al., 2021).

However, literature warns that defensive AI is vulnerable to adversarial adaptation, model drift, dataset biases, and false-positive escalation (Baniecki & Biecek, 2024; Costa et at., 2024; Goyal et al., 2023). Overreliance on AI-driven defence may itself create systemic risks if not combined with human oversight and strong governance structures (King et al., 2020).

2.5 Gaps in Existing Scholarship

Despite substantial scholarship on AI and cybercrime, several critical research gaps persist. First, existing studies frequently address isolated aspects of AI-driven crime but lack comprehensive frameworks integrating technological, human-centric, and organisational perspectives (Cascavilla, Tamburri & Heuvel, 2021; Malatji & Tolah, 2024; Zeng, 2022). Second, harmonised global regulation remains limited, and cross-border misuse of AI continues to outpace legal and institutional controls (Chang, 2024; Evang, 2022; OECD, 2019; Zaidan & Ibrahim, 2024). Third, although cybersecurity standards exist, few provide specific guidance on AI vulnerabilities, leaving organisations underprepared for model-specific threats (De Gregorio, 2025; ENISA, 2023b; Mcintosh et al., 2024; Rampášek et al., 2025). Fourth,

research offers limited translation of theoretical threat models into practical, actionable frameworks that organisations can operationalize (Cho & Kim, 2025; Tatam et al., 2021; Xiong & Lagerström, 2019)

This gap underscores the need for structured, comprehensive models that consolidate criminal elements of AI and map them to organisational defence requirements. Such frameworks provide critical value by integrating diverse threat categories, enabling systematic assessment, and guiding the development of targeted mitigation strategies.

3. METHODS

3.1 Research Design and Process

Our research adopted a qualitative, pragmatist research design, reflecting the need for methodological flexibility in a rapidly evolving threat landscape. Pragmatism was appropriate because AI-enabled cybercrime develops faster than theoretical models, requiring approaches that prioritise applicability and real-world utility. Employing an inductive approach, we explored emerging AI threats beyond existing theories and frameworks by deriving an artefact from literature and qualitative expert interviews (based on recommendations from Saunders et al., 2019).

Using qualitative methods, we extracted key themes literature as well as from cybersecurity and AI experts categorizing challenges and solutions (adopted on vom Brocke et al., 2020). Semi-structured interviews were conducted with eight subject matter experts selected for their roles in IT governance, risk management, and AI deployment. Interviews followed a standardized protocol focused on AI threat categorization and mitigation strategies. Thematic coding was conducted manually to identify recurring patterns and validate framework components.

The literature review (Section 2) laid the theoretical groundwork and identified research gaps, while qualitative interviews provided nuanced perspectives on threats and mitigation. Data were collected cross-sectionally to capture the current AI threat landscape, with future longitudinal studies recommended. As leading methodology, the Design Science Research (DSR) methodology as developed by Hevner and colleagues (Hevner et al., 2004; Hevner & Chatterjee, 2010). The approach from Kuechler & Vaishnavi (2008) was applied with its four iterative phases:

(1) Problem Awareness: an extensive literature review analysed AI cybercrime, threats, and (governance) frameworks, addressing RQs 1 and 2 and identifying gaps.
(2) Suggestion: Using the Double Diamond Model (Meinel et al., 2011), semi-structured interviews with Swiss experts informed AI-centric governance framework requirements. Thematic analysis and standards review guided development, addressing RQ3.
(3) Development: Requirements were translated into an initial governance framework v0.1, refined iteratively with expert feedback to the final governance framework v1.0, addressing RQ4.
(4) Evaluation: Cybersecurity experts assessed the governance framework's strengths and limitations through interviews, enabling refinements to finalize governance framework v1.0, ensuring responsiveness to evolving threats and addressing RQ5.

This structure ensured systematic alignment between empirical findings, conceptual development, and practical applicability, consistent with established guidance for artefact-oriented information systems research.

3.2 Data Collection and Analysis

A structured literature review was conducted to consolidate existing knowledge on AI enabled threats, AI-driven threat behaviours, cybercrime trends, and organisational defence measures. The review followed established guidelines for systematic identification and synthesis of sources (vom Brocke et al., 2020).

Peer-reviewed publications, standards documents, and practitioner reports were included to provide a comprehensive evidence base. The review served three purposes: (1) establishing the conceptual basis for identifying criminal AI elements, (2) deriving initial organisational requirements, and (3) informing the design of interview questions.

Empirical data were collected through semi-structured interviews with cybersecurity professionals, including CISOs, AI risk managers, incident response leads, and information security officers. Expert sampling was used to ensure participants possessed relevant operational experience, consistent with established qualitative research practice (Easterbrook et al., 2008). Two expert groups were engaged: (1) group 1 contributed to identifying AI-related criminal elements and refining organisational requirements; (2) group 2 evaluated the prototypical governance framework and provided feedback for improving v0.1. All interviews followed a predefined guide, were recorded, and were transcribed for analysis.

Interview transcripts and literature findings were analysed using thematic coding. Coding progressed from descriptive categories to more abstract analytical themes, following recognised qualitative analysis principles (Saunders et al., 2019). The analysis produced three main outputs: (1) criminal elements of AI-enabled threats, (2) organisational requirements for mitigating AI-driven threats, and (3) structured measures translating requirements into actionable activities. These results were integrated into the DSR development cycle and informed subsequent refinements of the governance framework.

3.3 Governance framework Development

The governance framework was constructed through an iterative design process that progressively refined categories and measures based on findings from literature, expert interviews, and corresponding frameworks. As leading frameworks, we used ISO/IEC 27001 (2023a), NIST CSF (2024), NIST AI RMF (2023), ENISA material (2023a, 2023b), and relevant sources identified throughout the research.

The first version v0.1 emphasized modularity, allowing users to apply measures selectively. Measures were refined and categorized by best practices, focusing on AI-augmented defences, human factors, legal compliance, and technical controls (Hu et al., 2017). Each measure details implementation steps, risk assessments, and use cases. The initial version was systematically compared to requirements to ensure completeness before evaluation.

The initial development began by defining a process and methodology grounded in DSR, enabling repeated cycles of refinement in response to practical insights. The first iteration produced an initial framework structure, centred on an overview (landing) page and detailing its purpose, intended users, and application context.

3.4 Governance Framework Evaluation

Starting with the landing page of the first draft of the governance framework v0.1, evaluation assessed the framework's effectiveness, scalability, robustness, and usability through expert interviews. Using a semi-structured interview design based on established cybersecurity standards (e.g., NIST CSF, 2024; NIST AI RMF, 2023; ENISA, 2023a, 2023b; ISO/IEC 42001, 2023b), the evaluation examined whether the governance framework adequately addresses AI-driven cyber threats and aligns with contemporary defensive practices.

Experts consistently confirmed the governance framework's logical structure, clear organisation, and practical applicability for medium to large organisations. They highlighted its strength in integrating established cybersecurity principles with AI-specific risk considerations—such as AI-powered threat detection, adversarial attacks, data privacy risks, and exposure to misinformation and deepfakes (referencing interview insights and sources like CIS Controls (Center for Internet Security, 2024), MITRE ATT&CK (2024), SANS (2012) or the WEF Global Risks Report (Bueermann & Rohrs, 2024). The governance framework was considered accessible and adaptable, offering actionable measures that bridge theory and practice. However, several improvement areas emerged. Interviewees recommended for example clarifying the framework's purpose, refining the terminology, consolidating overlapping categories, and enhancing practical guidance through examples or workflows (Table 1). These inputs informed targeted adaptations for the final version, including clearer goals, improved usage guidance, reaffirmed relevance of core cybersecurity controls, streamlined categories, and adjusted terminology.

In summary, the evaluation confirmed the framework's usefulness and relevance while addressing RQ5 by demonstrating its value for organisations facing AI-driven risks and highlighting the need for ongoing refinement toward a mature governance framework version v1.0.

Table 1. Adaptations resulting from the evaluation

#	Conceptional Governance Framework v0.1 – Adaptions after Evaluation
1	Clarified the framework's mission and defined its target organisational context
2	Reinforced foundational cybersecurity as the base for AI-specific controls
3	Improved usage guidance with practitioner-oriented guiding questions
4	Standardised terminology to remove ambiguity and ensure consistency
5	Consolidated overlapping categories and measures for greater coherence
6	Streamlined the taxonomy by reducing redundant activities

4. INITIAL GOVERNANCE FRAMEWORK v0.1

4.1 Conceptual Structure

The first version v0.1 of the governance framework was developed to provide a solid basis for the evaluation to discuss the approach, to gather feedback and finally to improve. An landing page was developed which serves as an introduction and overview to introduce the governance framework's purpose ("What is it for?"), intended audience ("Who is it for?"), usage

instructions ("How to use it?"), core cybersecurity principles for AI contexts ("Essentials"), and categories of measures ("Measures Overview" and "Measure Sheet").

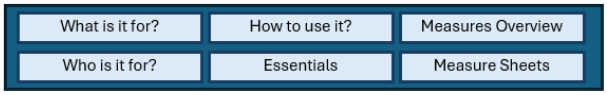

Figure 1. Governance Framework v0.1 – Landing Page

The landing page was foreseen to enable users to assess the framework's relevance and suitability for their organisational context. Target users include cybersecurity teams, IT managers, compliance officers, and risk managers in European organisations, while also serving as an awareness tool for IT professionals. Figure 1 presents the structure of the landing page.

The landing page was the starting point for the evaluation group of experts for understanding. Subsequently, the "Overview Page" was shown to discuss the entire structure (Figure 2) followed by an exemplary "Measure Sheet" (Figure 3) to identify weaknesses and add further fields if necessary.

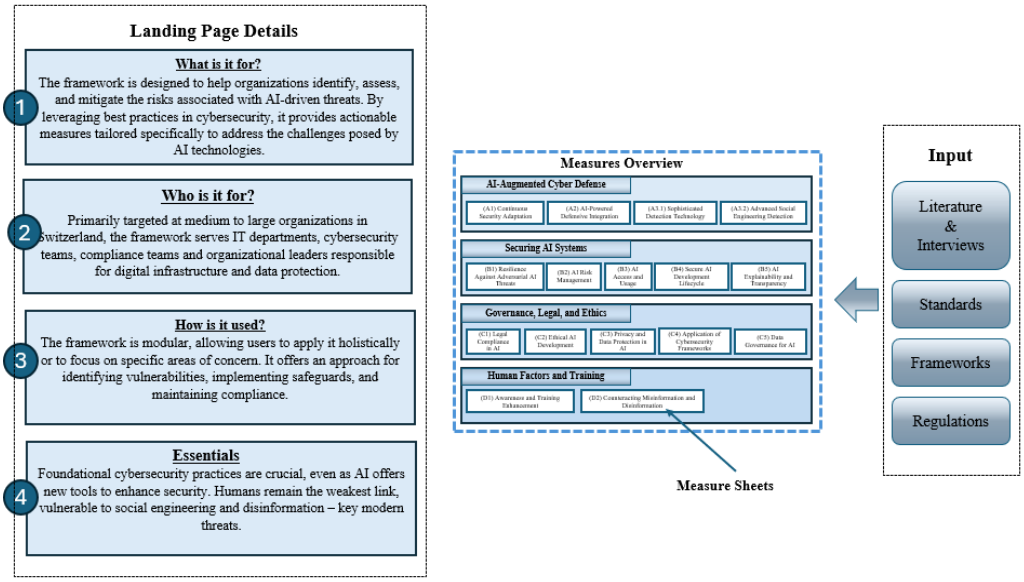

Figure 2. Governance Framework v0.1 – Overview Page

A2 - AI-Powered Defensive Integration	
Description	Integrating AI technologies into existing cybersecurity defenses to enhance capabilities in threat detection, response, and prevention.
Goal	To improve the effectiveness and efficiency of cybersecurity measures through AI.
Activities	• **Assessing and Planning AI Integration**: Evaluate current cybersecurity infrastructure to identify gaps that AI can fill and develop a roadmap for integration aligned with organizational goals. • **Researching and Acquiring Suitable AI Solutions**: Identify AI tools and platforms that align with specific security needs, ensuring compatibility and scalability. • **Implementing Technologies and Training Personnel**: Deploy AI solutions effectively and ensure that the security team is trained to use and manage them properly. • **Monitoring System Performance and Optimizing AI Integrations**: Continuously assess the performance of AI tools, adjusting parameters and configurations for optimal results. • **Updating Security Policies and Conducting Regular Audits**: Revise policies to reflect new technologies and perform audits to ensure compliance and effectiveness of AI integrations.
Example Metric	• **Increase in Threats Identified Due to AI**: Compare the number of threats detected before and after AI integration. • **Decrease in Response Time to Incidents**: Measure the reduction in time taken to respond to threats. • **Accuracy of AI Detections (False Positives/Negatives)**: Track the rate of incorrect threat identifications to assess AI reliability.
Core Sources	ISO/IEC 42001, ENISA's Cybersecurity of AI and Standardization, CIS Critical Security Controls
Additional sources	AI Maturity Model from Deloitte, Gartner Magic Quadrant for AI in Security, SANS Framework, Mlflow, COBIT, NIST CSF.

Figure 3. Governance Framework v0.1 – Exemplary Measure Sheet A2

4.2 Addressed Requirements

Table 2 shows that all initial requirements and corresponding measure resulted from the data collection and analysis (section 3.2) have been addressed in the governance framework v0.1.

Table 2 Adaptations resulting from the Governance Framework v0.1 evaluation

Mapping of requirements to measures	Addressed in Measure Sheet
(1) Incorporate adaptive security strategies	A1, A2
(2) Implement advanced detection mechanisms for social engineering attacks	A3, A4
(3) Detect and mitigate misinformation and disinformation	D2
(4) Strengthen AI models against adversarial attacks	B1
(5) Integrate AI into defensive measures	A2, A3, A4
(6) Control access to AI tools and promote responsible usage	B3, B5, D1
(7) Enhance awareness and training programs	D1, D2
(8) Develop sophisticated detection methods	A3, A4
(9) Adapt existing cybersecurity frameworks	C4
(10) Establish risk management practices for AI systems	B2, B4
(11) Ensure that AI systems comply with ethical guidelines & legal regulations	C1, C2, C3, C5

Version v0.1 was developed through an iterative refinement of categories and measures, aligned with requirements derived from expert interviews. Its structure comprises an overview page, inputs, measure categories, and detailed measure sheets. Mapping requirements to implemented measures confirmed full coverage. This process addressed RQ4 by defining the framework's initial structure and establishing the foundation for its later evaluation and refinement into version v1.0.

5. RESULTING GOVERNANCE FRAMEWORK v1.0

The initial governance framework (v0.1) was developed iteratively, with categories and measures refined to align with requirements from the literature review and expert interviews. Two development cycles were conducted: one to identify core requirements and a second to evaluate the prototype (section 3.4). The framework included a landing page outlining purpose, audience, and use, followed by input sources, measure categories, and detailed measure sheets.

All requirements from the data collection and analysis (sections 3.2 and 4.2) were mapped to specific measures. As shown in Table 2, each requirement resulted in at least one organisational measure, confirming that v0.1 met the identified needs and addressed RQ4. The evaluation also underscored strengths such as clear structure and practical relevance for medium-to-large organisations confronting AI-driven cyber threats.

Experts also identified areas for refinement, such as clarifying the framework's purpose, standardising terminology, consolidating overlapping categories, and enhancing practical guidance—for example, by adding examples or workflows to support application. These insights informed targeted improvements in version v1.0.

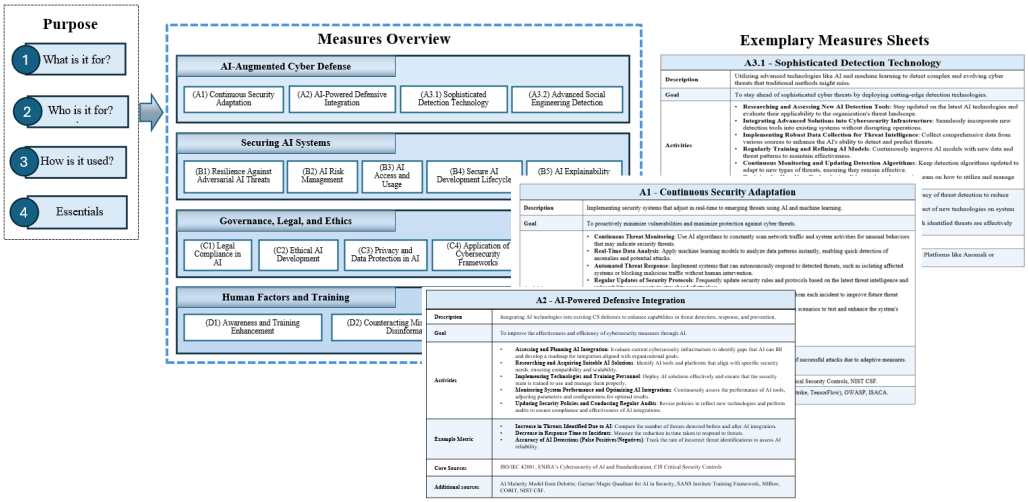

Figure 4. Resulting Governance Framework against AI-Driven Threats version v1.0

The resulting governance framework v1.0 (Figure 4) presents clarified objectives, refined terminology, streamlined categories, and strengthened implementation guidance. It integrates established cybersecurity principles with AI-specific defensive strategies to support organisations in addressing emerging threat patterns.

The framework combines traditional security practices—such as authentication, patch management, and awareness training—with measures for adaptive response, ethical oversight, and regulatory compliance.

The resulting governance framework´s modular structure enables both holistic application and targeted use, allowing it to complement existing cybersecurity programs. Developed for cybersecurity teams, IT managers, compliance officers, policymakers, and risk professionals, the framework provides clear measures across four main categories (Figure 4; detailed in Table 3).

Table 3. Governance Framework v1.0 with four main categories, measures and detailed context

#	Governance Framework Main Categories		Detailed Description
A	**AI-Augmented Cyber Defence**		**Enhancing defence via real-time detection, automated responses, and predictive analytics.**
	Measure		
	A1	Continuous Security Adaption	Implementing security systems that adjust in real-time to emerging threats using AI and machine learning.
	A2	AI-Powered Defensive Integration	Integrating AI technologies into existing cybersecurity defences to enhance capabilities in threat detection, response, and prevention.
	A3.1	Sophisticated Detection Technology	Utilizing advanced technologies like AI and machine learning to detect complex and evolving cyber threats that traditional methods might miss.
	A3.2	Advanced Social Engineering Detection	Using AI and behavioural analytics to identify and prevent social engineering attacks like phishing and spear-phishing.
B	**Securing AI Systems**		**Protecting AI from threats with dedicated security measures.**
	Measure		
	B1	Resilience Against Adversarial AI Threats	Building systems robust against attacks that exploit AI systems, such as adversarial machine learning attacks.
	B2	AI Risk Management	Identifying, assessing, and mitigating risks associated with the use of AI technologies within an organisation.
	B3	AI Access and Usage	Managing who has access to AI systems and how they are used to prevent unauthorized use or abuse.
	B4	Secure AI Development Lifecycle	Incorporating security best practices throughout the AI development lifecycle to prevent vulnerabilities.
	B5	AI Explainability and Transparency	Ensuring that AI systems are transparent, and their decision-making processes are understandable to humans.
C	**Governance, Legal, and Ethics**		**Ensuring that AI complies with laws, ethics, and strong governance.**
	Measure		
	C1	Legal Compliance in AI	Ensuring that AI technologies comply with relevant laws, regulations, and ethical guidelines.
	C2	Ethical AI Development	Ensuring that AI systems are designed and deployed according to ethical principles like fairness and respect for human rights.
	C3	Privacy and Data Protection in AI	Ensuring that AI systems handle personal and sensitive data in compliance with data protection laws like GDPR.
	C4	Application of Cybersecurity Frameworks	Updating existing cybersecurity frameworks to incorporate AI considerations and address new technological challenges.
	C5	Data Governance for AI	Establishing robust data governance frameworks to manage the quality, security, and ethical use of data in AI systems.
D	**Human Factors and Training**		**Emphasizes awareness and education to reduce AI-driven risks.**
	Measure		
	D1	Awareness and Training Enhancement	Enhancing employee awareness and training regarding cybersecurity, with a focus on AI-related threats and tools.
	D2	Counteracting Misinformation and Disinformation	Strategies and tools to detect, analyse, and mitigate the spread of false or misleading information.

Each measure includes defined objectives, implementation steps, and references to established standards, including ISO/IEC 27001 (2023a) and the NIST CSF (2024) therefore offers a coherent, practical, and governance-aligned tool to strengthen organisational defence in environments increasingly affected by AI-driven cyber threats.

Table 3 presents the structure of the governance framework v1.0, organized into four main categories that reflect the core domains of AI-related organisational defence. Each category contains specific measures with detailed descriptions, outlining the objectives, scope, and operational context required for implementation.

The table provides a concise overview of how AI-augmented defence capabilities, secure AI system practices, governance and ethical requirements, and human-centric measures collectively support a comprehensive approach to mitigating AI-driven cyber risks. Together, these components form the actionable foundation of the governance framework v1.0.

6. CONCLUSION AND OUTLOOK

This research examined the accelerating misuse of AI in cybercrime and addressed five research questions through a governance-oriented lens. It identified the key criminal elements of AI (RQ1)—such as automated phishing, deepfake-enabled deception, adversarial manipulation, and AI-generated malware—and organised them into systematic categories that clarify how AI reshapes attack dynamics (RQ2). The research then derived the essential organisational and policy requirements for mitigating these risks (RQ3), highlighting the importance of adaptive defence strategies, responsible access controls, robust governance mechanisms, and strengthened human-centric protections.

These insights informed the conceptualisation of the initial governance framework v0.1 (RQ4), which was subsequently refined through expert evaluation into version v1.0. The resulting framework integrates regulatory expectations (e.g., GDPR, the EU AI Act) with recognised cybersecurity standards and fills a critical policy gap by offering operational, AI-specific mitigation measures for cybersecurity teams, IT managers, and risk officers.

The evaluation confirmed the framework's practical value while underscoring the need for continuous refinement as AI capabilities, attacker techniques, and regulatory conditions evolve (RQ5). Nonetheless, the study is constrained by its regional expert base and the absence of full-scale organisational deployment. Future research should therefore expand the empirical foundation through multi-sector case studies, involve broader international stakeholders, and explore alignment with emerging global regulatory developments. As forthcoming regulatory instruments—such as the EU AI Act's risk-based obligations—take effect, future versions of the framework may increasingly serve as a practical bridge between compliance requirements and day-to-day defensive operations.

To support adoption, organisations can integrate the framework directly into existing cybersecurity and governance structures (e.g., ISO/IEC 27001, 2023a; NIST CSF, 2024), use it to strengthen risk assessments for AI systems, and embed its measures into training, procurement, and incident-response processes.

Ultimately, this research demonstrates that strengthening organisational defence against AI-driven cyber threats is no longer optional: timely policy action, grounded in structured frameworks such as this one, is essential for safeguarding organisational resilience in an era of rapidly evolving AI-enabled risks.

REFERENCES

Adewale, D. S. and Segun, V. S., 2024. The intersection of artificial intelligence and cybersecurity: Challenges and opportunities. *World Journal of Advanced Research and Reviews*, Vol. 21, No. 2, pp. 1720-1736. https://doi.org/10.30574/wjarr.2024.21.2.0607

Admass, W. S., Munaye, Y. Y. and Diro, A. A., 2024. Cyber security: State of the art, challenges and future directions. *Cyber Security and Applications*, Vol. 2, 100031. https://doi.org/10.1016/j.csa.2023.100031

Afghani, A., 2025. Enhancing operational effectiveness in security and defence within the UAE: The strategic role of artificial intelligence. *International Journal of Technology and Systems*, Vol. 10, No. 1, pp. 1-22. https://doi.org/10.47604/ijts.3194

AL-Hawamleh, A., 2024. Cyber resilience framework: Strengthening defenses and enhancing continuity in business security. *International Journal of Computing and Digital Systems,* Vol. 15, No. 1, pp. 1315-1331. Doi:10.12785/ijcds/15019

Baniecki, H. and Biecek, P., 2024. Adversarial attacks and defenses in explainable artificial intelligence: A survey. *Information Fusion,* Vol. 107. https://doi.org/10.1016/j.inffus.2024.102303

Blauth, T., Gstrein, O., & Zwitter, A., 2022. Artificial intelligence crime: An overview of malicious use and abuse of AI. *IEEE Access*, Vol. 10, pp. 77110-77122. https://doi.org/10.1109/access.2022.3191790

Brundage, M. et al., 2018. The malicious use of artificial intelligence: Forecasting, prevention, and mitigation. *arXiv*. https://doi.org/10.48550/ARXIV.1802.07228

Bueermann, G. and Rohrs, M., 2024. Global cybersecurity outlook 2024. *World Economic Forum (WEF)*. https://www.weforum.org/publications/global-cybersecurity-outlook-2024/

Caldwell, M., Andrews, J., Tanay, T. and Griffin, L., 2020. AI-enabled future crime. *Crime Science*, Vol. 9. https://doi.org/10.1186/s40163-020-00123-8

Cascavilla, G., Tamburri, D. and Heuvel, W. (2021). Cybercrime threat intelligence: A systematic multi-vocal literature review. *Computers & Security*, Vol. 105, 102258. https://doi.org/10.1016/j.cose.2021.102258

Center for Internet Security, 2024. *CIS Critical Security Controls (Version 8.1)*. https://www.cisecurity.org/insights/white-papers/cis-critical-security-controls-v8-1

Chang, Q., 2024. The legal and regulatory issues of AI technology in cross-border data flow in international trade. *Transactions on Economics, Business and Management Research*, Vol. 8. https://doi.org/10.62051/cyw9y102

Cho, H. and Kim, S., 2025. Threat modeling for the defense industry: Past, present, and future. *IEEE Access*, Vol. 13, pp. 53276-53304. https://doi.org/10.1109/access.2025.3550337

Concepcion, J. and Palaoag, T., 2024. An assessment of cybersecurity awareness among academic employees at Quirino State University: Promoting cyber hygiene. *Journal of Electrical Systems*, Vol. 20, No. 7s. https://doi.org/10.52783/jes.3445

Costa, J., Roxo, T., Proença, H. and Inácio, P. (2024). How deep learning sees the world: A survey on adversarial attacks & defenses. *IEEE Access*, Vol. 12, pp. 61113-61136. https://doi.org/10.1109/access.2024.3395118

De Gregorio, A., 2025. Mitigating cyber risk in the age of open-weight LLMs: Policy gaps and technical realities. *arXiv, abs/2505.17109*. https://doi.org/10.48550/arxiv.2505.17109

Dwork, C. and Roth, A., 2014. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science,* Vol. 9 No. 3-4, pp. 211-407. https://www.cis.upenn.edu/~aaroth/Papers/privacybook.pdf

Easterbrook, S., Singer, J., Storey, M. A. and Damian, D., 2008. Selecting empirical methods for software engineering research. In F. Shull, J. Singer and D. I. K. Sjøberg (eds.) *Guide to advanced empirical software engineering*. Springer, pp. 285-311. https://doi.org/10.1007/978-1-84800-044-5_11

European Parliament & Council of the European Union, 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union, L 2024/1689*, pp. 1-144.

ENISA, 2023a. *ENISA threat landscape 2023: July 2022 to June 2023*. Available at: https://data.europa.eu/doi/10.2824/782573

ENISA, 2023b. *Artificial intelligence and cybersecurity research: ENISA research and innovation brief*. Available at: https://data.europa.eu/doi/10.2824/808362

Evang, J. M., 2022. ISO 27001 as a tool for availability management. *2022 International Conference on Advanced Enterprise Information System (AEIS)*, pp. 82-85. https://doi.org/10.1109/AEIS59450.2022.00018

Fard, N., Selmic, R. and Khorasani, K., 2023. A review of techniques and policies on cybersecurity using artificial intelligence and reinforcement learning algorithms. *IEEE Technology and Society Magazine,* Vol. 42, pp. 57-68. https://doi.org/10.1109/mts.2023.3306540

Feretzakis, G., Papaspyridis, K., Gkoulalas-Divanis, A. and Verykios, V., 2024. Privacy-preserving techniques in generative AI and large language models: A narrative review. *Information*, Vol. 15, No. 11. https://doi.org/10.3390/info15110697

Golop, V. and Săvulescu, N., 2024. Profile of persons who act in the field of computer criminality. *Proceedings of the International Conference on Cybersecurity and Cybercrime (IC3)*. https://doi.org/10.19107/cybercon.2024.12

Goyal, S., Doddapaneni, S., Khapra, M. and Ravindran, B., 2023. A survey of adversarial defenses and robustness in NLP. *ACM Computing Surveys,* Vol. 55, pp. 1-39. https://doi.org/10.1145/3593042

Hadnagy, C., 2018. *Social engineering: The science of human hacking* (2nd ed.). Wiley.

He, X., Xu, G., Han, X., Zhang, Y., Li, Z. and Wang, S., 2025. Artificial intelligence security and privacy: A survey. *Science China Information Sciences,* Vol. 68, No. 181101. https://doi.org/10.1007/s11432-025-4388-5

Hevner, A. R. and Chatterjee, S., 2010. *Design research in information systems: Theory and practice.* Springer.

Hevner, A. R., March, S. T., Park, J. and Ram, S., 2004. Design science in information systems research. *MIS Quarterly*, Vol. 28, No. 1, pp. 75-105.

Hu, V. C., Kuhn, R. and Yaga, D., 2017. Verification and test methods for access control policies models. *NIST Special Publication 800-192.* National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-192

ISO, 2023a. Information security, cybersecurity and privacy protection — Information security management systems — Requirements (ISO/IEC 27001:2023). Available at: https://www.iso.org/isoiec-27001-information-security.html

ISO, 2023b. *ISO/IEC 42001:2023 Information technology — Artificial intelligence — Management system*. Available at: https://www.iso.org/standard/42001

Kairouz, P. et al., 2021. Advances and open problems in federated learning. *Foundations and Trends in Machine Learning,* Vol. 14, No. 1-2, pp. 1-210.

Kaloudi, N. and Li, J., 2020. The AI-based cyber threat landscape. *ACM Computing Surveys*, Vol. 53, pp. 1-34. https://doi.org/10.1145/3372823

Kazimierczak, M., Habib, N., Chan, J. and Thanapattheerakul, T., 2024. Impact of AI on the cyber kill chain: A systematic review. *Heliyon*, Vol. 10. https://doi.org/10.1016/j.heliyon.2024.e40699

Khan, A. et al., 2025. Artificial intelligence in computer science: Evolution, techniques, challenges, and multidisciplinary applications. *Scholars Journal of Engineering and Technology*, Vol. 13, No. 04. https://doi.org/10.36347/sjet.2025.v13i04.006

King, T. C., Aggarwal, N., Taddeo, M. and Floridi, L., 2020. Artificial intelligence crime: An interdisciplinary analysis of foreseeable threats and solutions. *Science and Engineering Ethics,* Vol. 26, No. 1, pp. 89-120. https://doi.org/10.1007/s11948-018-00081-0

Kuechler, W. and Vaishnavi, V., 2008. The emergence of design research in information systems in North America. *Journal of Design Research*, Vol. 7, No. 1, pp. 1-16. https://doi.org/10.1504/JDR.2008.021205

Liu, Y., Huang, J., Li, Y., Wang, S. and Chen, X., 2025. Generative AI model privacy: A survey. *Artificial Intelligence Review*, Vol. 58, No. 33. https://doi.org/10.1007/s10462-024-11024-6

Loh, P. K. K., Lee, A. Z. Y. and Balachandran, V., 2024. Towards a hybrid security framework for phishing awareness education and defense. *Future Internet*, Vol. 16. No. 3, 86. https://doi.org/10.3390/fi16030086

Makridakis, S., 2017. The Forthcoming Artificial Intelligence (AI) Revolution: Its Impact on Society and Firms. *Futures*, Vol. 90, pp. 46-60. https://doi.org/10.1016/j.futures.2017.03.006

Malatji, M., 2023. Offensive artificial intelligence: Current state of the art and future directions. *2023 International Conference on Digital Applications, Transformation & Economy (ICDATE)*, pp. 1-6. https://doi.org/10.1109/icdate58146.2023.10248780

Malatji, M. and Tolah, A., 2024. Artificial intelligence (AI) cybersecurity dimensions: A comprehensive framework for understanding adversarial and offensive AI. *AI and Ethics*, pp. 1-28. https://doi.org/10.1007/s43681-024-00427-4

Martineau, M., Spiridon, E. and Aiken, M., 2023. A comprehensive framework for cyber behavioral analysis based on a systematic review of cyber profiling literature. *Forensic Sciences*, Vol. 3, No. 3. https://doi.org/10.3390/forensicsci3030032

Maurya, R., 2023. Analyzing the role of AI in cyber security threat detection & prevention. *International Journal for Research in Applied Science and Engineering Technology*, Vol. 11, No. 11, pp. 514-519. https://doi.org/10.22214/ijraset.2023.56510

McIntosh, T., Sušnjak, T., Liu, T., Watters, P., Nowrozy, R. and Halgamuge, M., 2024. From COBIT to ISO 42001: Evaluating cybersecurity frameworks for opportunities, risks, and regulatory compliance in commercializing large language models. *arXiv, abs/2402.15770*. https://doi.org/10.1016/j.cose.2024.103964

Meinel, C., Leifer, L. and Plattner, H. (eds.), 2011. *Design thinking*. Springer. https://doi.org/10.1007/978-3-642-13757-0

Melaku, H. M., 2023. A dynamic and adaptive cybersecurity governance framework. *Journal of Cybersecurity and Privacy*, Vol. 3, No. 3, pp. 327-350. https://doi.org/10.3390/jcp3030017

MITRE, 2023. *ATLAS: Adversarial threat landscape for artificial-intelligence systems*. Available at: https://atlas.mitre.org/

MITRE, 2024. *MITRE ATT&CK*. Available at: https://attack.mitre.org/

Mohamed, N., 2023. Current trends in AI and ML for cybersecurity: A state of the art survey. *Cogent Engineering*, Vol. 10, No. 2, 2272358. https://doi.org/10.1080/23311916.2023.2272358

Murdoch, B., 2021. Privacy and artificial intelligence: Challenges for protecting health information in a new era. *BMC Medical Ethics*, Vol. 22. https://doi.org/10.1186/s12910-021-00687-3

NIST, 2023. Artificial intelligence risk management framework (AI RMF 1.0). *NIST Special Publication 1270*.

NIST, 2024. *Cybersecurity framework (CSF) 2.0*. Available at: https://doi.org/10.6028/NIST.CSWP.29

Obi, O., Akagha, O., Dawodu, S., Anyanwu, A., Onwusinkwue, S. and Ahmad, I., 2024. Comprehensive review on cybersecurity: Modern threats and advanced defense strategies. *Computer Science & IT Research Journal*, Vol. 5, No. 2. https://doi.org/10.51594/csitrj.v5i2.758

Obioha-Val, O., Olaniyi, O., Gbadebo, M., Balogun, A. and Olisa, A., 2025. Cyber Espionage in the Age of Artificial Intelligence: A Comparative Study of State-Sponsored Campaign. *Asian Journal of Research in Computer Science*, Vol. 18, No. 1. https://doi.org/10.9734/ajrcos/2025/v18i1557

OECD (2019). *OECD principles on artificial intelligence*. Available at: https://doi.org/10.1787/eedfee77-en

Pawlicka, A., Choraś, M. and Pawlicki, M., 2021. The stray sheep of cyberspace a.k.a. the actors who claim they break the law for the greater good. *Personal and Ubiquitous Computing*, Vol. 25, pp. 843-852. https://doi.org/10.1007/s00779-021-01568-7

Radanliev, P., 2025. Privacy, ethics, transparency, and accountability in AI systems for wearable devices. *Frontiers in Digital Health*, Vol. 7. https://doi.org/10.3389/fdgth.2025.1431246

Rampášek, M., Mesarčík, M. and Andraško, J., 2025. Evolving cybersecurity of AI-featured digital products and services: Rise of standardisation and certification? *Computer Law & Security Review,* Vol. 56, 106093. https://doi.org/10.1016/j.clsr.2024.106093

Rao, G. R. K., Battu, V. V., Anupama, V., Allada, A., Krishna, S. V. R. and Hema, C., 2023. Modern progressive pitfalls of cyber attacks on the digital world. *2023 2nd International Conference on Edge Computing and Applications (ICECAA)*, pp. 244-248. https://doi.org/10.1109/ICECAA58104.2023.10212303

Rawat, R., Oki, O., Chakrawarti, R., Adekunle, T., Lukose, J. and Ajagbe, S., 2023. Autonomous Artificial Intelligence Systems for Fraud Detection and Forensics in Dark Web Environments. *Informatica (Slovenia)*, Vol. 47. https://doi.org/10.31449/inf.v46i9.4538

SANS, 2012. *SANS incident handler's handbook*. Available at: https://www.sans.org/white-papers/33901

Saunders, M. N. K., Lewis, P. and Thornhill, A., 2019. *Research methods for business students* (8th ed.). Pearson.

Shareef, O., 2024. Building organizational defense: A comprehensive approach to implementing IT controls for SOX compliance. *International Journal of Computer Science and Mobile Computing,* Vol. 13., No. 2, pp. 69-71. https://doi.org/10.47760/ijcsmc.2024.v13i02.006

Tatam, M., Shanmugam, B., Azam, S. and Kannoorpatti, K., 2021. A review of threat modelling approaches for APT-style attacks. *Heliyon*, Vol. 7. https://doi.org/10.1016/j.heliyon.2021.e05969

vom Brocke, J., Hevner, A. and Maedche, A., 2020. Introduction to design science research. In J. vom Brocke, A. Hevner and A. Maedche (eds.) *Design science research. Cases. Progress in IS*. Springer, pp. 1-16. https://doi.org/10.1007/978-3-030-46781-4_1

Xiong, W. and Lagerström, R.,2019. Threat modeling: A systematic literature review. *Computers & Security,* Vol. 84, pp. 53-69. https://doi.org/10.1016/j.cose.2019.03.010

Xu, Y. et al., 2021. Artificial intelligence: A powerful paradigm for scientific research. *The Innovation,* Vol. 2. https://doi.org/10.1016/j.xinn.2021.100179

Yeung, K., 2024. International governance of advancing artificial intelligence. *AI & Society*, Vol. 40, pp. 3019-3044. https://doi.org/10.1007/s00146-024-02050-7

Zaidan, E. and Ibrahim, I., 2024. AI governance in a complex and rapidly changing regulatory landscape: A global perspective. *Humanities and Social Sciences Communications*, Vol. 11, 1121. https://doi.org/10.1057/s41599-024-03560-x

Zeng, Y., 2022. AI empowers security threats and strategies for cyber attacks. *Procedia Computer Science,* Vol. 208, pp. 170-175. https://doi.org/10.1016/j.procs.2022.10.025