iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

AI DRIVEN CYBER THREATS AND DATA PRIVACY: A RESILIENT FRAMEWORK FOR DETECTING AND RESPONDING TO EMERGING CYBERSECURITY RISKS

Rashid Hussain
University of East London
Nour Internet Company for communication & IT, Riyadh, Saudi Arabia
corporate.lawyer.ksa@gmail.com
00966-566751233

ABSTRACT
Artificial intelligence (AI) is simultaneously reshaping the offensive and defensive dimensions of cybersecurity while introducing new categories of risk that arise from AI systems themselves. Generative and adversarial techniques have lowered the cost of convincing phishing, deepfake enabled deception, and automated exploitation, while AI based detection systems offer defenders faster and more adaptive analytics. At the same time, the data intensive nature of AI driven security tooling creates privacy exposures including model inversion, membership inference, training data leakage, and prompt injection that existing cybersecurity, privacy, and AI governance frameworks address only partially and largely in isolation. This paper undertakes a structured review of the scholarly and institutional literature on AI enabled cyberattacks, AI enabled cyber defence, privacy preserving machine learning, explainable AI, and adaptive incident response, in order to identify the fragmentation between these bodies of work and the resulting gap in integrated guidance. Building on that synthesis, the paper proposes the Resilient AI Driven Cybersecurity and Data Privacy Framework (RA CyDPF), a six-layer conceptual architecture spanning threat intelligence and data acquisition, privacy and data governance, AI based detection and analytics, explainability and human oversight, adaptive incident response, and resilience through continuous learning. The framework is developed through thematic synthesis and comparative analysis of existing frameworks rather than empirical experimentation, and every proposed metric and evaluation approach is presented as conceptual and illustrative rather than as measured results. The paper's contribution lies in the explicit integration of security detection, privacy protection, explainability, and organisational resilience within a single architecture, together with a defensible strategy for its future empirical evaluation. RA CyDPF is intended to complement, not replace, established standards such as NIST CSF 2.0, NIST AI RMF, ISO/IEC 27001, and ISO/IEC 42001. Implications for security operations centres, privacy officers, AI developers, and policymakers are discussed, alongside the limitations of a literature derived conceptual framework and the empirical validation required before operational adoption.

Keywords:
Artificial intelligence; cybersecurity; data privacy; explainable AI; adversarial machine learning; privacy preserving machine learning; cyber resilience; incident response; AI governance; conceptual framework

1. INTRODUCTION

Cybersecurity has always evolved in step with the technologies it seeks to protect, but the diffusion of artificial intelligence across both attacker and defender toolchains has compressed that evolutionary cycle into months rather than years. Large language models can now draft convincing phishing content, clone voices from short audio samples, and assist in vulnerability discovery at a pace that outstrips many organisations' capacity to adapt their defences (Ahi and Valizadeh, 2026; Begou et al., 2023). At the same time, the same generative and analytical capabilities are being embedded into security operations centres to accelerate anomaly detection, threat triage, and automated response (Kaur et al., 2023; Sarker, 2023). The result is a threat landscape in which AI is not a peripheral tool but a structural feature of both attack and defence.

This dual role is complicated by a third dimension that is frequently underexamined in isolation: the risks that AI systems create for the confidentiality and integrity of data merely by existing within an organisation's security stack. Machine learning models trained on network telemetry, authentication logs, or behavioural data can themselves

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

become privacy liabilities, since they may be susceptible to model inversion, membership inference, training data extraction, and, in the case of large language model based security assistants, prompt injection (Fredrikson et al., 2015; Shokri et al., 2017; Vassilev et al., 2025). A security architecture that improves detection by ingesting ever more granular personal and behavioural data may therefore increase, rather than decrease, an organisation's overall privacy risk, a tension that is rarely made explicit in either cybersecurity or the AI governance literature.

Existing guidance tends to treat these three dimensions, AI enabled attack, AI enabled defence, and AI created risk, as separate problems addressed by separate communities. Cybersecurity frameworks such as the NIST Cybersecurity Framework 2.0 (National Institute of Standards and Technology [NIST], 2024) provide a structured account of governance, protection, detection, and response, but were not written with AI specific risks such as adversarial manipulation or model level privacy leakage as a central concern. Conversely, AI risk management guidance such as the NIST AI Risk Management Framework (Tabassi, 2023) and the associated adversarial machine learning taxonomy (Vassilev et al., 2025) addresses AI trustworthiness and adversarial robustness in considerable technical depth but does not offer an operational cybersecurity architecture. Privacy preserving machine learning research spanning differential privacy, federated learning, and secure aggregation has matured substantially (Dwork and Roth, 2014; McMahan et al., 2017; Mothukuri et al., 2021) but is rarely connected explicitly to incident response workflows or to the explainability requirements that security analysts need in order to trust and act on AI generated alerts (Arrieta et al., 2020; Mohale and Obagbuwa, 2025).

This paper addresses that fragmentation. It first synthesises the literature across AI enabled threats, AI enabled defence, privacy preserving machine learning, explainable AI, and existing governance frameworks, in order to substantiate the claim that an integrated operational architecture is currently missing rather than merely under documented. It then proposes the Resilient AI Driven Cybersecurity and Data Privacy Framework (RA CyDPF), a six-layer conceptual architecture intended to connect threat intelligence, privacy governance, AI based detection, human oversight, adaptive response, and continuous organisational learning into a single coherent structure. The paper is explicit throughout the boundary between what the literature establishes empirically and what the proposed framework offers as a conceptual synthesis: no new experiments were conducted, and all evaluation metrics presented in later sections are illustrative targets for future empirical work rather than measured outcomes. RA CyDPF is positioned as a complementary integration layer that operationalises considerations across established standards, not as a replacement for NIST CSF 2.0, NIST AI RMF, ISO/IEC 27001, ISO/IEC 42001, or applicable privacy frameworks.

The intended scope of RA CyDPF is organisations that deploy AI based cybersecurity tooling, particularly security operations centres (SOCs), enterprise security teams, and regulated entities in sectors such as finance, healthcare, and critical infrastructure, where the intersection of AI driven threat detection, data privacy obligations, and human oversight creates operational complexity that existing frameworks do not resolve in an integrated manner. The framework is not presented as universally applicable to all organisations regardless of maturity; rather, it offers a conceptual reference architecture that organisations can adapt to their existing cybersecurity and privacy governance maturity.

The paper is organised as follows. Section 2 sets out the conceptual foundations distinguishing AI enabled attack, AI enabled defence, and AI created risk. Section 3 synthesises the relevant literature critically rather than descriptively. Section 4 explains the review methodology in detail, including its limitations. Section 5 analyses the AI driven threat landscape in detail. Section 6 examines AI for cyber defence. Section 7 addresses the privacy and security risks that AI systems themselves introduce. Section 8 compares existing frameworks, substantiates the research gap, and articulates the distinctive contribution of RA CyDPF. Section 9 presents the proposed framework architecture in full, including operational definitions for each layer. Section 10 sets out a conceptual evaluation and validation strategy. Sections 11 through 13 discuss implications, limitations, and future research, and Section 14 concludes.

1.1 Research Questions

The following five research questions guide the synthesis and framework development:

- RQ1: How are AI technologies transforming the nature and sophistication of emerging cyber threats?
- RQ2: What cybersecurity and data privacy vulnerabilities arise from AI enabled threat detection and response systems themselves?

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

- RQ3: What limitations exist in current cybersecurity, AI risk management, and privacy frameworks when applied to AI driven threats?
- RQ4: How can AI based detection, privacy preserving techniques, human oversight, and adaptive incident response be integrated into a single resilient cybersecurity architecture?
- RQ5: How might the effectiveness and resilience of such an integrated framework be evaluated in future empirical work?

2. BACKGROUND AND CONCEPTUAL FOUNDATIONS

A clear analytical account of AI and cybersecurity requires separating three dimensions that are frequently conflated in both trade and academic writing: AI used by attackers, AI used by defenders, and risks created by AI systems themselves. Conflating these dimensions produces imprecise recommendations, because a control that mitigates one dimension, for example, restricting model access to reduce extraction risk, may have little bearing on another, such as an organisation's exposure to AI generated phishing.

2.1 AI Used by Attackers

Generative models have measurably changed the economics of social engineering. Large language models can draft grammatically fluent, contextually personalised phishing messages at a scale that previously required substantial human effort, and several studies report that AI assisted phishing content achieves higher click through and conversion rates than traditional templates (Begou et al., 2023; Schmitt and Flechais, 2024). Voice and video based deepfakes extend this deception into synchronous communication channels, enabling impersonation of executives and colleagues in ways that bypass many trained heuristics for detecting fraud (Mubarak et al., 2023; Mustak et al., 2023; Wang et al., 2024). Beyond social engineering, adversaries increasingly use automated tools for reconnaissance, vulnerability discovery, and the generation of polymorphic malware capable of evading signature-based detection (Coppolino et al., 2025; Kumar and Sinha, 2023). Adversarial machine learning techniques, evasion, poisoning, and extraction attacks against defenders' own models, represent a further offensive capability directed specifically at the AI systems that organisations deploy for protection (Goodfellow et al., 2015; Vassilev et al., 2025).

2.2 AI Used by Defenders

On the defensive side, machine learning and deep learning techniques underpin contemporary intrusion detection, malware classification, and behavioural analytics, offering the capacity to identify anomalies within volumes of network and endpoint telemetry that exceed human analytical capacity (Kaur et al., 2023; Kilincer et al., 2021; Zhang, Ning et al., 2022). Threat intelligence platforms increasingly apply natural language processing to unstructured sources to accelerate the correlation of indicators of compromise (Alevizos and Dekker, 2024; Tounsi and Rais, 2018), while automated triage and orchestration reduce the burden on security analysts facing high volumes of alerts (Saeed et al., 2023). Reinforcement learning has also been explored for adaptive defence, in which a system incrementally improves its response policy in response to observed attacker behaviour (Sewak et al., 2023).

2.3 Risks Created by AI Systems Themselves

A distinct category of risk arises not from how AI is used by either side of the conflict, but from the vulnerabilities inherent to AI systems as software artefacts. Predictive models are susceptible to evasion attacks that induce misclassification, and to poisoning attacks that corrupt training data to implant backdoors or degrade accuracy (Goodfellow et al., 2015; Vassilev et al., 2025). Privacy attacks, membership inference, model inversion, training data extraction, and property inference, can reveal whether specific records were used in training or can reconstruct sensitive attributes of the training population (Carlini et al., 2021; Fredrikson et al., 2015; Shokri et al., 2017). Generative and agentic systems introduce further risks specific to their architecture, including direct and indirect prompt injection, in which adversarial instructions embedded in user input or retrieved content manipulate model behaviour (Vassilev et al., 2025). Model extraction attacks allow adversaries to reconstruct a functional copy of a proprietary model through repeated querying (Tramer et al., 2016), while excessive permissions granted to autonomous AI agents create a further operational risk that is only beginning to receive sustained attention in the literature. Retrieval augmented generation (RAG) systems introduce additional vulnerabilities, including poisoned knowledge sources and sensitive information disclosure through retrieved context. Because these vulnerabilities are properties of the AI system rather than of the attacker or the defender specifically, they require governance and technical controls distinct from conventional network security measures.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

3. LITERATURE REVIEW

The literature on AI and cybersecurity has grown rapidly since 2020, with bibliometric analyses confirming a marked acceleration in publication volume and a shift in dominant themes toward privacy, explainability, and generative AI in the years 2023 and 2024 in particular (Ilic et al., 2024). This section synthesises that literature critically across five interconnected themes, rather than presenting a study by study summary. Contemporary AI governance instruments beyond NIST and ISO are also considered where they illuminate the regulatory context within which the proposed framework operates.

3.1 AI Enabled Threat Detection and the Persistence of the Black Box Problem

A substantial body of survey literature documents the effectiveness of machine learning and deep learning techniques for intrusion detection, malware classification, and anomaly detection (Dasgupta et al., 2022; Kaur et al., 2023; Zhang, Ning et al., 2022). Comparative studies consistently report that deep learning architectures outperform classical signature based methods on benchmark datasets, particularly for zero day and previously unseen attack variants (Kilincer et al., 2021). However, this literature converges on a recurring limitation: detection accuracy is frequently reported in isolation from interpretability, operational integration, and false positive burden on analysts, and comparatively few studies examine how detection models behave under adversarial conditions specifically engineered to evade them (Sarker, 2023). The explainability literature has grown partly in response to this gap. Reviews of explainable AI in intrusion detection converge on the finding that post hoc techniques such as SHAP and LIME improve analyst trust and support regulatory auditability, but that a persistent trade off remains between model interpretability and raw detection accuracy, and that standardisation of evaluation for explanation quality remains immature (Mahbooba et al., 2021; Mohale and Obagbuwa, 2025; Rudin, 2019).

3.2 Generative AI and the Escalation of Social Engineering

A second cluster of literature documents the effect of generative AI on social engineering. Schmitt and Flechais (2024) propose a conceptual model in which generative AI amplifies social engineering along three pillars, realistic content creation, advanced targeting, and automated attack infrastructure, and this framing is corroborated by controlled studies showing that AI generated phishing content can outperform human crafted equivalents in open and click through rates (Begou et al., 2023). Deepfake research similarly documents a widening gap between the sophistication of synthetic media and the reliability of detection tools, with survey evidence indicating that multimodal deepfake detection remains an unsolved problem at scale (Mubarak et al., 2023; Wang et al., 2024). This literature is largely descriptive of the threat and comparatively thin on integrated organisational countermeasures that combine technical detection with governance and human verification protocols, a gap the present paper addresses directly in the proposed framework's human oversight layer.

3.3 Privacy Preserving Machine Learning in Security Contexts

Privacy preserving machine learning, principally differential privacy, federated learning, and secure multi party computation, has developed into a mature technical field (Bonawitz et al., 2017; Dwork and Roth, 2014; McMahan et al., 2017). Systematic reviews confirm that federated learning is increasingly applied to intrusion detection and malware classification specifically because it avoids centralising sensitive network and endpoint data (Mothukuri et al., 2021; Yin et al., 2021). At the same time, this literature documents that privacy preserving techniques introduce a measurable accuracy privacy trade off, and that federated architectures remain vulnerable to gradient leakage and membership inference attacks unless combined with differential privacy or secure aggregation (Demelius et al., 2025; Zhang, Zhu et al., 2022; Zhu et al., 2019). Critically, this technical literature is rarely connected to incident response or governance frameworks; privacy preserving ML is typically evaluated on accuracy and privacy loss metrics in isolation from how a security operations centre would operationalise, explain, or act upon its outputs.

3.4 Explainability and Human Oversight

The explainable AI literature converges on the argument that opacity in AI driven security tools erodes analyst trust, complicates regulatory compliance, and can itself become an attack surface if explanation mechanisms are exploited to reverse engineer detection logic (Arrieta et al., 2020; Rudin, 2019). Recent work has begun to treat explainability not merely as a usability feature but as a security relevant control that supports auditability and accountability under emerging AI governance requirements (Mohale and Obagbuwa, 2025). However, the literature offers limited guidance on how explainability mechanisms should be embedded structurally within a detection to response pipeline, as opposed to being applied post hoc to individual model outputs.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

3.5 Governance and Institutional Frameworks

Institutional guidance has evolved substantially in recent years. The NIST Cybersecurity Framework 2.0 introduced a dedicated Govern function addressing organisational risk strategy, supply chain risk, and cross functional accountability (NIST, 2024), while the NIST AI Risk Management Framework provides a parallel structure, Govern, Map, Measure, Manage, specifically for AI risk (Tabassi, 2023). The associated adversarial machine learning taxonomy offers the most comprehensive available vocabulary for AI specific attack types (Vassilev et al., 2025). The NIST Privacy Framework (NIST, 2020) provides a risk based approach to privacy engineering that aligns with the CSF functions but is often treated separately from cybersecurity operations. Internationally, ISO/IEC 27001:2022 continues to anchor information security management, and ISO/IEC 42001:2023 extends management system principles specifically to AI governance. The European Union's AI Act (European Parliament and Council of the European Union, 2024) introduces risk based classification of AI systems with corresponding conformity obligations, while the GDPR (European Parliament and Council of the European Union, 2016) remains the primary data protection reference for organisations processing personal data. These frameworks are individually well developed, but each was authored by a different community with a different central concern, cybersecurity risk, AI trustworthiness, information security certification, or data protection regulation, and none of them was designed as an operational architecture that connects AI based detection, privacy governance, explainability, and adaptive incident response into a single implementable structure. This fragmentation, substantiated across Sections 3.1 through 3.5, forms the basis for the research gap articulated in Section 8.

4. METHODOLOGY

This paper adopts a structured literature review and thematic synthesis methodology combined with comparative framework analysis, consistent with accepted approaches for conceptual and framework development research in information systems and cybersecurity scholarship (Vom Brocke et al., 2009; Wolfswinkel et al., 2013). No primary experimental data were collected. The review was not conducted as a formally registered systematic review with a documented database level record count; rather, it represents a structured, transparent synthesis designed to maximise reproducibility within the constraints of a conceptual framework development study. The methodological process is reported below in detail, including its acknowledged limitations.

4.1 Search Strategy and Sources

Searches were conducted between January 2024 and March 2026, with the final literature update performed on 15 March 2026. The following academic databases and search platforms were queried: IEEE Xplore, ACM Digital Library, Scopus, Web of Science, and Google Scholar. Institutional repositories consulted included the NIST Publications Portal, the ISO Online Browsing Platform, and the arXiv preprint server (limited to papers later published in peer reviewed venues or widely cited institutional reports).

Search terms combined core concept clusters using Boolean logic, following the general pattern: ("artificial intelligence" OR "machine learning" OR "deep learning") AND ("cybersecurity" OR "cyberattack" OR "intrusion detection") AND ("privacy" OR "data protection" OR "explainability" OR "adversarial"). Searches were supplemented with targeted queries for institutional publications (for example, "NIST AI Risk Management Framework", "NIST Cybersecurity Framework 2.0", "NIST adversarial machine learning taxonomy", "ISO/IEC 42001:2023", "EU AI Act") and for specific technical subtopics (federated learning intrusion detection, membership inference, deepfake detection, explainable intrusion detection systems, retrieval augmented generation security, agentic AI risks). Sources were drawn primarily from peer reviewed journals (including Information Fusion, Computers and Security, IEEE Communications Surveys and Tutorials, IEEE Access, Sensors, Electronics, ACM Computing Surveys, and Nature Machine Intelligence), IEEE and ACM conference proceedings, and official publications of NIST, ISO/IEC, and the European Union.

4.2 Inclusion and Exclusion Criteria

Sources were included if they: (a) addressed AI or machine learning applied to cybersecurity threat detection, defence, or adversarial risk; (b) addressed privacy preserving machine learning, differential privacy, or federated learning in a security relevant context; (c) addressed explainable AI in intrusion detection or security decision making; (d) constituted an authoritative institutional framework (NIST, ISO/IEC, EU regulatory instruments) relevant to cybersecurity, AI risk, or privacy governance; or (e) provided empirical evidence of AI driven threats (for example,

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

controlled phishing studies, deepfake detection benchmarks, adversarial attack demonstrations). Sources were prioritised from the period 2020 to 2026, with seminal earlier technical papers (for example, foundational work on differential privacy, federated learning, and adversarial examples) retained where they remain the primary reference for a technique still in active use. Sources were excluded if they were non peer reviewed opinion content without a substantive evidentiary or technical basis, or if they duplicated the specific findings of an already included source without adding distinct content.

4.3 Screening and Selection Process

Titles and abstracts were screened independently against the inclusion criteria. Full texts were retrieved for all sources that passed initial screening and were assessed for relevance, technical depth, and evidentiary quality. Duplicate publications identified across databases were resolved by retaining the most complete or most recently updated version. Where multiple versions of the same work existed (for example, arXiv preprints subsequently published in peer reviewed venues), the peer reviewed version was preferred. Exact historical counts of records identified, screened, excluded, and finally included were not recorded in a PRISMA compliant format; this is acknowledged as a methodological limitation. The absence of a formal record count means that the review cannot claim exhaustive coverage of every potentially relevant source, and readers should treat the synthesis as representative rather than census complete.

4.4 Quality and Relevance Assessment

Source quality was assessed hierarchically. Peer reviewed journal articles and major IEEE/ACM/USENIX conference papers were treated as the strongest evidence base. Official NIST, ISO/IEC, and EU regulatory sources were treated as authoritative for framework level claims. Recognised institutional threat reports (for example, CrowdStrike, World Economic Forum) were used where empirical industry evidence was necessary but were not relied upon for core technical or novelty claims where stronger peer reviewed sources existed. Preprints were used only where they represented the most current available source for an emerging threat and were clearly identified as such. Lower quality publications, non peer reviewed blog posts, vendor white papers without independent validation, or opinion pieces, were excluded.

4.5 Thematic Synthesis and Framework Development Process

Included sources were coded thematically against the five review clusters presented in Section 3 (detection and explainability, generative AI enabled threats, privacy preserving machine learning, human oversight, and governance frameworks). Thematic coding was performed iteratively: an initial coding scheme was derived from the research questions and refined as new cross cutting themes emerged during full text review. Recurring points of convergence and divergence across clusters were extracted, with particular attention to instances where one literature identified a limitation that a different, non overlapping literature had already addressed technically, for example, the detection and explainability literature's concern with analyst trust corresponds closely to unresolved usability questions in the explainable AI literature, while the accuracy privacy trade off documented in the privacy preserving machine learning literature is rarely discussed by authors writing about incident response workflows. These cross cluster gaps directly informed the six layer structure of the proposed framework in Section 9, with each layer corresponding to a thematic cluster and the connections between layers designed specifically to close the fragmentation documented in Section 3. Disagreements or subjective judgments in assigning studies to themes were resolved through iterative discussion and reconciliation against the inclusion criteria. Where a source spanned multiple themes, it was coded in each relevant cluster to preserve cross references. The framework development process followed the comparative logic described by Vom Brocke et al. (2009): existing institutional frameworks were mapped against their stated scope, AI coverage, privacy coverage, and threat response capability to substantiate, rather than assert, the identified gap. The six layers of RA CyDPF were then derived systematically from the capabilities required to close that gap, as documented in the Framework Derivation table (Table 8).

4.6 Limitations of the Review Method

The review methodology is subject to several limitations that should inform interpretation of the findings. First, the absence of a formally registered systematic review with database level record counts means that the synthesis may be subject to publication and search bias: sources indexed in the searched databases are overrepresented relative to grey literature, non English language scholarship, or national regulatory guidance not translated into English. Second, the rapid evolution of AI threat literature means that sources published before the final search date may already be

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

superseded by newer findings, particularly in the generative AI domain. Third, the thematic synthesis necessarily involves interpretive judgment; alternative coders might emphasise different cross cluster connections. Fourth, the quality assessment prioritised peer reviewed sources, which may introduce a lag bias: the most current threat intelligence sometimes appears first in institutional reports rather than in peer reviewed journals. These limitations are acknowledged transparently and do not invalidate the conceptual contribution of RA CyDPF, but they do mean that the framework should be treated as a literature derived starting point for future empirical refinement rather than as a definitive, exhaustively validated architecture.

5. THE AI DRIVEN CYBER THREAT LANDSCAPE

The AI driven threat landscape is best understood as a set of pathways connecting attacker capability, through specific attack vectors, to organisational targets and their downstream cybersecurity and privacy consequences. Figure 1 depicts this pathway conceptually, distinguishing four AI enabled attacker capabilities, generative content synthesis, automated reconnaissance and open source intelligence gathering, adversarial model manipulation, and behavioural mimicry, from the specific attack vectors they enable.

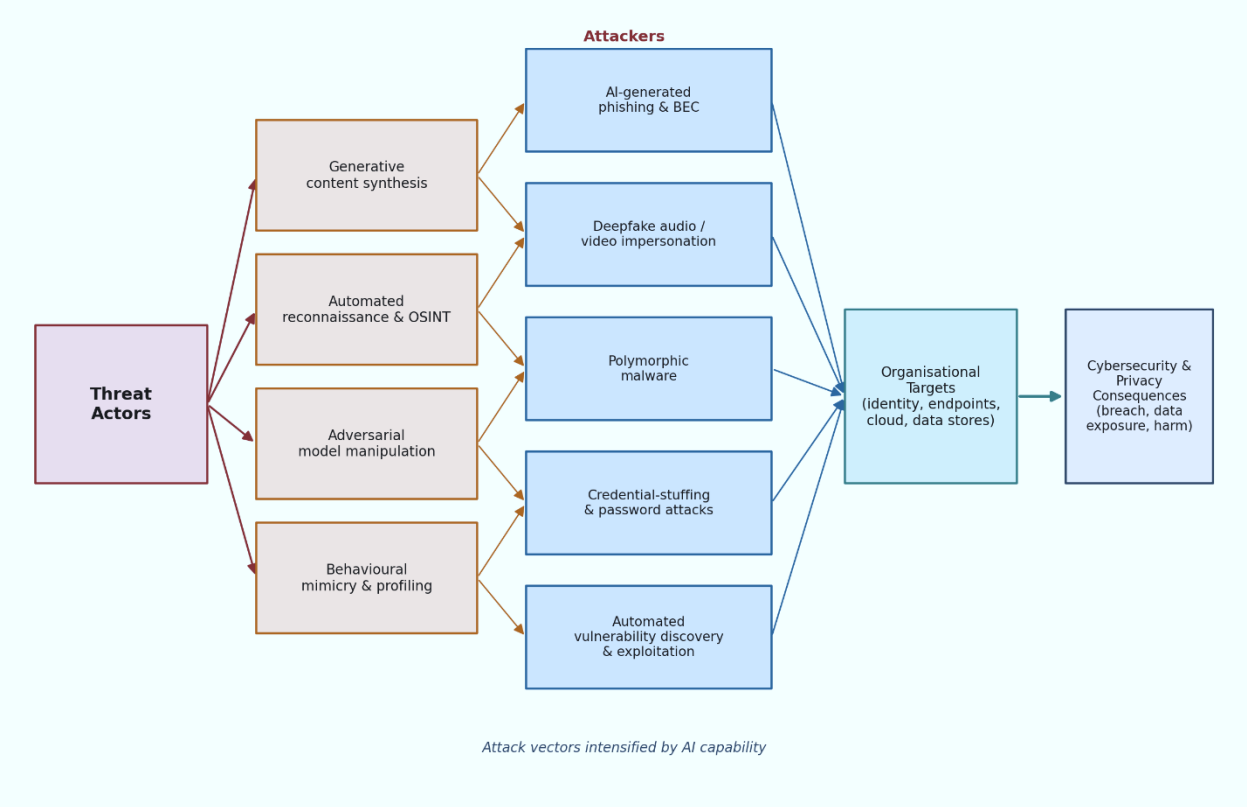

Figure 1. AI-driven cyber threat landscape and organisational impact pathways.

Generative content synthesis underpins AI enhanced phishing and business email compromise, in which large language models produce grammatically fluent, contextually personalised messages that mimic an individual's writing style and reference legitimate organisational detail gathered through automated reconnaissance (Begou et al., 2023; Schmitt and Flechais, 2024). The same synthesis capability extends to deepfake audio and video impersonation, which has already been implicated in high value fraud incidents involving fabricated executive video conferences (Mustak et al., 2023; Wang et al., 2024). Automated reconnaissance and open source intelligence tools compress what previously required hours of manual profiling into an automated pipeline, enabling attackers to construct detailed

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

personalised vulnerability profiles of individual targets at scale. Adversarial model manipulation, by contrast, is directed specifically at the AI systems organisations deploy defensively, encompassing evasion, poisoning, and extraction techniques documented in the NIST adversarial machine learning taxonomy (Vassilev et al., 2025). Behavioural mimicry and profiling techniques support both credential based attacks and the construction of synthetic identities that evade conventional identity verification controls.

These capabilities converge on organisational targets, principally identity systems, endpoints, cloud infrastructure, and data stores, with consequences spanning direct financial loss, operational disruption, regulatory exposure, and the erosion of trust in digital communication channels. A taxonomy of these threats, organised by category, underlying AI capability, mechanism, potential impact, and the specific challenge each poses for detection, is presented in Table 1.

Table 1. Taxonomy of AI Driven Cyber Threats

Threat category	AI capability	Attack mechanism	Potential impact	Detection challenge
AI generated phishing / BEC	Generative language modelling	Personalised, contextually fluent messages referencing OSINT derived detail	Credential theft, fraudulent transfers, initial access	Content is linguistically indistinguishable from legitimate communication
Deepfake impersonation	Generative audio/video synthesis	Real time or pre recorded synthetic media of trusted individuals	Authorisation fraud, reputational and financial harm	High quality deepfakes evade both automated and human detection
Polymorphic / AI assisted malware	Generative and adversarial ML	Code or behaviour mutated to evade signature detection	System compromise, persistence, data exfiltration	Signatures degrade quickly; requires behavioural detection
Automated vulnerability discovery and exploitation	ML assisted fuzzing and code analysis	Automated scanning and exploit generation at scale	Faster time to exploit, reduced attacker skill barrier	Speed of exploitation may outpace patch cycles
Adversarial evasion of detection models	Adversarial machine learning	Crafted inputs causing misclassification of malicious activity	Undetected intrusion, false sense of security	Requires adversarial robustness testing rarely performed operationally
Credential and identity attacks	Behavioural mimicry / synthetic identity	Automated credential stuffing, synthetic identity construction	Account takeover, fraud, bypass of verification	Synthetic identities can pass conventional verification checks
Indirect prompt injection	Generative AI / RAG systems	Adversarial instructions embedded in retrieved external content	Data exfiltration, unauthorised actions, privilege escalation	Input sanitisation cannot fully trust external knowledge sources
Model supply chain poisoning	Adversarial ML	Compromised training data, pre trained weights, or third party plugins	Backdoored detection, degraded accuracy, persistent compromise	Supply chain verification of AI artefacts is immature

Note. BEC = business email compromise; OSINT = open source intelligence; RAG = retrieval augmented generation.

6. AI FOR CYBER DEFENCE AND THREAT DETECTION

Where Section 5 examined AI as an offensive capability, this section examines the corresponding defensive toolkit. Machine learning and deep learning techniques are now embedded across the detection pipeline, from network level

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

anomaly detection to endpoint behavioural analytics and automated threat triage. Supervised classifiers trained on labelled traffic remain widely used for known attack detection, while unsupervised and semi supervised techniques including autoencoders and clustering based anomaly detection are increasingly applied to identify previously unseen attack patterns without requiring exhaustive labelled datasets (Kaur et al., 2023; Kilincer et al., 2021). Behavioural analytics extends this capability to user and entity behaviour, establishing baselines of normal activity against which deviations are flagged for investigation.

Threat intelligence platforms increasingly apply natural language processing to unstructured sources, security advisories, forums, and incident reports, to accelerate the extraction and correlation of indicators of compromise, reducing the manual burden historically associated with threat intelligence curation (Alevizos and Dekker, 2024; Tounsi and Rais, 2018). Automated incident triage systems apply classification and prioritisation models to the high volume of alerts generated by modern security stacks, a capability of particular importance given well documented analyst alert fatigue in security operations centres (Saeed et al., 2023). Predictive and adaptive defence techniques, including reinforcement learning based response policies, represent a further frontier in which the defensive system incrementally adjusts its posture based on observed outcomes (Sewak et al., 2023).

Despite these advances, the literature converges on several recurring limitations. First, detection accuracy reported on benchmark datasets frequently does not generalise to production environments with different traffic characteristics and evolving attack patterns (Sarker, 2022). Second, the interpretability of complex models, particularly deep neural networks, remains a persistent barrier to analyst trust and regulatory auditability, motivating the explainability requirements incorporated into the proposed framework (Section 9, Layer 4). Third, few detection systems are evaluated for robustness against adversarial manipulation specifically designed to evade them, despite the existence of a mature taxonomy of such attacks (Vassilev et al., 2025). Table 2 summarises the principal AI techniques applied in cyber defence contexts alongside their characteristic strengths, limitations, and privacy implications.

Table 2. AI Techniques for Cyber Threat Detection

AI technique	**Security application**	**Strength**	**Limitation**	**Privacy implication**
Supervised classification (e.g., random forest, gradient boosting)	Known attack and malware classification	High accuracy on labelled, well characterised threats	Poor generalisation to novel or evolving attacks	Requires labelled data that may include sensitive attributes
Deep learning (CNN, LSTM, autoencoders)	Network intrusion detection, behavioural anomaly detection	Captures complex, non linear patterns; effective on high dimensional data	Limited interpretability; computationally intensive	Risk of memorising and later leaking sensitive training patterns
Unsupervised / semi supervised anomaly detection	Zero day and previously unseen attack detection	Does not require exhaustive labelled datasets	Higher false positive rates than supervised approaches	Behavioural baselining can constitute continuous personal monitoring
Natural language processing for threat intelligence	Correlating indicators of compromise from unstructured sources	Accelerates analyst workflow and reduces manual curation	Quality depends heavily on source reliability	May process personal or organisational data present in open sources
Reinforcement learning for adaptive response	Automated containment and response policy adaptation	Can improve response strategy iteratively based on outcomes	Immature for high stakes production deployment; limited explainability	Automated actions may affect individuals without direct human review

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

7. DATA PRIVACY AND AI SECURITY RISKS

AI enabled cybersecurity tooling is inherently data intensive: effective anomaly detection depends on granular network telemetry, authentication logs, and behavioural signals that are frequently personal or quasi identifying in nature. This data intensity creates a structural tension, illustrated conceptually in Figure 4, between the volume and granularity of data required for effective AI based detection and the data minimisation principles that underpin contemporary privacy regulation and privacy by design practice.

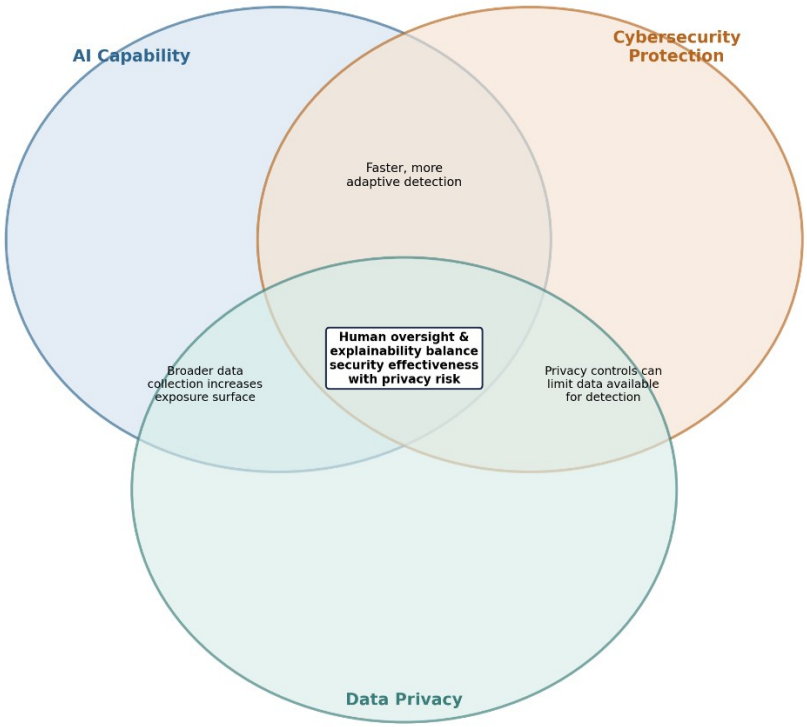

Figure 2. Interactions and trade-offs among AI capability, cybersecurity protection, and data privacy.

AI capability and cybersecurity protection reinforce one another where richer data enables faster and more adaptive detection. However, the intersection of AI capability and data privacy is characterised by tension rather than synergy: broader data collection required to train and operate effective detection models directly increases the surface available for privacy attack. Model inversion techniques can reconstruct approximations of sensitive training inputs from model outputs (Fredrikson et al., 2015); membership inference attacks can determine whether a specific individual's data was used in training, which is itself a disclosure in contexts such as health or financial monitoring (Shokri et al., 2017); and training data extraction attacks against generative models have been shown to recover verbatim fragments of training data, including personally identifying content, from large language models (Carlini et al., 2021). Model extraction attacks pose a related but distinct risk, in which an adversary reconstructs a functional copy of a proprietary detection model through repeated querying, potentially exposing both intellectual property and any sensitive patterns the model encodes (Tramer et al., 2016).

The intersection of cybersecurity protection and data privacy exhibits the converse tension: privacy preserving controls such as data minimisation, aggregation, or differential privacy, while necessary to limit exposure, can reduce the

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

granularity of data available for detection and thereby degrade a model's sensitivity to subtle attack indicators (Demelius et al., 2025). Federated learning offers a partial resolution by keeping raw data local to its source and sharing only model updates, but remains vulnerable to gradient leakage attacks unless combined with differential privacy or secure aggregation, each of which reintroduces some accuracy cost (Bonawitz et al., 2017; Mothukuri et al., 2021; Zhu et al., 2019). Generative and agentic AI systems introduce a further, architecturally distinct risk in the form of direct and indirect prompt injection, in which adversarial instructions embedded in user input or in retrieved external content manipulate model behaviour, potentially causing the exfiltration of sensitive context or the execution of unauthorised actions (Vassilev et al., 2025). Excessive permissions granted to autonomous AI agents, insecure tool or plugin use, and poisoned training or knowledge sources in retrieval augmented generation systems represent additional emerging risks that the framework must accommodate. As illustrated in Figure 4, human oversight and explainability mechanisms, formalised as Layer 4 of the proposed framework, are positioned at the intersection of all three domains precisely because they offer a mechanism for balancing detection effectiveness against privacy risk without requiring either to be sacrificed unconditionally.

Table 3 summarises the principal privacy risks associated with AI enabled cybersecurity, their underlying cause, potential consequence, available mitigation, and the residual concern that remains even after mitigation is applied.

Table 3. Data Privacy Risks in AI Enabled Cybersecurity

Privacy risk	**Cause**	**Potential consequence**	**Mitigation**	**Residual concern**
Membership inference	Model overfitting to training data	Disclosure that an individual's data was used in training	Differential privacy, regularisation, reduced overfitting	Utility trade off; residual risk persists at non zero privacy budgets
Model inversion / training data extraction	Model memorisation of specific training records	Reconstruction of sensitive attributes or verbatim data	Differential privacy, output perturbation, access restriction	Detection remains difficult; extraction may go unnoticed
Model extraction	Repeated querying of a deployed model	Theft of proprietary detection logic and embedded patterns	Rate limiting, query monitoring, watermarking	Determined adversaries can distribute queries to evade limits
Gradient leakage in federated learning	Shared model updates encode information about local data	Partial reconstruction of local training data from gradients	Secure aggregation, gradient clipping, differential privacy	Adds communication and computational overhead
Prompt injection (direct and indirect)	Untrusted input or retrieved content processed by generative models	Manipulated model behaviour, sensitive data exfiltration	Input sanitisation, privilege separation, output monitoring	No fully reliable technical defence currently established
Over collection for detection purposes	Broad telemetry collection to maximise detection sensitivity	Expanded surface for breach or secondary misuse	Data minimisation, purpose limitation, retention limits	May reduce detection sensitivity for subtle or novel threats
Agentic action risks	Excessive permissions granted to autonomous AI agents	Unauthorised containment actions, privilege escalation, cascade failures	Principle of least privilege, human in the loop for high impact actions	Autonomy speed trade off may delay response in time critical incidents

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

8. EXISTING FRAMEWORKS AND RESEARCH GAP

Section 3.5 introduced the principal institutional frameworks relevant to this study. This section compares them systematically to substantiate, rather than assert, the fragmentation that motivates the proposed framework in Section 9. Table 4 compares seven widely referenced frameworks against their primary purpose, AI coverage, privacy coverage, threat response capability, and their major limitation when considered as a candidate for governing AI driven cybersecurity and data privacy in an integrated manner.

Table 4. Comparison of Existing Cybersecurity, AI Risk, and Privacy Frameworks

Framework	Primary purpose	AI coverage	Privacy coverage	Threat response capability	Major limitation
NIST Cybersecurity Framework 2.0 (NIST, 2024)	Organisation wide cybersecurity risk management	Limited; addresses AI as a supply chain and governance topic, not a technical detection architecture	Indirect, via general risk management alignment	Strong (Identify, Protect, Detect, Respond, Recover)	Not designed for AI specific technical risks such as adversarial manipulation
NIST AI Risk Management Framework (Tabassi, 2023)	Managing risk across the AI system lifecycle	Comprehensive (Govern, Map, Measure, Manage)	Addressed as one of several trustworthiness characteristics	Limited; not an operational incident response architecture	Does not specify a cybersecurity detection to response pipeline
NIST Adversarial ML Taxonomy (Vassilev et al., 2025)	Standardising terminology for adversarial ML attacks and mitigations	Comprehensive for predictive and generative AI	Addressed through privacy attack taxonomy	Not applicable; a taxonomy rather than an operational framework	Provides vocabulary, not an implementable organisational architecture
NIST Privacy Framework (NIST, 2020)	Privacy risk management through enterprise risk management	Not addressed as a technical AI security concern	Comprehensive (Identify P, Govern P, Control P, Communicate P, Protect P)	Limited; privacy risk management, not cybersecurity incident response	Not designed as an operational cybersecurity architecture with AI detection
ISO/IEC 27001:2022	Information security management system certification	Minimal; general controls applicable but not AI specific	Addressed through general information security controls	Strong at the governance and control catalogue level	Certification oriented; not a technical AI security architecture
ISO/IEC 42001:2023	AI management system certification	Comprehensive AI governance and lifecycle management	Indirect; privacy addressed as one of several AI system impacts	Limited; management system focus rather than operational threat response	Not an operational cybersecurity detection to response architecture
Zero Trust Architecture, NIST SP 800	Identity and resource centric access	Not addressed; predates widespread AI	Indirect, via access minimisation	Strong for access control; not a detection	Does not address AI based detection, explainability, or

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

207 (Rose et al., 2020)	control architecture	specific threat modelling		or analytics framework	privacy preserving analytics

This comparison illustrates a consistent pattern: each framework is authoritative and well developed within its own scope, but none was designed to connect AI based threat detection, privacy preserving data handling, explainability, and adaptive incident response within a single operational architecture. The NIST Cybersecurity Framework 2.0 provides the strongest account of organisational cybersecurity risk management but treats AI largely as a governance and supply chain concern rather than as a technical detection capability requiring its own controls. The NIST AI Risk Management Framework and the associated adversarial machine learning taxonomy provide the most detailed treatment of AI specific risk but are not structured as an operational cybersecurity architecture with defined detection to response workflows. The NIST Privacy Framework offers a structured approach to privacy risk management but is not integrated with cybersecurity operations. ISO/IEC 27001:2022 anchors certifiable information security governance but was not written with AI specific technical risk in mind. ISO/IEC 42001:2023 extends management system principles to AI governance but does not prescribe operational cybersecurity controls. Zero Trust Architecture offers a strong access control philosophy but does not address AI based analytics or privacy preserving detection design.

This pattern is corroborated by the applied research literature. Table 5 summarises a representative set of recent studies selected to illustrate the thematic fragmentation documented in Section 3. These studies were chosen because they are highly cited, published in reputable peer reviewed venues, and each represents a distinct thematic cluster (detection, explainability, privacy preserving ML, adversarial taxonomy, and generative AI threats). They are not presented as an exhaustive census of the literature; rather, they serve as exemplars of the broader pattern in which individual research efforts typically advance one thematic cluster without integrating the others, and few studies explicitly connect their technical contribution to an organisational incident response or governance workflow.

Table 5. Representative Studies Illustrating Thematic Fragmentation in the Literature

Study	**Year**	**Research focus**	**AI method**	**Privacy consideration**	**Key finding / research gap**
Kaur, Gabrijelcic and Klobucar	2023	AI for cybersecurity literature review	Multiple ML/DL techniques	Discussed conceptually	Synthesises AI applications across the security lifecycle; does not propose an integrated technical architecture
Mohale and Obagbuwa	2025	Explainable AI in intrusion detection	XAI techniques (SHAP, LIME, rule based)	Not central focus	Rule and tree based XAI preferred for interpretability; limited connection to privacy or incident response workflow
Mothukuri et al.	2021	Security and privacy of federated learning	Federated learning	Central focus	FL reduces raw data exposure but remains vulnerable to gradient leakage; not connected to cybersecurity detection use cases specifically
Vassilev et al. (NIST)	2025	Adversarial ML taxonomy	Taxonomy across PredAI and GenAI	Addressed via privacy attack category	Provides the most comprehensive AI attack vocabulary available; taxonomy only, no operational implementation guidance
Schmitt and Flechais	2024	Generative AI in social engineering	Generative language and media models	Not central focus	Proposes a three pillar model of AI amplified social engineering; countermeasures

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

					discussed conceptually, not architecturally
Demelius et al.	2025	Differential privacy in centralized deep learning	Differential privacy, deep learning	Central focus	Documents accuracy privacy trade offs systematically; not connected to SOC operational workflows or explainability requirements
Yao et al.	2024	LLM security and privacy survey	Large language models	Addressed as one of several risk dimensions	Comprehensive taxonomy of LLM risks including prompt injection and data extraction; lacks operational integration with SOC incident response

Note. FL = federated learning; GenAI = generative AI; LLM = large language model; PredAI = predictive AI; SOC = security operations centre; XAI = explainable artificial intelligence.

Taken together, the institutional comparison in Table 4 and the applied research comparison in Table 5 substantiate the central claim of this paper: existing literature and frameworks treat AI enabled detection, privacy protection, explainability, and adaptive response as related but largely separate concerns. The reviewed literature did not identify a sufficiently integrated operational architecture simultaneously connecting AI driven threat detection, privacy preserving data governance, explainability and human oversight, adaptive incident response, and continuous resilience. What is missing is not further technical depth within any single stream, each is individually well developed, but an integrating architecture that makes the interactions between these streams explicit and operational. This is the gap the Resilient AI Driven Cybersecurity and Data Privacy Framework, presented in Section 9, is intended to address.

8.1 Novelty and Distinctive Contribution of RA CyDPF

The novelty of RA CyDPF lies not in the individual techniques it incorporates, each draws on established literature, but in the explicit operational integration of four domains that existing frameworks address separately. The following distinctions clarify what RA CyDPF provides that existing standards do not, and why the six layer structure constitutes a useful conceptual contribution rather than merely repackaging existing frameworks

8.1.1 Distinction from NIST CSF 2.0

NIST CSF 2.0 provides a mature, organisation wide cybersecurity risk management structure organised around the functions Govern, Identify, Protect, Detect, Respond, and Recover. RA CyDPF does not duplicate these functions; rather, it complements them by adding: (a) explicit AI specific technical controls (for example, adversarial robustness testing, model extraction defences) within the detection and response pipeline; (b) a dedicated privacy governance layer positioned upstream of AI analytics, operationalising data minimisation and purpose limitation principles as structural preconditions rather than compliance checklists; (c) an explainability and human oversight layer that defines risk based escalation criteria for automated versus human reviewed decisions; and (d) a continuous learning feedback loop that closes the cycle from incident outcome back to threat intelligence and model retraining. Where NIST CSF 2.0 treats AI as a governance and supply chain concern, RA CyDPF treats AI as a structural feature of the detection to response pipeline requiring its own operational controls.

8.1.2 Distinction from NIST AI RMF

The NIST AI Risk Management Framework (AI RMF) provides a comprehensive lifecycle approach to AI trustworthiness through the functions Govern, Map, Measure, and Manage. RA CyDPF complements the AI RMF by translating its trustworthiness principles into an operational cybersecurity architecture. Specifically, RA CyDPF adds: (a) a detection to response workflow that the AI RMF does not prescribe; (b) explicit integration of privacy preserving machine learning techniques (differential privacy, federated learning, secure aggregation) as structural components of the data pipeline; (c) defined human oversight escalation logic that connects explanation quality to response authorisation; and (d) adversarial robustness testing as a continuous organisational practice rather than a one time measurement activity. The AI RMF asks organisations to measure and manage AI risk; RA CyDPF proposes one way to operationalise that mandate within a SOC context.

8.1.3 Distinction from ISO/IEC 27001 and ISO/IEC 42001

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

ISO/IEC 27001:2022 anchors information security management system certification through a control based approach, while ISO/IEC 42001:2023 extends management system principles to AI governance. Both are certification oriented and designed to demonstrate conformity rather than to prescribe operational technical architectures. RA CyDPF complements these standards by: (a) mapping specific AI security controls (for example, adversarial testing, gradient leakage defences) to operational layers rather than to abstract control objectives; (b) defining inputs, processes, outputs, and responsible roles for each layer, making the framework actionable for SOC designers rather than audit ready for certifiers; and (c) explicitly modelling the accuracy privacy explainability trade off as a governance decision rather than assuming that all three objectives can be maximised simultaneously. RA CyDPF is not a substitute for ISO/IEC 27001 or 42001 certification; it is a conceptual architecture that organisations can use to operationalise the intent of those standards within their detection and response infrastructure.

8.1.4 Distinction from Zero Trust Architecture

Zero Trust Architecture (ZTA) offers a powerful identity and resource centric access control philosophy based on the principle of never trusting and always verifying. RA CyDPF does not replace ZTA; rather, it addresses domains that ZTA was not designed to cover. Specifically, RA CyDPF adds: (a) AI based behavioural analytics and anomaly detection as a detection capability beyond access control enforcement; (b) privacy preserving analytics that enable detection without centralising sensitive data; (c) explainability mechanisms that support analyst trust in AI generated alerts; and (d) adaptive incident response that goes beyond access revocation to include containment, remediation, and organisational learning. ZTA governs who and what can access resources; RA CyDPF governs how AI driven detection, privacy protection, and human oversight interact to detect and respond to threats that may bypass or exploit access controls.

8.1.5 Why Six Interacting Layers Constitute a Conceptual Contribution

The six layer structure is not an arbitrary decomposition. It was derived systematically from the cross cluster gaps identified in the literature review (see Table 8, Framework Derivation), with each layer closing a specific fragmentation documented in Sections 3.1 to 3.5. The layering is conceptual rather than physical: a given organisation might implement multiple layers within a single platform or distribute them across separate systems. The contribution lies in making the interactions between layers explicit, particularly the feedback from Layer 6 (resilience) back to Layer 1 (threat intelligence), and the cross cutting governance requirements that act on every layer simultaneously. By defining inputs, processes, outputs, responsible roles, decision points, and feedback mechanisms for each layer, RA CyDPF moves beyond a high level conceptual diagram toward an operationally meaningful architecture that practitioners can map against existing infrastructure, identify gaps, and prioritise investments. The framework is not claimed to be the first to mention any of its constituent concepts; it is claimed to be the first literature derived conceptual architecture that explicitly integrates these four domains, AI driven detection, privacy governance, explainability, and adaptive resilience, into a single cyclical operational structure with defined governance responsibilities.

9. PROPOSED RESILIENT AI DRIVEN CYBERSECURITY AND DATA PRIVACY FRAMEWORK

This section presents the paper's central contribution: the Resilient AI Driven Cybersecurity and Data Privacy Framework (RA CyDPF), a six layer conceptual architecture synthesised from the literature reviewed in Sections 3 and 8 and structured to make explicit the interactions between AI based detection, privacy governance, explainability, and adaptive response that existing frameworks leave implicit or unaddressed. The framework is a conceptual synthesis rather than an empirically validated system; Section 10 sets out how its components could be evaluated in future work.

The framework is organised as six interacting layers rather than as isolated technical modules. Data and decisions flow sequentially from threat intelligence acquisition through to resilience and continuous learning, but the architecture is explicitly cyclical: outputs of the resilience layer feed back into the acquisition layer, and cross cutting governance and accountability concerns act on every layer simultaneously rather than being confined to a single stage. This cyclical structure directly addresses the fragmentation identified in Section 8, in which existing frameworks typically address either the detection pipeline or the governance layer, but not their ongoing interaction.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

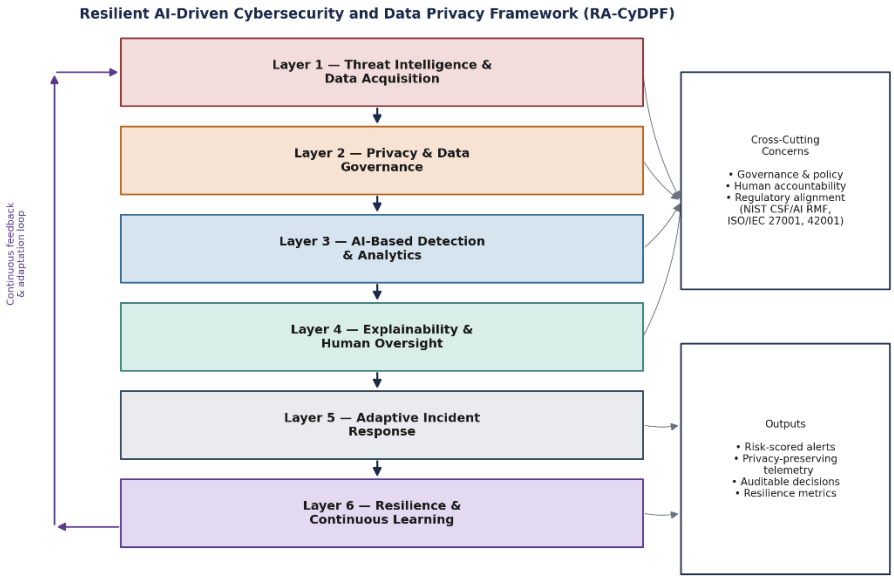

Figure 3. Architecture of the proposed resilient AI-driven cybersecurity and data privacy framework.

9.1 Layer 1, Threat Intelligence and Data Acquisition

Core function: Aggregate and classify raw security and privacy relevant data from internal and external sources.
Inputs: External threat feeds, internal security logs, network telemetry, endpoint data, identity and access information, vulnerability intelligence, and regulatory data source catalogues.
Main processes/controls: Automated feed correlation, data tagging and classification at point of acquisition, source reliability scoring, initial privacy sensitivity labelling.
Outputs: Classified data streams ready for privacy governance; threat intelligence packages; data source provenance records.
Responsible organisational role: Threat intelligence team; data steward.
Decision/escalation point: Governance sign off required for new data sources before ingestion; high sensitivity sources escalated to privacy officer.
Link to next layer: Classified data feeds forwarded to Layer 2 for privacy governance transformation.
Feedback mechanism: Layer 6 resilience outcomes inform which sources proved most valuable and which should be deprioritised.
Relevant external framework/control: NIST CSF 2.0 Identify function; ISO/IEC 27001 A.12.6 (information gathering).

9.2 Layer 2, Privacy and Data Governance

Core function: Apply minimisation, access control, and privacy preserving transformation before data reaches AI analytics.
Inputs: Classified data streams from Layer 1; privacy impact assessment triggers; regulatory requirements (GDPR, sector specific rules).
Main processes/controls: Data minimisation, purpose limitation, access control, classification, anonymisation or pseudonymisation, encryption, retention limits, privacy impact assessment prior to operationalisation, federated aggregation pipelines, differential privacy noise injection where appropriate.
Outputs: Privacy transformed datasets; privacy impact assessment records; data provenance and lineage documentation.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

Responsible organisational role: Privacy officer; data protection officer; data governance committee.
Decision/escalation point: Privacy impact assessment must be approved before any new data source or detection model is operationalised; acceptable privacy accuracy trade offs defined by governance committee.
Link to next layer: Privacy transformed data forwarded to Layer 3 for AI based detection.
Feedback mechanism: Layer 6 incident reviews may reveal privacy control failures, triggering reassessment of Layer 2 controls.
Relevant external framework/control: NIST Privacy Framework Control P; GDPR Articles 5, 25, 35; ISO/IEC 27001 A.8.2 (information classification).

9.3 Layer 3, AI Based Detection and Analytics

Core function: Detect and score anomalies and threats using AI/ML techniques.
Inputs: Privacy transformed data from Layer 2; threat intelligence context from Layer 1; analyst defined detection thresholds.
Main processes/controls: ML/DL classification, behavioural analytics, anomaly detection, threat classification, risk scoring, adversarial robustness monitoring; operates on privacy transformed data where feasible, accepting accuracy trade offs as deliberate governance decisions.
Outputs: Alert scores; threat classifications; confidence metrics; feature attribution vectors for explainability.
Responsible organisational role: Security operations centre analysts; AI/ML engineers.
Decision/escalation point: Detection thresholds set by analysts; model deployment approved by AI governance board; models failing adversarial robustness checks blocked from production.
Link to next layer: Alert packages with confidence scores and explanations forwarded to Layer 4 for validation.
Feedback mechanism: Layer 6 performance monitoring informs model retraining and threshold recalibration.
Relevant external framework/control: NIST CSF 2.0 Detect function; NIST AI RMF Measure and Manage functions; ISO/IEC 42001 A.7 (AI system operation).

9.4 Layer 4, Explainability and Human Oversight

Core function: Validate and interpret AI outputs before action, applying risk based human oversight principles.
Inputs: Alert packages from Layer 3; organisational risk appetite definitions; business impact assessments.
Main processes/controls: Feature attribution generation (SHAP, LIME, rule based justification), confidence scoring, risk based escalation logic. The framework introduces a risk based human oversight principle: automated response is authorised only when an alert satisfies a multi factor threshold combining confidence, severity, reversibility, potential business impact, privacy impact, and criticality of the affected system. High confidence alone does not automatically justify automated action. For example, an AI system might have high confidence but still require human approval before disabling a critical production system.
Outputs: Validated alerts with explanation records; automated or human authorised action directives; audit logs of all decisions.
Responsible organisational role: Security analysts (validation); incident commander (escalation); AI ethics board (oversight policy).
Decision/escalation point: Analyst validation required for alerts not meeting the multi factor automated response threshold; incident commander authorises high impact actions; AI ethics board reviews acceptable privacy accuracy trade offs.
Link to next layer: Validated alerts forwarded to Layer 5 for adaptive incident response.
Feedback mechanism: Analyst feedback on explanation quality informs Layer 3 model selection and Layer 6 training data review.
Relevant external framework/control: NIST AI RMF Govern function; EU AI Act Article 14 (human oversight); ISO/IEC 42001 A.8 (AI system impact evaluation).

9.5 Layer 5, Adaptive Incident Response

Core function: Contain, remediate, and recover from validated threats.
Inputs: Validated alerts from Layer 4; containment playbooks; recovery targets.
Main processes/controls: Automated containment for alerts meeting the multi factor threshold; threat prioritisation and isolation for ambiguous cases; remediation and recovery workflows; structured threat hunting informed by Layer 1 intelligence; all response actions logged and access controlled.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

Outputs: Containment confirmations; remediation reports; recovery time objective attainment records; threat hunting findings.
Responsible organisational role: Incident response team; SOC manager; system owners.
Decision/escalation point: Human authorisation required for high impact containment actions (for example, disconnecting critical infrastructure); SOC manager approves playbook deviations.
Link to next layer: Incident outcomes and response metrics forwarded to Layer 6 for organisational learning.
Feedback mechanism: Response effectiveness data feeds Layer 6 resilience assessment and Layer 1 threat intelligence refinement.
Relevant external framework/control: NIST CSF 2.0 Respond and Recover functions; ISO/IEC 27001 A.16 (information security incident management).

9.6 Layer 6, Resilience and Continuous Learning

Core function: Learn from outcomes and adapt the system.
Inputs: Incident outcomes from Layer 5; model performance metrics; governance review findings; adversarial test results.
Main processes/controls: Structured feedback loops, ongoing model performance monitoring, adversarial robustness testing informed by the taxonomy in Vassilev et al. (2025), periodic model retraining, organisational resilience assessment, privacy review of retraining data, review of adversarial testing results, acceptance of residual AI and privacy risks by governance.
Outputs: Updated threat intelligence (fed back to Layer 1); retrained models; revised detection thresholds; updated privacy controls; resilience metrics dashboard.
Responsible organisational role: Governance review board; CISO; privacy officer; AI model risk committee.
Decision/escalation point: Governance review board approves model retraining; CISO accepts residual security risks; privacy officer accepts residual privacy risks; AI model risk committee approves models for continued production.
Link to next layer: Outputs feed back into Layer 1, completing the cycle.
Feedback mechanism: This layer is the feedback mechanism for the entire framework; its outputs directly inform all preceding layers.
Relevant external framework/control: NIST CSF 2.0 Govern function; NIST AI RMF Manage function; ISO/IEC 27001 A.18 (compliance) and A.12.1 (operational procedures).

9.7 Governance Responsibilities (RACI Style Conceptual Table)

Table 6 presents a conceptual mapping of governance responsibilities across the six layers. It is not a formal RACI matrix but rather an illustrative assignment of accountability to clarify who conceptually owns each decision.

Table 6. Conceptual Governance Responsibilities Across RA CyDPF Layers

Governance activity	Layer 1	Layer 2	Layer 3	Layer 4	Layer 5	Layer 6
Approving new data sources	A	C	I	I	I	R
Conducting privacy impact assessments	C	A	I	I	I	R
Approving AI models for production	I	C	R	C	I	A
Defining acceptable privacy/accuracy trade offs	C	A	C	C	I	R
Authorising automated containment	I	I	C	A	R	C
Reviewing high impact AI decisions	I	C	C	A	R	C
Approving model retraining	I	C	R	I	I	A
Reviewing adversarial testing results	C	I	R	C	I	A
Accepting residual AI/privacy risks	I	A	C	C	C	R

Note. A = accountable (final decision maker); R = responsible (performs the work); C = consulted (provides input); I = informed (receives updates). This is a conceptual assignment for illustrative purposes.

9.8 Detection and Response Workflow

Figure 3 depicts the sequential workflow through which the six layers interact during a single detection to response cycle, showing explicitly where privacy controls and explainability mechanisms operate within the pipeline rather than being applied only at the architectural level.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

Data collection and pre processing correspond to Layer 1; privacy filtering, nonymizatio, masking, nonymization, and access control enforcement corresponds to Layer 2 and is applied before the data reaches the AI analysis stage. AI analysis, threat classification, and risk scoring correspond to Layer 3. The human or automated decision stage corresponds to Layer 4, at which explainability mechanisms, confidence scores, feature attribution, and analyst validation determine whether a case proceeds to automated containment or is escalated for human review. Containment and recovery correspond to Layer 5. Feedback and model updating close the loop into Layer 6, and the resulting insights re enter the workflow at the data collection stage, forming the continuous cycle.

Figure 4. Conceptual AI-driven cyber-threat severity and response-priority matrix.

9.9 Threat Prioritisation Within the Framework

Because organisations cannot treat every AI driven threat with equal urgency, the framework incorporates a conceptual risk prioritisation matrix (Figure 5) to support triage decisions within Layer 4. The positions of illustrative threat categories on this matrix are conceptual placements informed by the relative characteristics described in the threat landscape literature synthesised in Section 5, rather than measurements derived from an empirical incident dataset, and are presented explicitly as such.

Threats combining high engineered sophistication with high potential impact, such as model poisoning directed at a training pipeline, or deepfake enabled executive fraud, are placed toward the critical region of the matrix, warranting the most stringent combination of automated containment and mandatory human review under Layer 4. Lower impact,

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

higher volume threats, such as automated credential stuffing attempts, are positioned to support a higher degree of automated handling, consistent with the adaptive response logic of Layer 5. This placement logic is intended as a starting heuristic for organisations adopting the framework, to be recalibrated against an organisation's own incident history once available.

The conceptual risk prioritisation matrix can be calibrated using variables such as: threat likelihood; business impact; AI sophistication; detectability; privacy impact; affected asset criticality; confidence level; and response urgency. These variables are presented as a conceptual scoring approach rather than as a mathematical formula, recognising that quantitative calibration would require empirical validation beyond the scope of this paper.

9.10 Framework Derivation

Table 8 demonstrates that the six layers were systematically derived from the literature rather than selected subjectively. Each row connects a literature finding to an identified gap, a required capability, and the proposed RA CyDPF layer that addresses it.

Table 8. Framework Derivation from Literature Review

Literature finding	Identified gap	Required capability	Proposed RA CyDPF layer	Supporting sources
AI driven threats are increasingly sophisticated but threat intelligence is fragmented	No unified intake for AI specific threat data	Structured threat intelligence and data acquisition with privacy classification at ingestion	Layer 1: Threat Intelligence and Data Acquisition	Begou et al., 2023; Vassilev et al., 2025
Privacy preserving ML exists but is rarely connected to SOC workflows	Privacy controls applied as afterthoughts rather than structural preconditions	Privacy governance upstream of analytics, with PIAs and data minimisation	Layer 2: Privacy and Data Governance	Dwork and Roth, 2014; Mothukuri et al., 2021
AI detection accuracy is high but adversarial robustness is rarely tested operationally	Detection models deployed without ongoing adversarial validation	AI based detection with embedded robustness monitoring	Layer 3: AI Based Detection and Analytics	Goodfellow et al., 2015; Vassilev et al., 2025
Explainability tools exist but are not integrated into response workflows	Analysts cannot trust or act on black box alerts	Risk based human oversight with multi factor escalation logic	Layer 4: Explainability and Human Oversight	Arrieta et al., 2020; Mohale and Obagbuwa, 2025
Incident response is often disconnected from detection and intelligence	Response actions lack context from detection and threat intelligence	Adaptive, intelligence informed incident response with automated and human pathways	Layer 5: Adaptive Incident Response	Saeed et al., 2023; Sewak et al., 2023
Organisations struggle to learn from incidents and adapt defences	No feedback from incident outcomes to threat intelligence or model improvement	Continuous learning, model retraining, and resilience assessment	Layer 6: Resilience and Continuous Learning	NIST, 2024; Tabassi, 2023

9.11 Mapping Tables 1 to 3 to RA CyDPF

Tables 1 to 3 provide threat, AI technique, and privacy risk taxonomies. The following mapping makes explicit how findings from these tables directly feed into the six RA CyDPF layers:

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

Table 9. Traceability from Taxonomies to Framework Layers

Threat/technique/risk identified (from Tables 1 to 3)	Corresponding control	RA CyDPF layer	Responsible actor	Expected resilience outcome
AI generated phishing / BEC (Table 1)	NLP based detection + human verification protocol	Layer 3 + Layer 4	SOC analyst	Reduced mean time to detect; fewer successful social engineering breaches
Deepfake impersonation (Table 1)	Multimodal detection + executive verification protocol	Layer 3 + Layer 4	SOC analyst + communications team	Reduced authorisation fraud; faster executive alert
Adversarial evasion (Table 1)	Adversarial robustness testing + model retraining	Layer 3 + Layer 6	AI/ML engineer	Sustained detection accuracy under adversarial perturbation
Membership inference (Table 3)	Differential privacy + output perturbation	Layer 2 + Layer 3	Privacy officer	Bounded privacy loss; reduced inference success rate
Model extraction (Table 3)	Rate limiting + query monitoring	Layer 2 + Layer 3	Security architect	Reduced intellectual property exposure
Gradient leakage in FL (Table 3)	Secure aggregation + gradient clipping	Layer 2	Privacy officer + AI engineer	Protected local training data in distributed settings
Prompt injection (Table 3)	Input sanitisation + privilege separation	Layer 2 + Layer 4	AI security engineer	Reduced model manipulation risk
Supervised classification privacy risk (Table 2)	Data minimisation + pseudonymisation	Layer 2	Data steward	Reduced sensitive attribute exposure
Deep learning interpretability gap (Table 2)	Feature attribution + analyst validation	Layer 4	SOC analyst	Improved trust and auditability
RL adaptive response risk (Table 2)	Human in the loop for high impact actions	Layer 4 + Layer 5	Incident commander	Prevented unauthorised automated containment

10. FRAMEWORK EVALUATION AND VALIDATION STRATEGY

Because RA CyDPF is developed through literature synthesis rather than experimentation, this section deliberately separates what can be stated now, a conceptual evaluation model and a set of candidate metrics, from what would require future empirical work. No experiment has been conducted for this paper, and no accuracy, precision, recall, false positive, or response time figures are reported as findings; the material below defines how such figures could be obtained and interpreted in subsequent empirical research.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

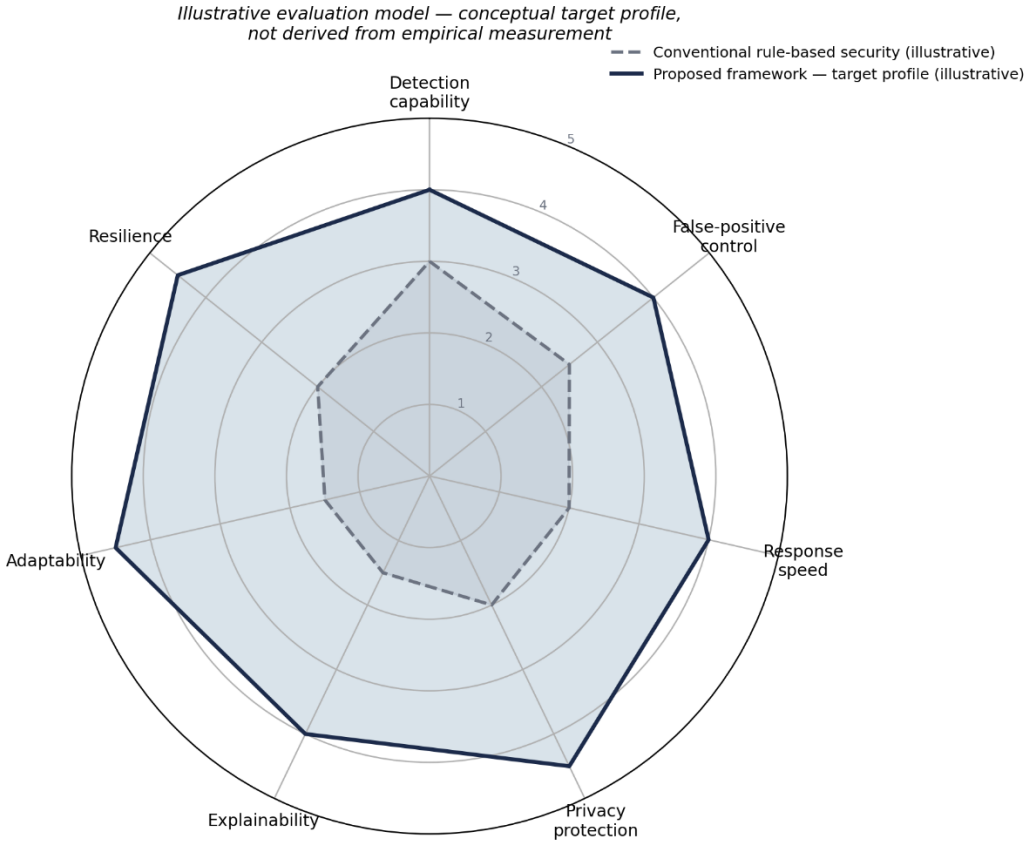

Figure 5. Proposed evaluation dimensions for assessing resilient AI-driven cybersecurity capability (illustrative target profile).

10.1 Conceptual Evaluation Dimensions

Figure 6 presents an illustrative evaluation model comparing a conventional rule based security posture against the target profile the proposed framework is intended to achieve across seven evaluation dimensions. The values plotted are conceptual target positions reflecting the qualitative expectations documented in the literature synthesised in Sections 3 and 6, for example, that explainability and privacy protection are structurally prioritised in RA CyDPF in a way that conventional rule based systems generally are not, and are explicitly not the output of a benchmark experiment. The figure should be labelled "Conceptual Target Profile, Not Empirical Results" in its title, caption, legend, and surrounding discussion.

Table 10. Proposed Evaluation Metrics by Dimension

Dimension	Metric	Measurement approach	Expected interpretation	Relevant framework component
Security effectiveness	Precision, recall, F1 score	Benchmark against labelled attack/benign datasets, including AI generated attack samples	Higher values indicate more reliable identification of true threats	Layer 3: AI Based Detection and Analytics

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

Privacy effectiveness	Differential privacy budget (epsilon); membership inference attack success rate	Formal privacy accounting; adversarial privacy attack simulation	Lower attack success rate and bounded epsilon indicate stronger protection	Layer 2: Privacy and Data Governance
Explainability/human factors	Analyst rated explanation usefulness; explanation consistency; inter rater reliability	Structured analyst survey and inter rater reliability testing	Higher ratings indicate explanations support, rather than hinder, decisions	Layer 4: Explainability and Human Oversight
Operational efficiency	Mean time to detect; mean time to contain; analyst alert burden	Timestamped workflow logging from detection to containment; production shadow deployment	Shorter intervals and reduced burden indicate improved operational resilience	Layer 5: Adaptive Incident Response
Adversarial robustness	Detection performance retained after adversarial perturbation or concept drift	Adversarial robustness testing per NIST AI 100 2e2025 taxonomy	Smaller performance degradation indicates greater adaptability	Layer 3 and Layer 6
Adaptability	Time to integrate new threat signatures; model retraining cycle time	Controlled simulation of novel attack introduction	Shorter integration times indicate stronger adaptive capacity	Layer 6: Resilience and Continuous Learning
Organisational resilience	Recovery time objective attainment; recurrence rate of similar incidents; post incident review scores	Post incident review against defined recovery targets	Faster recovery and lower recurrence indicate stronger organisational resilience	Layer 6: Resilience and Continuous Learning

10.2 Baseline Comparison Approach

Where possible, future evaluation should distinguish between:

- **Security effectiveness:** Conventional SOC / existing security architecture versus SOC implementing RA CyDPF controls.
- **Privacy effectiveness:** Centralised detection with raw data versus privacy preserving detection under RA CyDPF Layer 2.
- **Explainability/human factors:** Black box detection alerts versus explainable alerts with Layer 4 oversight.
- **Operational efficiency:** Manual triage workflow versus AI assisted triage with human oversight.
- **Adversarial robustness:** Standard trained models versus adversarially tested models under Layer 6.
- **Adaptability:** Static model deployment versus continuous learning under Layer 6.
- **Organisational resilience:** Pre framework incident response versus post framework structured feedback.

Any future comparison would require controlled or appropriately matched environments. Organisations should not be compared across different sectors, maturity levels, or threat exposure profiles without statistical matching or stratification.

10.3 Continuous Resilience Cycle

Figure 6 depicts the continuous resilience cycle that Layer 6 is intended to sustain over time, showing how lessons learned from a given incident feed structurally back into threat intelligence for the next cycle, rather than resilience being treated as a one time property established during initial deployment.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

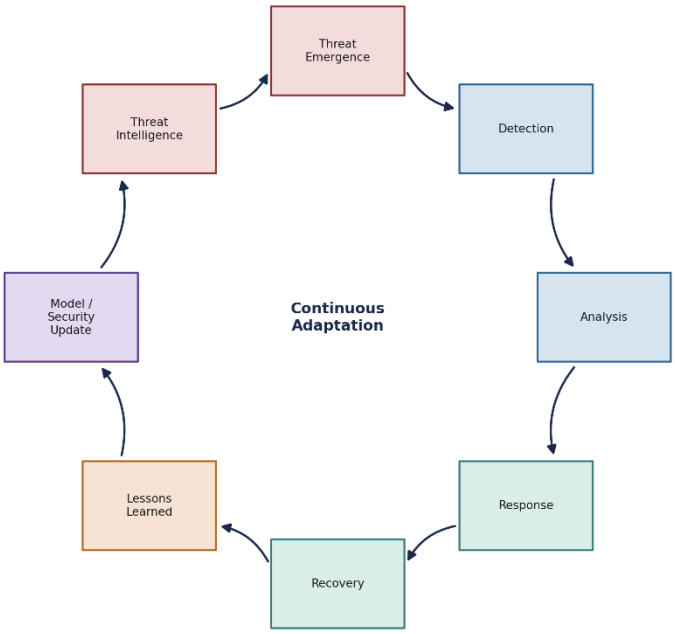

Figure 6. Continuous learning and cyber-resilience feedback cycle.

11. DISCUSSION

The synthesis presented in this paper carries theoretical and practical implications beyond the specific architecture proposed in Section 9. Theoretically, the paper contributes to the cybersecurity literature by making explicit an interaction that is frequently treated implicitly: the accuracy-privacy-explainability trade-off that emerges when AI is embedded within security operations. Sections 3 and 7 demonstrated that this trade-off is documented piecemeal across separate literatures privacy-preserving machine learning researchers document the accuracy cost of differential privacy and secure aggregation; explainability researchers document the interpretability cost of high-accuracy deep learning; and cybersecurity researchers document the detection-sensitivity cost of data minimisation but is rarely modelled as a single, three-way trade-off requiring an explicit governance decision. Figure 4's interaction model, and the layered structure of RA-CyDPF more broadly, offer one way of representing this trade-off as a design decision made deliberately at Layer 2 and Layer 4, rather than as an unexamined default inherited from whichever technique an organisation happens to adopt first.

Practically, the framework's emphasis on embedding privacy controls before, rather than after, AI-based analytics (Layer 2 preceding Layer 3) responds directly to a pattern observed across the reviewed literature, in which privacy-preserving techniques are frequently evaluated as an add-on to an already-designed detection pipeline rather than as a structural precondition for it. Similarly, the explicit placement of explainability and human oversight as a distinct architectural layer, rather than as a post-hoc interpretability technique applied to individual model outputs, responds

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

to the recurring finding in the explainable-AI literature that interpretability tools are most effective when integrated into the decision workflow rather than appended to it (Mohale & Obagbuwa, 2025; Rudin, 2019).

The privacy implications of the proposed framework are addressed directly through Layer 2 and through the residual-concern column of Table 3, which acknowledges that no combination of currently available techniques eliminates privacy risk from AI-enabled security tooling; the framework instead offers a structure for managing that risk explicitly and transparently rather than for eliminating it. The cybersecurity implications centre on the framework's attempt to close the adversarial-robustness gap identified in Section 6, by embedding adversarial testing within Layer 6 as an ongoing organisational practice rather than a one-time pre-deployment check. Organisationally, the framework implies a need for cross-functional coordination between security operations, privacy or data-protection functions, and AI governance functions that, in many organisations, currently operate with limited structural connection a coordination requirement that is itself a practical implication of the fragmentation documented in Section 8 rather than a novel observation specific to this framework.

AI governance considerations run throughout the framework rather than being confined to a single layer, consistent with the cross-cutting concerns depicted in Figure 2. The framework's reliance on human oversight at Layer 4 is intended to operationalise, rather than merely reference, the human-in-the-loop principle emphasised in the NIST AI Risk Management Framework's Govern function (Tabassi, 2023) and in the broader explainable-AI literature. Resilience, in the sense developed through Layer 6 and Figure 7, is treated throughout this paper as a continuously maintained property rather than a static certification, consistent with the adaptive orientation of the NIST Cybersecurity Framework's Respond and Recover functions (NIST, 2024).

12. IMPLICATIONS

12.1 Implications for Cybersecurity Professionals and Security Operations Centres

For practitioners, the framework offers a structured way to evaluate whether an existing security stack addresses AI-specific risk comprehensively or only partially for example, whether adversarial-robustness testing (Layer 6) is performed with the same rigour as conventional penetration testing, and whether explainability requirements (Layer 4) are applied consistently across automated and human-reviewed decisions.

12.2 Implications for Organisations and Risk Owners

For organisational risk owners, the framework highlights that AI adoption within cybersecurity is not risk-neutral: the same capabilities that improve detection can expand privacy exposure if not governed deliberately. The framework supports a risk-based justification for investment in privacy-preserving analytics and explainability tooling as core security infrastructure rather than optional compliance overhead.

12.3 Implications for Policymakers and Regulators

For policymakers, the fragmentation documented in Section 8 suggests that guidance addressing cybersecurity, AI risk, and privacy in isolation may leave organisations without a clear path for reconciling conflicting requirements for example, data-retention obligations that support incident investigation may conflict with data-minimisation obligations intended to limit privacy exposure. Cross-referencing between cybersecurity, AI-governance, and privacy regulatory guidance, of the kind partially achieved by NIST's own cross-framework mappings, appears to merit continued institutional investment.

12.4 Implications for AI Developers

For developers of AI-based security tools, the framework underscores that detection accuracy alone is an insufficient product requirement; explainability outputs, privacy-preserving training options, and adversarial-robustness documentation are structural expectations under the proposed architecture rather than differentiating features.

12.5 Implications for Privacy Professionals and Security Operations Centres

For privacy professionals, the framework's Layer 2 provides a concrete point of engagement with security-tooling procurement and deployment decisions, positioning privacy impact assessment as a precondition for, rather than a review of, AI-based detection deployment.

13. Limitations and Future Research

This paper is subject to several limitations that should inform how its contribution is interpreted. First, and most fundamentally, RA-CyDPF has not been empirically validated: no organisation has implemented the framework, and the evaluation dimensions and illustrative target profile presented in Section 10 describe what should be measured,

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

not what has been measured. Second, the literature review, while structured and covering a substantial cross-section of the relevant scholarly and institutional literature, was not conducted as a formally registered systematic review with a documented database-level record count; the source base, while extensive, may not be fully exhaustive of every relevant national or sector-specific guidance document. Third, the framework's six-layer structure, while synthesised from convergent findings across the reviewed literature, necessarily involves interpretive judgment in how boundaries between layers are drawn, and alternative decompositions may be equally defensible. Fourth, the conceptual risk-prioritisation matrix (Figure 5) and illustrative evaluation model (Figure 6) are explicitly qualitative placements intended to support reasoning about the framework rather than calibrated instruments, and should not be treated as validated risk-scoring tools.

Future research should prioritise: (a) empirical validation of RA-CyDPF through pilot deployment within one or more security operations centres, using the metrics proposed in Table 7; (b) construction or identification of benchmark datasets that specifically represent AI-generated attacks (AI-crafted phishing corpora, adversarially perturbed intrusion-detection traffic, deepfake-derived social-engineering scenarios) against which the framework's detection layer could be tested; (c) structured, real-world security operations centre testing of the explainability and human-oversight layer, including analyst trust and workload studies; (d) adversarial testing of the framework's detection and response layers against the attack taxonomy documented in Vassilev et al. (2025); (e) deeper technical integration of federated learning and differential-privacy techniques within the privacy-governance layer, building on the trade-offs documented in Demelius et al. (2025) and Mothukuri et al. (2021); (f) formal evaluation of explainability quality using established human-computer interaction methods rather than technical interpretability metrics alone; and (g) cross-sector deployment studies examining whether the framework's six-layer structure requires sector-specific adaptation for contexts such as healthcare, finance, or critical infrastructure.

14. CONCLUSION

Artificial intelligence has become a structural feature of the contemporary cybersecurity landscape, simultaneously expanding attacker capability, improving defensive detection, and introducing privacy and security risks intrinsic to AI systems themselves. This paper has argued, on the basis of a structured synthesis of the scholarly and institutional literature, that these three dimensions are addressed by largely separate bodies of work and separate institutional frameworks, producing a fragmentation that leaves organisations without integrated guidance for managing AI-driven cybersecurity risk alongside data privacy. In response, the paper has proposed the Resilient AI-Driven Cybersecurity and Data Privacy Framework, a six-layer conceptual architecture connecting threat intelligence acquisition, privacy governance, AI-based detection, explainability and human oversight, adaptive incident response, and continuous resilience within a single cyclical structure. The framework's contribution lies not in any individual technical component each of which draws on established literature but in the explicit integration of these components and in the identification of where their interactions require deliberate governance decisions rather than default technical configuration.

The framework, as presented, is a conceptual synthesis rather than a validated system, and its practical value will ultimately depend on the empirical work outlined in Section 13. It is offered as a structured starting point for organisations, researchers, and policymakers seeking to reason about AI-driven cybersecurity and data privacy as a single interconnected problem rather than as a set of adjacent but separately governed concerns.

REFERENCES

1) Ahi, K., & Valizadeh, S. (2026). Large language models (LLMs) and generative AI in cybersecurity and privacy: A survey of dual-use risks, AI-generated malware, explainability, and defensive strategies. arXiv preprint arXiv:2607.06963.
2) Ahsan, M., Nygard, K. E., Gomes, R., Chowdhury, M. M., Rifat, N., & Connolly, J. F. (2022). Cybersecurity threats and their mitigation approaches using machine learning—A review. Journal of Cybersecurity and Privacy, 2(3), 527–555.
3) Alevizos, L., & Dekker, M. (2024). Towards an AI-enhanced cyber threat intelligence processing pipeline. Electronics, 13(11), 2021. https://doi.org/10.3390/electronics13112021

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

4) Alnajim, A. M., Habib, S., Islam, M., Thwin, S. M., & Alotaibi, F. (2023). A comprehensive survey of cybersecurity threats, attacks, and effective countermeasures in industrial internet of things. Technologies, 11(6), 161. https://doi.org/10.3390/technologies11060161
5) Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.
6) Begou, N., Vinoy, J., Duda, A., & Korczyński, M. (2023). Exploring the dark side of AI: Advanced phishing attack design and deployment using ChatGPT. In 2023 IEEE Conference on Communications and Network Security (CNS) (pp. 1–6). IEEE. https://doi.org/10.1109/CNS59707.2023.10288940
7) Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 1175–1191). ACM.
8) Carlini, N., Tramèr, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, Ú., Oprea, A., & Raffel, C. (2021). Extracting training data from large language models. In Proceedings of the 30th USENIX Security Symposium (pp. 2633–2650).
9) Coppolino, L., D'Antonio, S., Mazzeo, G., & Uccello, F. (2025). The good, the bad, and the algorithm: The impact of generative AI on cybersecurity. Neurocomputing, 623, 129406.
10) CrowdStrike. (2025). 2025 Global Threat Report. CrowdStrike Holdings.
11) Dasgupta, D., Akhtar, Z., & Sen, S. (2022). Machine learning in cybersecurity: A comprehensive survey. The Journal of Defense Modeling and Simulation, 19(1), 57–106.
12) de Azambuja, A. J. G., Plesker, C., Schützer, K., Anderl, R., Schleich, B., & Almeida, V. R. (2023). Artificial intelligence-based cyber security in the context of Industry 4.0—A survey. Electronics, 12(8), 1920. https://doi.org/10.3390/electronics12081920
13) Demelius, L., Kern, R., & Trügler, A. (2025). Recent advances of differential privacy in centralized deep learning: A systematic survey. ACM Computing Surveys, 57(6), 1–28.
14) Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.
15) European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 (General Data Protection Regulation). Official Journal of the European Union, L119, 1–88.
16) Fredrikson, M., Jha, S., & Ristenpart, T. (2015). Model inversion attacks that exploit confidence information and basic countermeasures. In Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (pp. 1322–1333). ACM.
17) Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In Proceedings of the International Conference on Learning Representations (ICLR).
18) Ilić, L., Šijan, A., Predić, B., Viduka, D., & Karabašević, D. (2024). Research trends in artificial intelligence and security—Bibliometric analysis. Electronics, 13(12), 2288. https://doi.org/10.3390/electronics13122288
19) International Organization for Standardization. (2022). ISO/IEC 27001:2022 — Information security, cybersecurity and privacy protection — Information security management systems — Requirements. ISO.
20) International Organization for Standardization. (2023). ISO/IEC 42001:2023 — Information technology — Artificial intelligence — Management system. ISO.
21) Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. Information Fusion, 97, 101804.
22) Kilincer, I. F., Ertam, F., & Sengur, A. (2021). Machine learning methods for cyber security intrusion detection: Datasets and comparative study. Computer Networks, 188, 107840.
23) Kumar, V., & Sinha, D. (2023). Synthetic attack data generation model applying generative adversarial network for intrusion detection. Computers & Security, 125, 103054.

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

24) Mahbooba, B., Timilsina, M., Sahal, R., & Serrano, M. (2021). Explainable Artificial Intelligence (XAI) to enhance trust management in intrusion detection systems using decision tree model. Complexity, 2021, 6634811. https://doi.org/10.1155/2021/6634811
25) McMahan, H. B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS) (pp. 1273–1282).
26) Mohale, V. Z., & Obagbuwa, I. C. (2025). A systematic review on the integration of explainable artificial intelligence in intrusion detection systems to enhancing transparency and interpretability in cybersecurity. Frontiers in Artificial Intelligence, 8, 1526221. https://doi.org/10.3389/frai.2025.1526221
27) Mohamed, N. (2023). Current trends in AI and ML for cybersecurity: A state-of-the-art survey. Cogent Engineering, 10(2), 2272358. https://doi.org/10.1080/23311916.2023.2272358
28) Mothukuri, V., Parizi, R. M., Pouriyeh, S., Huang, Y., Dehghantanha, A., & Srivastava, G. (2021). A survey on security and privacy of federated learning. Future Generation Computer Systems, 115, 619–640.
29) Moustafa, N., Koroniotis, N., Keshk, M., Zomaya, A. Y., & Tari, Z. (2023). Explainable intrusion detection for cyber defences in the Internet of Things: Opportunities and solutions. IEEE Communications Surveys & Tutorials, 25(3), 1775–1807. https://doi.org/10.1109/COMST.2023.3280465
30) Mubarak, R., Alsboui, T., Alshaikh, O., Inuwa-Dutse, I., Khan, S., & Parkinson, S. (2023). A survey on the detection and impacts of deepfakes in visual, audio, and textual formats. IEEE Access, 11, 144497–144529.
31) Mustak, M., Salminen, J., Mäntymäki, M., Rahman, A., & Dwivedi, Y. K. (2023). Deepfakes: Deceptions, mitigations, and opportunities. Journal of Business Research, 154, 113368.
32) National Institute of Standards and Technology. (2020). NIST Privacy Framework: A tool for improving privacy through enterprise risk management (Version 1.0). NIST CSWP 01162020. https://doi.org/10.6028/NIST.CSWP.01162020
33) National Institute of Standards and Technology. (2024). The NIST Cybersecurity Framework (CSF) 2.0. NIST CSWP 29. https://doi.org/10.6028/NIST.CSWP.29
34) Proofpoint. (2024). 2024 State of the Phish report. Proofpoint Inc.
35) Raji, A., Olawore, A., Mustapha, A., & Joseph, J. (2023). Integrating artificial intelligence, machine learning, and data analytics in cybersecurity: A holistic approach to advanced threat detection and response. World Journal of Advanced Research and Reviews, 20(3), 2005–2024.
36) Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero Trust Architecture. NIST Special Publication 800-207. https://doi.org/10.6028/NIST.SP.800-207
37) Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206–215.
38) Saeed, S., Suayyid, S. A., Al-Ghamdi, M. S., Al-Muhaisen, H., & Almuhaideb, A. M. (2023). A systematic literature review on cyber threat intelligence for organizational cybersecurity resilience. Sensors, 23(16), 7273. https://doi.org/10.3390/s23167273
39) Sarker, I. H. (2022). Machine learning for intelligent data analysis and automation in cybersecurity: Current and future prospects. Annals of Data Science, 10(6), 1473–1498.
40) Sarker, I. H. (2023). Multi-aspects AI-based modeling and adversarial learning for cybersecurity intelligence and robustness: A comprehensive overview. Security and Privacy, 6(5), e295.
41) Sarker, I. H., Janicke, H., Maglaras, L., & Camtepe, S. (2023). Data-driven intelligence can revolutionize today's cybersecurity world: A position paper. arXiv preprint arXiv:2308.05126.
42) Schmitt, M., & Flechais, I. (2024). Digital deception: Generative artificial intelligence in social engineering and phishing. AI & Society. Advance online publication.
43) Sewak, M., Sahay, S. K., & Rathore, H. (2023). Deep reinforcement learning in the advanced cybersecurity threat detection and protection. Information Systems Frontiers, 25(2), 589–611.
44) Shokri, R., Stronati, M., Song, C., & Shmatikov, V. (2017). Membership inference attacks against machine learning models. In 2017 IEEE Symposium on Security and Privacy (SP) (pp. 3–18). IEEE.
45) Tabassi, E. (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). NIST AI 100-1. https://doi.org/10.6028/NIST.AI.100-1

iJETRM
INTERNATIONAL JOURNAL OF ENGINEERING TECHNOLOGY RESEARCH & MANAGEMENT (IJETRM)
Peer-Reviewed | Open Access | International Journal
https://ijetrm.com/

46) Tounsi, W., & Rais, H. (2018). A survey on technical threat intelligence in the age of sophisticated cyber attacks. Computers & Security, 72, 212–233.
47) Tramèr, F., Zhang, F., Juels, A., Reiter, M. K., & Ristenpart, T. (2016). Stealing machine learning models via prediction APIs. In Proceedings of the 25th USENIX Security Symposium (pp. 601–618).
48) Vaddadi, S. A., Vallabhaneni, R., & Whig, P. (2023). Utilizing AI and machine learning in cybersecurity for sustainable development through enhanced threat detection and mitigation. International Journal of Sustainable Development through AI, ML and IoT, 2(2), 1–8.
49) Vassilev, A., Oprea, A., Fordyce, A., Anderson, H., Davies, X., & Hamin, M. (2025). Adversarial Machine Learning: A taxonomy and terminology of attacks and mitigations. NIST AI 100-2e2025. https://doi.org/10.6028/NIST.AI.100-2e2025
50) Wang, T., Liao, X., Chow, K. P., Lin, X., & Wang, Y. (2024). Deepfake detection: A comprehensive survey from the reliability perspective. ACM Computing Surveys, 57(3), 1–35.
51) World Economic Forum. (2024). Global Risks Report 2024. World Economic Forum.
52) Yao, Y., Duan, J., Xu, K., Cai, Y., Sun, Z., & Zhang, Y. (2024). A survey on large language model (LLM) security and privacy: The good, the bad, and the ugly. High-Confidence Computing, 4(2), 100211. https://doi.org/10.1016/j.hcc.2024.100211
53) Yin, X., Zhu, Y., & Hu, J. (2021). A comprehensive survey of privacy-preserving federated learning: A taxonomy, review, and future directions. ACM Computing Surveys, 54(6), 1–36.
54) Zhang, J., Zhu, H., Wang, F., Zhao, J., Xu, Q., & Li, H. (2022). Security and privacy threats to federated learning: Issues, methods, and challenges. Security and Communication Networks, 2022, 2886795.
55) Zhang, Z., Ning, H., Shi, F., Farha, F., Xu, Y., Xu, J., Zhang, F., & Choo, K.-K. R. (2022). Artificial intelligence in cyber security: Research advances, challenges, and opportunities. Artificial Intelligence Review, 55, 1029–1053.
56) Zhu, L., Liu, Z., & Han, S. (2019). Deep leakage from gradients. Advances in Neural Information Processing Systems, 32.

