Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 3, 2026

Artificial Intelligence in Cybersecurity and Data Protection: A Study of the Information Technology (IT) Sector

Dr. Mubeen Fatima[1], Junaid Shaukat*[2], Muhammad Faraz Tahir Bhatti[3]

[1]Associate Professor, Riphah International University, Islamabad
[2]Research Scholar, Master of Computer Science, Pir Mehr Ali Shah Arid Agriculture University, Rawalpindi
[3]Research Scholar, BS Information Technology, Minhaj University Lahore

DOI: https://doi.org/

Keywords
Artificial Intelligence; Cybersecurity; Data Protection; Information Technology Sector; Machine Learning; Generative AI; Threat Detection; Privacy; AI Governance

Article History
Received on 20 Jan 2026
Accepted on 20 Mar 2026
Published on 30 Mar 2026

Copyright @Author
Corresponding Author: *

Abstract
The growing dependence of the information technology (IT) sector on cloud platforms, software services, connected infrastructure, and digital customer data has increased the importance of effective cybersecurity and data protection. Artificial intelligence (AI), including machine learning, deep learning, natural language processing, and generative models, offers opportunities to detect anomalous activity, prioritize alerts, support incident response, and monitor sensitive information. At the same time, AI introduces risks related to privacy, model manipulation, bias, explainability, operational dependence, and misuse by attackers. This paper examines the role of AI in cybersecurity and data protection in the IT sector through a structured narrative review that is supported by secondary quantitative evidence from recent authoritative sources, including the IBM Cost of a Data Breach Report 2026, the Verizon 2026 Data Breach Investigations Report, the FBI Internet Crime Report 2025, the ENISA Threat Landscape 2025 and 2026, and the ISC2 2025 Cybersecurity Workforce Study. The evidence shows that the average global breach cost reached a record USD 4.99 million in 2026, that organizations making extensive use of security AI and automation reported savings of USD 1.93 million per breach, and that attackers increasingly use AI to scale phishing and fraud. It also shows that AI systems have become a new attack surface and that adoption is constrained by skills shortages and weak governance. The paper concludes that AI is best understood as an augmentation of, rather than a replacement for, established security controls and qualified human judgment. It proposes a seven-stage adoption framework, a governance mapping to current standards, and a set of measurable indicators, and it identifies directions for primary empirical research.

INTRODUCTION

Background

Information technology organizations operate the systems, applications, networks, cloud services, and data environments on which most other economic sectors now depend. Their responsibilities commonly include software development, infrastructure management, cloud hosting, managed services, technical support, cybersecurity services, and the processing of personal or commercially sensitive information. As these environments become more interconnected, a cyber incident at one provider can extend beyond a single organization to customers, business partners, and dependent public services. The European Union Agency for Cybersecurity (ENISA, 2026) makes this point directly, describing cyber dependencies, including supply chains and third-party service providers, as a defining feature of the current threat environment.

Conventional cybersecurity practice relies on controls such as access management, firewalls, endpoint protection, vulnerability management, logging, encryption, and human-led investigation. These controls remain essential, but the scale and speed of modern IT environments make it difficult for security teams to review every alert and to recognize subtle patterns across very large volumes of network, endpoint, identity, cloud, and application data. AI-based tools can assist by learning patterns from data, identifying deviations from normal behavior, correlating events, classifying suspicious content, summarizing incidents, and recommending or initiating responses. The commercial and institutional interest in these capabilities is considerable. In the ISC2 (2025) workforce study, 28% of surveyed professionals reported that AI tools were already integrated into their daily operations, and a further 69% reported that they were evaluating, testing, or implementing them.

The application of AI to cybersecurity is not without limits. Models may generate false positives or miss novel attacks, and their performance depends on the relevance and quality of the data on which they are trained and evaluated. Attackers may attempt to evade detection, poison training data, or exploit AI-enabled systems directly, and the U.S. National Institute of Standards and Technology has published a detailed taxonomy of such attacks (Vassilev et al., 2025). Attackers also use AI offensively. The FBI's Internet Crime Complaint Center (IC3) included a dedicated section on AI for the first time in its 2025 annual report, linking more than 22,000 complaints to losses of nearly USD 893 million (Federal Bureau of Investigation [FBI], 2026). Furthermore, security monitoring can involve sensitive employee, customer, and user information, which creates a need to balance threat prevention with privacy, proportionality, and lawful data processing. Therefore, the use of AI must be assessed against both the security outcomes it delivers and the new risks the technology introduces.

This paper adapts the broad topic of AI, cybersecurity, and data protection to the operational context of the IT sector. It explains the principal applications of AI, reviews the benefits and constraints reported in the scholarly and professional literature, synthesizes recent quantitative evidence on those benefits and constraints, and proposes practical guidance for responsible adoption. The paper is a literature-based analysis that draws on published secondary data. It does not claim to have collected primary survey data or to have experimentally measured the effectiveness of any specific AI tool.

Research Objectives

The study pursues the following objectives:

To examine the principal applications of AI in cybersecurity and data protection in the IT sector.

To discuss the potential contribution of AI to threat detection, incident response, and security

monitoring, using recent quantitative evidence where available.

To identify the technical, organizational, regulatory, and ethical challenges associated with AI-enabled security, including the security of AI systems themselves.

To propose practical recommendations and measurable indicators for responsible AI adoption by IT organizations.

Research Questions

Four research questions guide the analysis:

RQ1: How can AI support cybersecurity and data protection in IT organizations?

RQ2: What does recent evidence show about the benefits of AI for threat detection, alert prioritization, incident response, and the cost of breaches?

RQ3: What technical, organizational, and ethical challenges may limit AI adoption in the IT sector, and how is AI itself changing the threat landscape?

RQ4: What governance and implementation practices can help IT organizations use AI responsibly?

Significance of the Study

The study is relevant to IT managers, cybersecurity professionals, software and cloud service providers, data protection officers, and researchers. It brings together the operational promise of AI and its associated risks in one sector-focused discussion and anchors that discussion in current published evidence rather than in general assertions. The paper may help organizations consider AI adoption as part of a wider security program rather than as a stand-alone product or as a substitute for fundamental controls. It is also timely. The 2026 editions of several major industry and government reports, including those of IBM, Verizon, and ENISA, now treat AI as a central rather than a peripheral theme, and national and international guidance on AI security has matured considerably since 2023.

Literature Review

AI and Cybersecurity: Conceptual Foundations

AI is an umbrella term for methods that perform tasks such as classification, pattern recognition, prediction, language processing, and decision support. In cybersecurity, machine-learning models analyze network flows, endpoint events, authentication logs, malware characteristics, and other telemetry to distinguish routine behavior from potentially malicious activity. Sarker (2021) describes AI-driven cybersecurity as the use of data-driven models to generate what he calls security intelligence, and he groups the relevant techniques into supervised, unsupervised, and reinforcement-based families, with deep learning applied where large volumes of data are available. Kaur et al. (2023), in a structured review of AI for cybersecurity, similarly emphasize that the field spans intrusion detection, malware analysis, phishing detection, and threat intelligence, and that many proposed techniques remain difficult to transfer into operational settings.

Buczak and Guven (2016) reviewed data-mining and machine-learning methods for intrusion detection and described the range of approaches used to analyze security data. Their review also illustrates that detection performance depends on the characteristics of the dataset, the learning method, and the evaluation design. This dependence on data and evaluation is a recurring theme. For IT organizations, AI may help correlate events across endpoints, cloud workloads, identity systems, and applications. It can also support security operations center (SOC) analysts by grouping related alerts and ranking them according to estimated risk. These capabilities can reduce repetitive manual work, but the quality of recommendations should be evaluated against relevant operational conditions. A model tested in one environment may not perform equally well in another with different users, software, attack patterns, or logging practices.

Machine Learning for Detection: Promise and Practical Constraints

The deployment of machine learning for detection has a long history of both promise and disappointment. Sommer and Paxson (2010) argued that network intrusion detection is a considerably harder problem for machine learning than many other application areas. Their reasons included the high cost of errors, the diversity and constant change of legitimate traffic, the difficulty of obtaining representative training data, and the challenge of evaluating systems realistically. Because normal behavior is heterogeneous, an anomaly is not necessarily an attack, and a detector that is accurate on a benchmark dataset may generate an unmanageable number of alerts in production. These observations remain relevant to modern IT environments, in which cloud services, remote work, and rapid software release cycles continually change what counts as normal.

More recent work has examined how research practice itself can overstate real-world performance. Arp et al. (2022) reviewed machine-learning studies in computer security and identified recurring pitfalls, including sampling bias, label inaccuracy, data snooping, spurious correlations, unrealistic base rates, and inappropriate performance measures. They found that such pitfalls were widespread in top-tier security papers. The implication for practitioners is straightforward. Vendor claims and published accuracy figures should be treated as hypotheses to be tested in the buyer's own environment, using realistic class imbalance, time-based evaluation splits, and metrics that reflect analyst workload rather than accuracy alone. A related technical issue is concept drift, the change over time in the statistical relationship between model inputs and the outcomes the model is meant to predict. Gama et al. (2014) survey methods for detecting and adapting to drift. In security, drift is the normal condition rather than an exception, because both legitimate usage and attacker behavior evolve.

AI for Threat Detection and Incident Response

Machine-learning techniques are used in intrusion detection, malware classification, spam and phishing identification, and behavioral analytics. Deep-learning methods can identify complex patterns in large datasets, while natural language processing can help analyze suspicious messages, threat reports, and security tickets. AI-enabled systems may also assist incident response by summarizing alerts, suggesting investigative steps, enriching indicators with threat intelligence, and automating carefully bounded actions such as blocking a known-malicious address or quarantining a file.

Automation should be designed with appropriate safeguards. A false positive may interrupt legitimate services, while a false negative may leave a threat undetected. High-impact actions, such as disabling production systems, revoking privileged access, or isolating critical infrastructure, should therefore be governed by clear approval rules, escalation paths, rollback procedures, and audit logs. Human review remains important where the context is ambiguous or the consequences of an incorrect action are substantial. The NIST Cybersecurity Framework 2.0 organizes cybersecurity outcomes under the functions Govern, Identify, Protect, Detect, Respond, and Recover (National Institute of Standards and Technology [NIST], 2024b), and AI can contribute to several of them. Its contribution is most direct in Detect and Respond, but it also supports Identify through asset discovery and risk scoring, and Protect through data classification and access analytics.

AI and Data Protection

IT organizations may process source code, customer records, credentials, business documents, usage logs, and other sensitive

information. AI can support data protection through anomaly detection, data classification, monitoring of unusual access, identification of possible data exfiltration, and risk-based prioritization. It may also help security teams discover patterns that conventional rules do not readily capture, for example by learning the typical download volume of a developer account and flagging deviations from it.

However, AI systems themselves may require access to extensive operational data. If the collection, retention, or reuse of this data is poorly governed, the security tool can become a source of privacy risk. Research on privacy attacks against machine-learning models shows that the concern is not hypothetical. Shokri et al. (2017) demonstrated membership inference attacks, in which an adversary with query access to a model can determine whether a particular record was part of its training data. Carlini et al. (2021) showed that training data, including personal information, can be extracted from large language models. Organizations should therefore minimize data collection, limit access, define retention periods, protect training and inference data, and assess whether sensitive data can be masked or processed in a more privacy-preserving way. AI-generated outputs should also be checked for unintended disclosure of confidential information. The joint guidance issued by the U.S. National Security Agency, the Cybersecurity and Infrastructure Security Agency, the FBI, and partner agencies in 2025 identifies the data supply chain, maliciously modified data, and data drift as the principal data-centric risks across the AI lifecycle, and it recommends measures such as provenance tracking, digital signatures for trusted data revisions, and trusted infrastructure (National Security Agency Artificial Intelligence Security Center et al., 2025).

Generative AI and Large Language Models in Security Operations

Since 2023, large language models (LLMs) and other generative systems have been introduced into security workflows. Typical uses include summarizing alerts and incident timelines, translating natural-language questions into log queries, drafting incident reports, assisting with code review and vulnerability triage, and powering conversational assistants for analysts. The NIST Generative Artificial Intelligence Profile (NIST, 2024a) identifies risks that are specific to generative systems, among them confabulation (the generation of plausible but false content), information security risks, data privacy risks, and the possibility that such systems lower barriers for malicious actors. In security settings, confabulation is particularly consequential, because an analyst who acts on a fabricated indicator or an incorrect remediation step can waste scarce time or cause an outage.

LLM-integrated applications also add new attack paths. Greshake et al. (2023) described indirect prompt injection, in which malicious instructions embedded in content that an application retrieves, such as a web page, a document, or an email, can manipulate the behavior of an LLM-integrated application without any direct access by the attacker to the application's interface. This matters for security tooling that ingests untrusted content by design, such as phishing triage assistants that read suspicious emails. The OWASP Top 10 for LLM Applications (OWASP, 2025) catalogs related risks, including prompt injection, sensitive information disclosure, supply chain weaknesses, and excessive agency, the last of which arises when an AI system is granted more autonomy or permissions than its task requires. The rise of agentic systems, in which models take sequences of actions through tools and credentials, makes excessive agency and non-human identity management a growing concern. IBM (2026) reports that fewer than

half of the organizations it studied adequately secured non-human identities within AI workflows.

Adversarial Machine Learning and AI as an Attack Surface

Machine-learning models used for security are themselves targets. Biggio and Roli (2018) traced a decade of research on adversarial machine learning and showed that models can be fooled by carefully crafted inputs at inference time (evasion) or corrupted by manipulated training data (poisoning). The NIST taxonomy of adversarial machine learning (Vassilev et al., 2025) now provides a common vocabulary. It classifies attacks on predictive AI into evasion, poisoning, and privacy attacks, and attacks on generative AI into poisoning, privacy, and misuse attacks, including direct and indirect prompt injection and supply chain attacks. The taxonomy organizes attacks by system type, lifecycle stage, attacker goals, attacker capabilities, and attacker knowledge. For security teams, the practical lesson is that a detection model is part of the attack surface and must be protected, tested, and monitored with the same discipline as any other critical system. An attacker who can influence the logs that feed a detection model, or who can probe its outputs, may be able to degrade or evade it.

Offensive Uses of AI

The same advances that help defenders also assist attackers. Generative models can produce fluent, personalized phishing messages in many languages, assist with reconnaissance, help write or modify malicious code, and support voice and video impersonation. ENISA (2025) reported that AI-supported phishing campaigns reportedly represented more than 80% of observed social engineering activity worldwide by early 2025, while noting that attacks on the AI supply chain were also rising. The 2026 edition characterizes AI, for the time being, as an accelerator of existing techniques rather than a wholly new class of capability, improving speed, scale, reconnaissance, phishing, malware development, and vulnerability exploitation, and it expects more phases of the attack chain to be directly AI-enabled during 2026 (ENISA, 2026). This assessment is useful because it avoids both alarmism and complacency. The principal change is in tempo and cost, which erodes the time defenders have to detect and respond.

2.8 Governance, Standards, and Regulation

Risk-management guidance such as the NIST AI Risk Management Framework emphasizes that AI risks should be governed, mapped, measured, and managed throughout the system lifecycle (NIST, 2023). Cybersecurity frameworks similarly emphasize risk identification, protection, detection, response, and recovery (NIST, 2024b). In December 2025, NIST released a preliminary draft of its Cybersecurity Framework Profile for Artificial Intelligence (the Cyber AI Profile, NIST IR 8596), which applies CSF 2.0 to three overlapping focus areas: securing AI systems, conducting AI-enabled cyber defense, and thwarting AI-enabled cyberattacks (Megas et al., 2025). At the time of writing, the profile is a draft and should be treated as emerging guidance rather than a settled standard. The international management-system standard ISO/IEC 42001 provides a certifiable framework for establishing, operating, and improving an AI management system (International Organization for Standardization & International Electrotechnical Commission, 2023).

Legal and regulatory requirements add further obligations. The General Data Protection Regulation (European Parliament & Council of the European Union, 2016) imposes principles of lawfulness, purpose limitation, data minimization, and security on the processing of personal data, which applies to security monitoring that involves employees or customers. The NIS2 Directive (European Parliament & Council of the European Union,

2022) strengthens cybersecurity risk-management and incident-reporting duties for essential and important entities, a category that includes many digital infrastructure and managed service providers. The EU Artificial Intelligence Act (European Parliament & Council of the European Union, 2024) establishes a risk-based regulatory framework for AI systems. In May 2026 the European Council and Parliament reached a provisional agreement to postpone the application of the high-risk system obligations to December 2027 and August 2028, as part of the Digital Omnibus on AI; organizations should confirm the final adopted dates and the transparency obligations that remain in force. Taken together, these instruments mean that IT organizations should evaluate AI security products before deployment and monitor them continuously after deployment. Procurement decisions should consider vendor transparency, data handling, model updates, incident notification, and the ability to audit system behavior.

Explainability and Human Oversight

Explainability is a practical requirement in security operations, where an analyst must decide whether to act on a model's output. Rudin (2019) argues that for high-stakes decisions, organizations should prefer inherently interpretable models to opaque models accompanied by post hoc explanations, which may not faithfully represent what the underlying model does. Not every security task is high-stakes, but actions that restrict a user's access, trigger a breach notification, or affect a customer's service carry consequences. In these cases, evidence that the analyst can inspect, confidence indicators, and audit trails are more useful than a bare score. Human oversight is meaningful only when analysts have the time, skills, and authority to challenge a recommendation. The ISC2 (2025) data on skills gaps, discussed in Section 4, suggests that this condition cannot be assumed.

2.10 Research Gap and Sector Focus

Existing studies discuss AI applications in cybersecurity across many industries, and industry reports describe breach costs and attack trends in aggregate. A sector-focused examination of IT organizations is useful because IT providers both use AI-enabled security and manage digital infrastructure or services for other organizations. Their operational context includes cloud platforms, software delivery pipelines, identity infrastructure, application programming interfaces (APIs), managed services, and distributed workforces. The central concern is not simply whether AI can improve security, but under what conditions its use is reliable, proportionate, and maintainable. Three gaps motivate this paper. First, scholarly reviews of AI for security (for example, Kaur et al., 2023; Sarker, 2021) predate the 2025 and 2026 evidence on AI-enabled attacks and AI-related breaches. Second, quantitative industry evidence is rarely read alongside the methodological cautions of Sommer and Paxson (2010) and Arp et al. (2022). Third, few sources translate governance guidance into an implementable sequence of steps with measurable indicators for IT providers. This paper addresses these gaps through a focused narrative synthesis supported by secondary quantitative data and a set of implementation recommendations. It does not claim to be a systematic review or to establish causal effects.

Methodology

Research Design

This paper uses a qualitative, narrative literature-review design that is supplemented by a descriptive synthesis of secondary quantitative data. The review draws on established scholarly work on machine learning for cybersecurity and on recognized cybersecurity, AI risk-management, and regulatory guidance. The analysis is organized around four themes: (1) AI applications in cybersecurity, (2) data protection, (3)

implementation barriers, and (4) governance and responsible use. The secondary data are used to give empirical context to these themes. The approach is intended to synthesize concepts and practical implications for the IT sector, not to estimate adoption rates for any specific population or to measure causal relationships.

Sources and Selection Criteria

Sources were selected on the basis of five criteria: (a) authority, meaning that the source is a peer-reviewed publication, a standards body, a government or intergovernmental agency, or a long-running annual industry report with a documented methodology; (b) recency, with a preference for material published between 2023 and 2026 for empirical data and guidance, while retaining earlier foundational scholarship; (c) relevance to AI, cybersecurity, or data protection; (d) verifiability, meaning that the document has a persistent identifier such as a digital object identifier (DOI) or an official publisher address; and (e) transparency about scope and sample. Quantitative figures were extracted only from the primary publisher's own report or official announcement and, where a figure could not be verified in this way, it was not used. Table 1 summarizes the principal empirical sources.

Table 1

Principal Secondary Data Sources Used in the Analysis

Source	Coverage and scale	Use in this paper
IBM (2026) with the Ponemon Institute, Cost of a Data Breach Report	21st annual edition; global study of breach costs, security AI use, and AI-related incidents	Breach cost trends; effect of security AI and automation; AI-related breaches
Verizon (2026), Data Breach Investigations Report	19th edition; more than 31,000 security incidents, over 22,000 confirmed breaches, 145 countries	Initial access vectors; third-party involvement; data leakage to AI platforms
FBI IC3 (2026), Internet Crime Report 2025	1,008,597 complaints to the U.S. Internet Crime Complaint Center in 2025	Scale and cost of cybercrime; first AI-specific section
ENISA (2025, 2026), Threat Landscape	4,875 incidents (July 2024–June 2025) and 8,257 incidents (calendar year 2025) in the EU	AI use by attackers; dependency and supply chain risk; sector context
ISC2 (2025), Cybersecurity Workforce Study	Survey of more than 16,000 cybersecurity professionals worldwide	Skills gaps; AI adoption in security teams
NIST (2023, 2024a, 2024b); Vassilev et al. (2025); Megas et al. (2025)	Standards and guidance documents	Governance framework; adversarial ML taxonomy; Cyber AI Profile draft

Note. Sample sizes and coverage are as stated by the publishers. The reports differ in sampling frame, definitions, and time period and are therefore not directly comparable.

Analytical Procedure

The sources were examined thematically. Concepts relating to detection and

response were grouped under security operations; privacy and data handling were grouped under data protection; and skills, cost, data quality, explainability, and oversight were grouped under implementation and governance. Quantitative indicators were tabulated, compared across editions of the same report where a consistent series exists (for example, the IBM breach-cost series and the FBI IC3 loss series), and interpreted in light of each publisher's methodology. Recommendations were then derived from the convergence of the scholarly literature, the guidance documents, and the empirical indicators. The synthesis is interpretive. It does not apply meta-analytic procedures, because the underlying studies use different populations, measures, and definitions.

Handling of Secondary Statistics

Industry and government statistics are used with explicit caveats, since each has known limitations. The IBM report analyzes costs incurred by organizations that experienced breaches, so it does not describe organizations that were not breached and cannot by itself show that a given tool caused a lower cost. The Verizon report aggregates data from many contributors, and the mix of contributors affects the distribution of incident types. The IC3 figures are voluntary complaints submitted mainly by U.S. residents and businesses, and they do not capture all incidents or losses. ENISA's datasets are drawn from open-source reporting and information supplied by EU Member States, and ENISA itself cautions that open-source reporting provides an incomplete picture. The ISC2 data are self-reported by professionals who volunteered to participate. For these reasons, the statistics in Section 4 are presented as indicators of direction and magnitude. They are not treated as precise estimates, and figures from different reports are not combined into a single index.

3.5 Validity and Reliability Considerations

Several steps were taken to strengthen the validity of the synthesis. Figures were cross-checked against the publisher's own summary page or press material, and the edition and reporting period of each figure are stated. Where an earlier edition of a report is cited alongside a later one, the edition is identified so that readers do not mistake a change in methodology for a change in reality. For example, the ENISA Threat Landscape 2026 covers a different reporting period (calendar year 2025) and a larger incident count than the 2025 edition, and the incident totals of the two editions should not be read as a trend. Claims about the effect of AI on breach cost are reported as associations observed by the publisher. The paper also separates what the sources report from what the author infers, and inferences are flagged as such.

Scope and Limitations

The paper focuses on AI applications relevant to IT companies and IT service providers, including software, cloud, network, and managed-security environments. It does not evaluate a particular vendor, model, or organization. Because the review is narrative rather than systematic, it may not capture every relevant publication. The quantitative evidence is secondary, and the paper does not provide its own estimates of AI adoption or of improvement rates. Such estimates would require a clearly defined sample, validated measures, and transparent data collection. These limitations are discussed further in Section 7.

Empirical Evidence from Recent Reports

This section presents quantitative evidence bearing on RQ2 and RQ3. Subsection 4.1 describes the current threat environment, 4.2 examines breach costs and the observed association with security AI, 4.3 considers AI as both an attacker's tool and a target, 4.4 reports workforce and adoption indicators, and

4.5 synthesizes the findings. All figures are taken from the publications listed in Table 1 and carry the caveats described in Section 3.4.

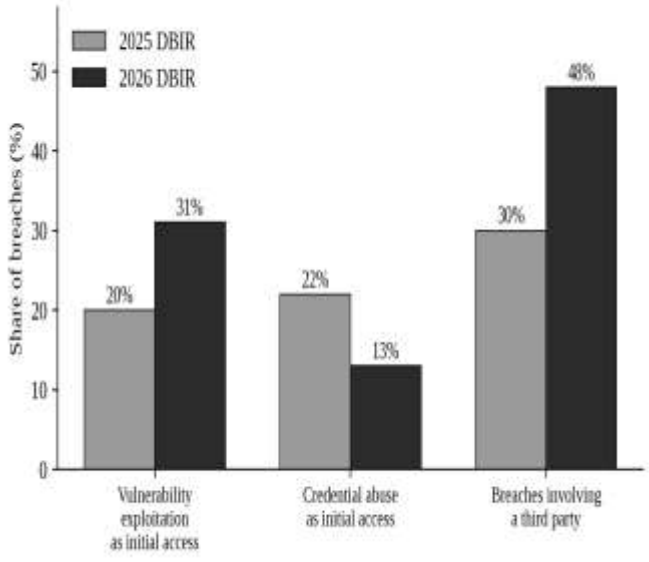
The Threat Environment Facing IT Organizations

Initial access and third-party exposure

The Verizon (2026) Data Breach Investigations Report draws on more than 31,000 security incidents, of which over 22,000 were confirmed data breaches, across 145 countries. Its most notable finding is a change in the leading way in which attackers gain initial access. Exploitation of software vulnerabilities accounted for 31% of breaches, overtaking credential abuse at 13% as the most common initial access vector. In the 2025 edition, credential abuse had stood at 22% and vulnerability exploitation at 20% (Verizon, 2025). The 2026 report also found that 48% of breaches involved a third party, compared with 30% in the previous year. Figure 1 compares the two editions.

Figure 1

Selected Breach Indicators in the 2025 and 2026 Verizon Data Breach Investigations Reports



Note. Data from Verizon (2025, 2026). Percentages refer to the share of breaches in each edition's dataset. Editions differ in contributors and incident samples.

For IT providers, these shifts carry specific implications. Vulnerability exploitation depends on how quickly organizations discover and remediate weaknesses in the software they run and the software they ship. Third-party involvement is directly relevant to providers of cloud, managed, and software services, because such providers are the third parties in their customers' breaches. The Verizon (2026) report also records that employees were uploading source code, technical documents, and other sensitive data to external AI platforms, with source code (28%), structured data (14%), and documents (13%) among the most frequently exfiltrated data types, creating a new channel for leakage of intellectual property and customer data. Taken together, these findings show that AI adoption by employees is itself a data-protection issue, separate from any AI tool the security team deploys.

European evidence on incidents, intrusion paths, and dependencies

ENISA's Threat Landscape 2026, published in September 2026, analyzed 8,257 incidents that affected the European Union during calendar year 2025 (ENISA, 2026). Distributed denial-of-service (DDoS) attacks made up 51.3% of the incidents and unauthorized access 39.5%, and phishing appeared in 77.8% of identified social engineering cases. ENISA characterizes financially motivated activity, ransomware above all, as the most impactful in the short term and identifies cyber dependencies, including supply chains, third-party service providers, and cloud concentration, as an amplifier of impact. The report's illustrative example of a dependency failure is a ransomware attack on an IT supplier in Sweden that reached around 200 municipalities and regional authorities, as summarized in secondary commentary on the report. The case

is a clear illustration of why IT providers carry a sector-specific responsibility.

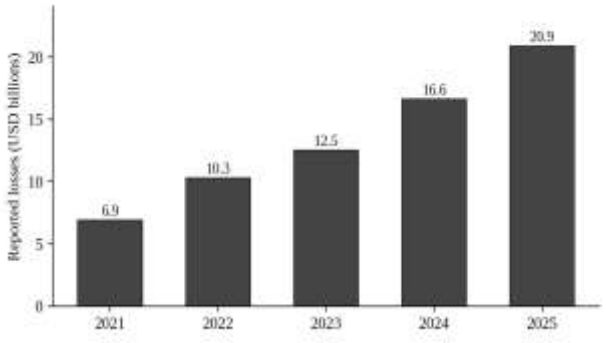
The previous edition covered a different period and used a different method. It analyzed 4,875 incidents between July 2024 and June 2025, found that phishing (about 60%) and vulnerability exploitation (21.3%) were the two leading intrusion access points, and reported that digital infrastructure and services accounted for 4.8% of incidents in the sectoral breakdown (ENISA, 2025). Because the two editions differ in reporting period and incident selection, their absolute totals should not be treated as a trend.

Cybercrime losses

The FBI's Internet Crime Complaint Center received 1,008,597 complaints in 2025, compared with 859,532 in 2024, and total reported losses reached USD 20.877 billion, an increase of 26% over 2024 (FBI, 2026). Cyber-enabled fraud accounted for 452,868 complaints and USD 17.697 billion, or 85% of reported losses. Phishing and spoofing were the most frequently reported crime type, with 191,561 complaints. Reported ransomware losses were comparatively small at USD 32.3 million, but this figure reflects only losses reported to the FBI and is widely regarded as an undercount of the true economic effect. Figure 2 plots the annual loss series for 2021–2025.

Figure 2

Cybercrime Losses Reported to the FBI Internet Crime Complaint Center, 2021–2025



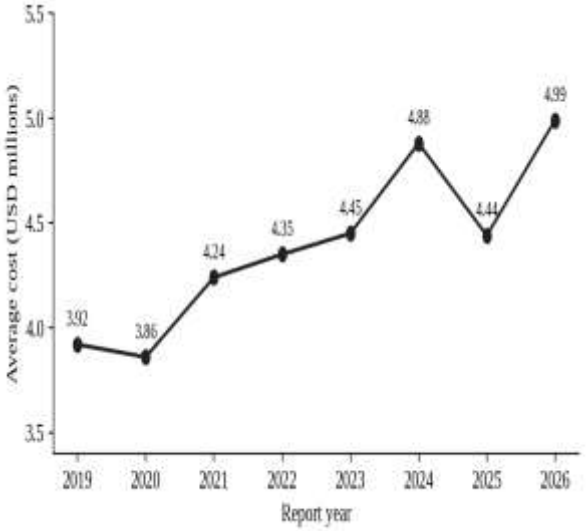
Note. Data from FBI (2026), which presents the historical series. Figures are voluntary complaints, primarily from the United States, and are rounded to one decimal place.

Breach Costs and the Association With Security AI

The IBM Cost of a Data Breach Report 2026, produced with the Ponemon Institute, reports that the global average cost of a data breach reached a record USD 4.99 million, a 12% increase over the previous year (IBM, 2026). In the United States the average reached USD 11.5 million, more than double the global figure and 11% higher than in the prior year. IBM attributes the increase to higher costs for detection, escalation, and lost business. The 2025 edition had reported a global average of USD 4.44 million, a decline from USD 4.88 million in 2024 (IBM, 2025). Figure 3 shows the series from 2019 to 2026. The figure shows that cost is not rising monotonically, which is a reminder that single-year changes reflect the composition of the sample as well as real shifts in risk.

Figure 3

Global Average Cost of a Data Breach, 2019–2026



Note. Data from IBM Security and Ponemon Institute reports (IBM, 2025, 2026). Values are

in millions of U.S. dollars as reported in each year's edition and are not adjusted for inflation.

The report's findings on AI are directly relevant to this paper. Organizations that made extensive use of security AI and automation reported breach costs that were lower by USD 1.93 million on average, and breach lifecycles that were shorter by 65 days, than those of organizations that did not use these capabilities extensively (IBM, 2026). The 2025 edition had reported savings of about USD 1.9 million and a mean breach lifecycle of 241 days across all respondents (IBM, 2025). IBM also reports that only about one-third of organizations used AI and automation extensively across the full security lifecycle, and that adoption is higher for detection and investigation than for prevention. Table 2 summarizes the main indicators.

Table 2
Selected Indicators From the IBM Cost of a Data Breach Report 2026

Indicator	Reported value	Interpretation for IT organizations
Global average cost of a breach	USD 4.99 million (+12%)	Record level; costs driven by detection, escalation, and lost business
U.S. average cost of a breach	USD 11.5 million (+11%)	Providers serving U.S. customers face more than twice the global average
Cost reduction with extensive security AI and automation	USD 1.93 million	Association, not proof of causation; depends on integration and governance
Shorter breach lifecycle with extensive security AI and automation	65 days	Faster identification and containment reduces dwell time
Growth in AI-driven attacks	56% year over year	Attackers scale phishing, impersonation, and malware with AI
Added cost per breach from AI-driven attacks	About USD 1 million	AI-enabled incidents cost about USD 6 million on average
Average cost of model inversion attacks	About USD 6 million	AI models are targets that expose training data
Organizations with an AI-related breach lacking AI access controls	92%	Governance of AI models and data lags adoption
Financial sector average cost	USD 6.29 million (prior year USD 5.56 million)	Sector context for IT providers serving financial clients

Note. Source: IBM (2026) and publisher summaries of the report. Cost savings are differences between groups of respondents and should not be read as the effect of a specific product.

Two cautions are needed in interpreting these numbers. First, the comparison is between groups of organizations that differ in many respects. Organizations that deploy security AI extensively may also have stronger logging,

mature incident-response processes, and larger budgets, and these factors may account for part of the difference. IBM's own analysis lists encryption, identity and access management, DevSecOps, offensive security testing, and governance among the factors associated with lower costs, so AI is one contributor in a set of reinforcing controls (IBM, 2026). Second, the study design (a survey of breached organizations) cannot establish how many organizations that used AI avoided a breach altogether. The figures therefore support a cautious reading: AI and automation are associated with lower cost and faster response when deployed as part of a mature program, but they are not a substitute for it.

AI as an Attacker's Tool and as a Target

The evidence in the 2025 and 2026 reports points in two directions at once. On the offensive side, IBM (2026) reports a 56% increase in AI-driven attacks, led by AI-enabled impersonation and malware, and notes that these incidents add roughly USD 1 million to the average cost of a breach. The FBI (2026) found that AI-related complaints involved nearly USD 893 million in reported losses, and that businesses lost more than USD 30 million to AI-related business email compromise. ENISA (2025) reported the high share of AI-supported phishing noted above, and ENISA (2026) judged that AI is mainly an accelerator today but expects more attack-chain phases to be AI-enabled. Taken together, these sources are consistent in direction. They differ in measurement, and none of them provides a precise count of the share of attacks that use AI.

On the defensive side, AI systems have become targets and sources of exposure. IBM (2026) reports that model inversion attacks, which extract sensitive training data, carried an average breach cost of around USD 6 million, and that 92% of organizations reporting an AI-related breach lacked proper access controls for AI models and data. One analysis of the report also lists prompt injection attacks, with an average cost of USD 5.89 million, and software supply chain or business-partner compromise as the factor that added the most to breach cost across the 30 factors the report examined. These findings reinforce the conclusions of the adversarial machine learning literature (Biggio & Roli, 2018; Vassilev et al., 2025) and show that the concerns raised in the academic literature are now visible in operational incident data.

Workforce and Adoption Indicators

The ISC2 (2025) Cybersecurity Workforce Study surveyed more than 16,000 cybersecurity professionals. Its headline conclusion is that skills, rather than headcount, are now the most pressing workforce concern. Roughly one-third of respondents (33%) said that their organizations lacked the resources to staff their teams adequately, and nearly 30% said that their organizations could not afford to hire staff with the necessary skills. AI was the most frequently cited critical skill need (41%), ahead of cloud security (36%). On adoption, 28% of respondents reported that AI tools were already integrated into daily operations, and 69% were at some stage of evaluating, testing, or implementing them. Respondents were broadly optimistic about the effect on their profession. Seventy-three percent believed AI would create more specialized cybersecurity skills, 72% believed it would require more strategic mindsets, and 66% believed it would require broader skill sets across the workforce.

Table 3 summarizes these indicators.

Table 3

Workforce and AI Adoption Indicators From the ISC2 2025 Cybersecurity Workforce Study

Indicator	Share of respondents (%)
Cited AI as a critical skill need	41
Cited cloud security as a critical skill need	36

Indicator	Share of respondents (%)
Reported insufficient resources to staff teams adequately	33
Reported AI tools already integrated into daily operations	28
Reported evaluating, testing, or implementing AI tools	69
Believed AI will create more specialized cybersecurity skills	73
Believed AI will require more strategic cybersecurity mindsets	72
Believed AI will require broader skill sets across the workforce	66

Note. Source: ISC2 (2025). Based on a survey of more than 16,000 cybersecurity professionals. Responses are self-reported and not from a probability sample.

These indicators matter for the interpretation of the rest of the evidence. If the principal shortage is skills, then AI tools that require specialist configuration, tuning, and oversight may add to the burden of scarce experts unless training keeps pace. The ISC2 data are consistent with the point made by Arp et al. (2022): a model is only as useful as the capacity of the team to evaluate it. The combination of high interest (69% evaluating or implementing) and low full integration (28%) also suggests that most organizations are still in a learning phase, in which governance choices made now will shape how safely AI is used later.

Synthesis of the Evidence

Table 4 relates the empirical findings to the research questions. Four patterns emerge. First, the threat environment is intensifying in scale and cost, and attackers are using AI to increase speed and volume, while the underlying attack techniques (phishing, exploitation of known vulnerabilities, abuse of third-party relationships) remain familiar. Second, security AI and automation are associated with materially lower breach costs and faster response, but the evidence is observational and the benefit is tied to maturity in other controls. Third, AI systems are now an explicit attack surface and a source of data exposure, and governance of AI access controls lags behind adoption. Fourth, the constraint on effective use is increasingly human and organizational, namely skills, evaluation capacity, and accountability, rather than the availability of tools.

Table 4

Relationship Between Empirical Findings and Research Questions

Research question	Principal evidence	Reading of the evidence
RQ2: Benefits of AI	Lower breach cost (USD 1.93 million) and 65 days faster lifecycle with extensive security AI and automation (IBM, 2026)	Plausible and sizable benefit, associated with mature programs; causal effect unproven
RQ3: Challenges and changing threats	56% rise in AI-driven attacks (IBM, 2026); AI-related complaints of nearly USD 893 million (FBI, 2026); 92% lacking AI access controls (IBM, 2026)	AI raises both attacker capability and defender exposure

Research question	Principal evidence	Reading of the evidence
RQ3: Organizational constraints	Skills are the leading concern; 41% cite AI as a critical skill need (ISC2, 2025)	Capacity to oversee AI is a limiting factor
RQ4: Governance	Exploitation overtakes credentials (Verizon, 2026); dependencies as systemic risk (ENISA, 2026)	Governance must cover fundamentals, third parties, and AI together

Note. Synthesis by the author based on the sources cited in the table.

Analysis and Discussion

The literature and the empirical evidence suggest that AI can contribute to several security activities in the IT sector, but the value of a deployment depends on the use case, the available data, the quality of integration, and the governance arrangements. The tables in this section summarize the main applications, data-protection roles, and implementation challenges as a qualitative synthesis. They are not survey results and do not represent measured percentages, except where a figure from a cited report is stated explicitly.

Applications of AI in IT Cybersecurity

Table 5 lists the principal applications, the contribution that each can plausibly make, and the control or limitation that should accompany it. The applications correspond to the areas in which the sources reviewed describe the most established use of machine learning (Buczak & Guven, 2016; Kaur et al., 2023) and in which the incident data show the greatest pressure, such as phishing, vulnerability exploitation, and identity abuse (ENISA, 2026; Verizon, 2026).

Table 5
Applications of AI in IT Cybersecurity

Application	Potential contribution	Important control or limitation
Network and intrusion detection	Identify unusual traffic patterns and potentially malicious activity.	Validate against representative traffic; monitor false positives and missed detections; expect drift (Gama et al., 2014; Sommer & Paxson, 2010).
Endpoint and malware analysis	Classify suspicious files or behaviors and help prioritize investigation.	Test against new variants and evasive behavior; retain endpoint controls and analyst review.
Phishing and social-engineering detection	Analyze message content, links, and sender patterns associated with suspicious communication.	Attackers adapt language with generative tools; provide reporting routes and user awareness training (ENISA, 2025).

Application	Potential contribution	Important control or limitation
Identity and access analytics	Flag unusual login patterns, privilege use, or account behavior, including non-human identities.	Avoid treating anomalies as proof of wrongdoing; apply privacy and access controls.
Vulnerability and exposure management	Prioritize vulnerabilities by likelihood of exploitation and asset context.	Does not replace patching discipline; exploitation is now the leading initial access vector (Verizon, 2026).
Security operations support	Correlate alerts, summarize events, and recommend investigative steps.	Audit outputs for confabulation; restrict automated actions according to potential impact (NIST, 2024a).
Cloud and application security	Help identify configuration anomalies, suspicious API activity, and risks in logs.	Integrate with secure development, vulnerability management, and cloud configuration controls.

Note. Synthesis by the author based on the literature reviewed in Section 2 and the evidence in Section 4.

AI for Data Protection in IT Organizations

Data protection is where the benefits and risks of AI are most closely intertwined. Table 6 summarizes data-protection areas in which AI can help and the implementation considerations that should accompany each.

Table 6
AI for Data Protection in IT Organizations

Data-protection area	Potential AI role	Implementation consideration
Data discovery and classification	Assist in identifying sensitive records and repositories, including source code and credentials.	Verify classification accuracy; limit unnecessary access to content; log what the classifier reads.
Anomaly and exfiltration detection	Highlight unusual downloads, transfers, or access patterns, including uploads to external AI platforms.	Set context-aware thresholds; investigate legitimate business activity before acting (Verizon, 2026).
Identity and access management	Support risk-based review of account activity and privileged access.	Keep access decisions explainable; maintain strong authentication (Rudin, 2019).
Incident triage	Summarize alerts and prioritize possible data exposure.	Confirm evidence before reporting an incident or taking disruptive action.

Data-protection area	Potential AI role	Implementation consideration
Privacy monitoring	Help detect possible policy violations or unexpected data flows.	Define lawful purposes, retention limits, and human oversight (European Parliament & Council of the European Union, 2016).
Protection of AI training and inference data	Track provenance, detect tampering and drift.	Apply data supply-chain controls and access restrictions (National Security Agency Artificial Intelligence Security Center et al., 2025).

Note. Synthesis by the author based on the sources cited in the table.

Key Challenges for AI Adoption

The challenges in Table 7 combine those identified in the literature with the empirical indicators in Section 4. Several are linked. For example, a shortage of skilled personnel (ISC2, 2025) makes it harder to evaluate models rigorously, which in turn weakens the ability to detect the pitfalls described by Arp et al. (2022).

Table 7

Key Challenges for AI Adoption in the IT Sector

Challenge	Why it matters in the IT sector	Recommended response
Shortage of skilled personnel	Skills gaps are now the leading workforce concern; AI is the most cited skill need (ISC2, 2025).	Provide role-based training and combine AI outputs with analyst expertise.
Data quality and coverage	Incomplete or inconsistent logs can weaken model performance and bias evaluation.	Improve logging, data governance, and validation using realistic scenarios.
Cost and integration complexity	Tools may require new infrastructure, licenses, engineering, and ongoing monitoring.	Start with a bounded use case and assess total cost and operational benefit.
Privacy and confidentiality	Security telemetry can contain personal, customer, or proprietary information; models can leak training data (Carlini et al., 2021).	Minimize data, restrict access, define retention, and assess vendor data practices.

Challenge	Why it matters in the IT sector	Recommended response
Explainability and trust	Analysts may be unable to assess why a model flagged an event.	Require useful evidence, confidence indicators, audit trails, and escalation paths.
Adversarial adaptation	Attackers may evade, manipulate, or exploit AI-enabled controls (Vassilev et al., 2025).	Red-team systems, monitor drift, update models, and maintain layered defenses.
Governance lag	92% of organizations with an AI-related breach lacked proper AI access controls (IBM, 2026).	Inventory AI systems, assign owners, and apply identity and access controls to models and agents.
Employee use of external AI tools	Source code and documents are uploaded to external AI platforms (Verizon, 2026).	Provide approved tools, set acceptable-use policy, and monitor data flows.

Note. Synthesis by the author based on the sources cited in the table.

Sector-Specific Implications for IT Providers

The preceding analysis applies broadly, but IT providers have five characteristics that change how it should be read.

The first is the dual role of user and provider. An IT provider uses AI for its own defense and also operates infrastructure or services on which customers rely. A compromise at the provider can therefore cascade, as ENISA (2026) illustrates through its example of an IT supplier whose ransomware incident disrupted roughly 200 public bodies. Verizon's finding that 48% of breaches involved a third party (Verizon, 2026) is the other side of the same relationship. For the provider, AI-enabled monitoring of its own environment is a means of protecting customers, and the provider's detection and notification capabilities affect its customers' regulatory obligations under instruments such as NIS2 (European Parliament & Council of the European Union, 2022).

The second is the software supply chain. IT organizations write, assemble, and distribute software, and the vulnerability exploitation trend in the Verizon (2026) data places weight on secure development and rapid remediation. AI can help prioritize vulnerabilities and review code, but it can also introduce risk if developers rely on generated code without review or paste proprietary code into external tools. IBM's analysis associates DevSecOps adoption with lower breach costs (IBM, 2026), which supports integrating AI assistance into, rather than around, secure development practice.

The third is identity infrastructure. IT providers manage large numbers of privileged accounts, service accounts, API keys, and, increasingly, non-human identities used by automation and AI agents. IBM (2026) reports that fewer than half of the organizations studied adequately secured non-human identities within AI workflows. Identity analytics is a strong use case for AI, but it requires accurate inventories of identities and their entitlements, which many organizations lack.

The fourth is the distributed and multi-tenant nature of operations. Managed security and cloud providers often analyze telemetry from many customers. This creates scale advantages, since a model can learn from a wide variety of environments, but it also raises

confidentiality questions when data from one customer influences models that serve another. Contractual terms, data segregation, and clear rules about the use of customer telemetry for model training are therefore essential.

The fifth is regulatory exposure across jurisdictions. A provider with customers in several regions may be subject to overlapping requirements on data protection, incident reporting, and AI use. The evolving timetable of the EU AI Act, including the provisional agreement in May 2026 to postpone certain high-risk obligations, illustrates the need for a compliance process that tracks change rather than one-time certification.

A Practical Framework for Responsible Adoption

A responsible adoption plan should connect AI tools to the organization's existing security and risk-management program. The following staged framework is proposed as a practical synthesis of the literature and guidance reviewed in this paper. It is not a validated measurement model. It is consistent in structure with the Govern, Map, Measure, and Manage functions of the NIST AI RMF (NIST, 2023) and with the focus areas of the draft Cyber AI Profile (Megas et al., 2025).

Define the security problem. Identify a specific use case, the assets at risk, the intended users, and the consequences of errors. A well-defined problem allows evaluation against a baseline.

Assess data readiness. Establish whether telemetry is relevant, sufficiently complete, lawfully processed, and protected against unauthorized access, and document the provenance of training and reference data.

Pilot in a controlled environment. Test with representative normal activity and simulated or historical incidents before enabling consequential automation, using time-based evaluation splits and realistic base rates to avoid the pitfalls described by Arp et al. (2022).

Measure operational performance. Evaluate detection quality, false-positive burden, investigation time, analyst workload, and unintended service disruption (see Table 9).

Set human oversight and response rules. Define which actions may be automated, which require approval, and how decisions can be reversed. Limit the permissions of AI agents to those required for the task, in line with the concern about excessive agency (OWASP, 2025).

Monitor and improve. Review performance drift, emerging attack techniques, access logs, vendor changes, and feedback from security analysts. Test the AI system itself against the attack classes in the NIST taxonomy (Vassilev et al., 2025).

Document accountability. Maintain records of system purpose, limitations, data use, approvals, incidents, and periodic review, and assign a named owner for each AI system.

Mapping Governance Instruments to Adoption Stages

Organizations often face several overlapping instruments. Table 8 maps the principal ones to the stages of the framework to show where each is most useful. The mapping is the author's interpretation and is intended as a planning aid, not as a compliance determination.

Table 8

Mapping of Governance Instruments to the Adoption Framework

Instrument	Principal focus	Most relevant stages
NIST AI RMF 1.0 (NIST, 2023)	Governing, mapping, measuring, and managing AI risk	All stages; especially 1, 4, 7

Instrument	Principal focus	Most relevant stages
NIST CSF 2.0 (NIST, 2024b)	Cybersecurity outcomes (Govern, Identify, Protect, Detect, Respond, Recover)	1, 5, 6
NIST AI 600-1 Generative AI Profile (NIST, 2024a)	Risks unique to generative AI, such as confabulation	2, 3, 5
NIST AI 100-2 E2025 (Vassilev et al., 2025)	Taxonomy of attacks on predictive and generative AI	3, 6
NIST IR 8596 draft Cyber AI Profile (Megas et al., 2025)	Securing AI, AI-enabled defense, thwarting AI-enabled attacks	1, 4, 6 (draft status)
AI Data Security guidance (NSA AISC et al., 2025)	Data supply chain, data poisoning, drift	2, 6
OWASP Top 10 for LLM Applications (OWASP, 2025)	Application-level risks for LLM systems	3, 5, 6
ISO/IEC 42001:2023	AI management system	1, 7
GDPR; NIS2; EU AI Act	Data protection, incident and risk duties, AI regulation	2, 5, 7

Note. Stage numbers refer to the seven stages in Section 5.5. Mapping by the author.

Measuring Success: Indicators for IT Organizations

The IBM (2026) findings suggest that the benefit of AI can be large, but, as noted in Section 4.2, that benefit has been observed across groups of organizations and has not been demonstrated for any particular deployment. Each organization should therefore measure outcomes in its own environment. Table 9 proposes indicators that correspond to the framework. The indicators are illustrative, and organizations should set thresholds and baselines that match their risk appetite and operating model.

Table 9

Suggested Indicators for Evaluating AI-Enabled Security Use Cases

Dimension	Example indicator	Why it matters
Detection quality	Precision and recall on labeled incidents, measured with realistic class imbalance	Avoids overstating accuracy (Arp et al., 2022)
Analyst workload	Alerts per analyst per shift; share of alerts auto-closed and later reopened	Captures the burden of false positives
Speed	Mean time to detect, to contain, and to recover compared with the pre-deployment baseline	Aligns with IBM's lifecycle metric (IBM, 2026)

Dimension	Example indicator	Why it matters
Safety of automation	Number of automated actions reversed or causing service disruption	Controls the cost of false positives
Model health	Drift indicators; frequency of retraining; performance by data source	Addresses concept drift (Gama et al., 2014)
Data protection	Volume of personal data processed by the tool; retention compliance; access reviews	Supports data minimization and lawful processing
AI system security	Findings from red-team tests against the model; inventory coverage of AI systems and agent identities	Responds to the governance lag reported by IBM (2026)
Skills	Share of analysts trained on the tool; time to onboard new analysts	Addresses the skills gap (ISC2, 2025)

Note. Indicators proposed by the author. They are not drawn from a validated instrument.

Discussion

AI can strengthen IT cybersecurity by helping teams process large amounts of telemetry and prioritize events that may otherwise be overlooked. The most useful role of AI is likely to be augmenting the work of security analysts: models can identify patterns and produce recommendations, while trained personnel assess context, confirm evidence, and decide how to respond. This combination can improve the consistency of workflows without assuming that automated decisions are always correct. The IBM (2026) association between extensive security AI use and lower breach cost is consistent with this view, but the same report's observation that fewer than half of organizations adequately secure non-human identities within AI workflows is a reminder that the benefit depends on how the tools are run.

Data protection is both a potential benefit and a source of risk. AI may identify suspicious data movement, but it may also require broad access to sensitive logs or content, and the models and the data they use may themselves be targeted (Carlini et al., 2021; Shokri et al., 2017). Organizations should therefore assess security and privacy together. Data minimization, role-based access, retention limits, secure model configuration, and careful vendor review are not optional additions. They are part of a trustworthy deployment. The 2026 Verizon finding on the upload of source code and documents to external AI platforms shows that data protection must also cover the way employees use AI, in addition to the way the security team uses it.

The IT sector also faces a dual-use problem. The same general advances that help defenders analyze threats can assist malicious actors in creating convincing phishing content, impersonating executives, or scaling reconnaissance. The FBI (2026) and ENISA (2025, 2026) evidence is consistent with this concern, although the data do not allow a precise estimate of how much of current attack activity depends on AI. ENISA's measured judgment that AI is so far an accelerator of known techniques has a useful implication. Defenders gain the most from improving the speed and reliability of fundamentals, namely asset visibility, patching, multifactor authentication, identity protection, backups, and tested incident response, because AI-enabled attackers shorten the time between a

vulnerability becoming known and being exploited. Layered security controls, secure software development, and incident-response readiness therefore remain necessary, and organizations should avoid dependence on a single AI product.

The analysis also supports a measured adoption strategy. Rather than deploying AI broadly because it is new, IT organizations should select high-value, bounded use cases and evaluate them against operational baselines. Claims of improved detection or reduced response time should be substantiated with documented evidence from the organization's own environment. The industry-wide association reported by IBM justifies investment in the capability but not a reduction in scrutiny. Without local measurement, benefits remain plausible but unverified.

5.9 Implications for Theory and Practice

For theory, the evidence supports treating AI in security as a socio-technical system in which model performance, data governance, human capability, and adversary behavior interact. Technical research on detection accuracy (Buczak & Guven, 2016; Kaur et al., 2023) is necessary but insufficient. Studies of operational outcomes should account for organizational maturity as a confounding and moderating factor, and for the adaptive behavior of attackers, which can erode the performance of a deployed model over time.

For practice, three implications stand out. First, leaders should budget for the complements of AI, including logging, identity hygiene, training, and evaluation, and not only for licenses. Second, procurement should require evidence of performance in comparable environments, transparency about data use, and the ability to audit model behavior. Third, boards and executives should treat AI security as part of enterprise risk management and assign clear accountability, consistent with the Govern function of CSF 2.0 (NIST, 2024b).

Recommendations

On the basis of the literature and the evidence reviewed, the following recommendations are offered to IT organizations. They are grouped by the audience that is best placed to act on them.

For Security and IT Leaders

Begin with bounded, high-value use cases, such as phishing triage, alert correlation, or identity anomaly detection, and evaluate them against a documented baseline before expanding.

Prioritize the fundamentals that current data identify as the main routes of compromise: rapid vulnerability remediation, strong authentication, third-party risk management, and tested incident response (ENISA, 2026; Verizon, 2026).

Create and maintain an inventory of AI systems, their owners, their data sources, and the identities and permissions that they use, including AI agents and other non-human identities.

Invest in training so that analysts can evaluate, tune, and challenge AI outputs, and allocate working time for it (ISC2, 2025).

For Data Protection and Privacy Teams

Conduct a data protection impact assessment before AI tools process personal telemetry, and define purpose limits, retention periods, and access rules.

Set and enforce policy on the use of external AI platforms by employees, provide approved alternatives, and monitor for uploads of source code and sensitive documents (Verizon, 2026).

Apply data supply-chain controls to training and reference data, including provenance tracking and integrity checks (National Security Agency Artificial Intelligence Security Center et al., 2025).

For Procurement and Vendor Management

Require vendors to disclose how customer data are used, whether they are used to train shared

models, how models are updated, and how incidents are notified.
Ask for evidence of testing against adversarial attack classes and for the ability to audit decisions and export logs.
Avoid single-vendor dependence for critical detection functions and plan for degraded operation if an AI service is unavailable.

For Governance and Executive Oversight
Align AI security governance with an established framework, such as the NIST AI RMF, CSF 2.0, and ISO/IEC 42001, and monitor the status of draft guidance such as the Cyber AI Profile.
Track regulatory change, including the final adopted timetable of the EU AI Act, and the obligations under data protection and NIS2 regimes that apply to the organization.
Require periodic reporting to senior management on indicators such as those in Table 9, including incidents involving AI systems.

Limitations and Directions for Future Research

Limitations

The study has several limitations that readers should bear in mind. First, it is a narrative review, so the selection of sources reflects the author's judgment against the stated criteria and may omit relevant work. Second, the quantitative evidence is secondary and derives from industry and government reports whose methods differ. These reports are subject to selection effects, such as reliance on breached organizations (IBM), voluntary complaints (FBI), open-source incident reporting (ENISA), and self-selected survey respondents (ISC2). Third, the reports were produced for different regions and periods, so their figures should not be pooled. Fourth, some figures, including the ENISA example of an IT supplier incident and the analysis of prompt-injection costs, were drawn from commentary on the primary reports and should be confirmed against the full publications before being relied upon for decisions. Fifth, the field is changing quickly. The Cyber AI Profile is still a draft, the EU AI Act timetable is being amended, and the capabilities of AI models and agents are advancing, so some conclusions may require revision. Finally, the proposed framework and indicators are conceptual syntheses and have not been empirically validated.

Directions for Future Research

The limitations point to several research opportunities. Survey research could examine IT-sector professionals using a transparent sampling design to estimate adoption, use cases, and governance practice, and could compare organizations of different sizes and in different regions, including developing economies in which the cybersecurity skills gap may be more acute. Case studies and field experiments could evaluate specific AI use cases using the indicators in Table 9, comparing outcomes with and without the tool while controlling for organizational maturity. Longitudinal studies could examine whether improvements in detection quality, response time, analyst workload, and data-protection outcomes persist as systems and threats change. Research on the security of AI-enabled security tools, including prompt injection in analyst assistants and manipulation of the logs that feed detection models, deserves particular attention. Finally, studies of human factors, including trust calibration, automation bias, and the effect of AI assistance on analyst skill development, would clarify the conditions under which human oversight is effective.

Conclusion

This paper examined the role of AI in cybersecurity and data protection in the IT sector through a narrative review that is supported by recent secondary data. AI can support network and endpoint monitoring, phishing detection, identity analytics, vulnerability prioritization, cloud security,

incident triage, and the identification of unusual data activity. The evidence reviewed indicates that these capabilities may be valuable: IBM (2026) reports that organizations using security AI and automation extensively had breach costs USD 1.93 million lower and breach lifecycles 65 days shorter than those that did not. At the same time, the threat environment is intensifying. Global average breach costs reached a record USD 4.99 million, cybercrime losses reported to the FBI exceeded USD 20 billion, vulnerability exploitation overtook credential abuse as the leading initial access vector, and third parties were involved in nearly half of breaches (FBI, 2026; IBM, 2026; Verizon, 2026).

Nevertheless, AI does not eliminate cybersecurity risk. Model errors, data-quality problems, privacy concerns, adversarial manipulation, integration costs, limited explainability, and shortages of skilled staff can reduce the value of AI-enabled tools. AI systems are themselves targets, and governance of access to AI models and data lags adoption. In some situations, poorly governed automation can create additional operational risk. IT organizations should consequently treat AI as one component of a layered cybersecurity strategy, supported by human oversight, secure data practices, continuous testing, clear accountability, and attention to the fundamentals that attackers continue to exploit.

The paper's main contribution is a sector-focused synthesis that links recent evidence to the scholarly literature and to current governance guidance, together with a staged adoption framework, a governance mapping, and a set of measurable indicators. It does not present new empirical measurement, and its framework has not been validated. Future research should survey IT-sector professionals using a transparent sampling design, evaluate specific AI use cases with clearly defined performance indicators, and follow organizations over time. Such research would provide stronger evidence about where AI produces measurable benefits and under what conditions those benefits justify the associated costs and risks.

References

Ahmad, G., Shaukat, J., & Bhatti, M. F. T. (2026). Strengthening pakistan's cybersecurity through cloud-based artificial intelligence and zero trust architecture: challenges, opportunities, and a strategic framework. *Journal Of Social Sciences Development*, *5*(1), 392-412.

Arp, D., Quiring, E., Pendlebury, F., Warnecke, A., Pierazzi, F., Wressnegger, C., Cavallaro, L., & Rieck, K. (2022). Dos and don'ts of machine learning in computer security. In *Proceedings of the 31st USENIX Security Symposium* (pp. 3971–3988). USENIX Association. https://www.usenix.org/conference/usenixsecurity22/presentation/arp

Arshad, K., Shaukat, J., & Bhatti, M. F. T. (2026). Cybersecurity and Public Trust in Autonomous Vehicles: Challenges, Mitigation Strategies, and Future Directions. *The Critical Review of Social Sciences Studies*, *4*(1), 8399-8414.

Bhatti, M. A. A., & Shaukat, J. (2025). Impact of Cybersecurity Risk Management Practices on Agile Project Performance: Mediating Role of Team Trust and Moderating Role of AI Adoption. *Research Consortium Archive*, *3*(2), 1308-1335.

Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition, 84*, 317–331. https://doi.org/10.1016/j.patcog.2018.07.023

Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications*

Surveys & Tutorials, 18(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

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). USENIX Association. https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting

European Parliament & Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). *Official Journal of the European Union, L 119*, 1–88. https://eur-lex.europa.eu/eli/reg/2016/679/oj

European Parliament & Council of the European Union. (2022). Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union (NIS 2 Directive). *Official Journal of the European Union, L 333*, 80–152. https://eur-lex.europa.eu/eli/dir/2022/2555/oj

European Parliament & Council of the European Union. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union, L, 2024/1689*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

European Union Agency for Cybersecurity. (2025). *ENISA threat landscape 2025*. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2025

European Union Agency for Cybersecurity. (2026). *ENISA threat landscape 2026*. https://www.enisa.europa.eu/publications/enisa-threat-landscape-2026

Federal Bureau of Investigation, Internet Crime Complaint Center. (2026). *Internet crime report 2025*. https://www.ic3.gov/AnnualReport/Reports/2025_IC3Report.pdf

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., & Bouchachia, A. (2014). A survey on concept drift adaptation. *ACM Computing Surveys, 46*(4), Article 44. https://doi.org/10.1145/2523813

Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., & Fritz, M. (2023). Not what you've signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security* (pp. 79–90). Association for Computing Machinery. https://doi.org/10.1145/3605764.3623985

IBM. (2025). *Cost of a data breach report 2025*. https://www.ibm.com/reports/data-breach

IBM. (2026). *Cost of a data breach report 2026*. https://www.ibm.com/reports/data-breach

International Organization for Standardization & International Electrotechnical Commission. (2023). *Information technology—Artificial intelligence—Management system* (ISO/IEC Standard No. 42001:2023). https://www.iso.org/standard/81230.html

ISC2. (2025). *2025 ISC2 cybersecurity workforce study*. https://www.isc2.org/research

Kaur, R., Gabrijelčič, D., & Klobučar, T. (2023). Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion, 97*, Article 101804. https://doi.org/10.1016/j.inffus.2023.101804

Megas, K., Cuthill, B., Dotter, M., Garris, M., Khemani, I., Patrick, B., Schiro, N., Snyder, J. N., & Zarei, M. (2025). *Cybersecurity framework profile for artificial intelligence (Cyber AI Profile): NIST community profile* (NIST IR 8596 initial preliminary draft). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8596.ipd

National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)* (NIST AI 100-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.100-1

National Institute of Standards and Technology. (2024a). *Generative artificial intelligence profile* (NIST AI 600-1). U.S. Department of Commerce. https://doi.org/10.6028/NIST.AI.600-1

National Institute of Standards and Technology. (2024b). *The NIST cybersecurity framework (CSF) 2.0* (NIST CSWP 29). U.S. Department of Commerce. https://doi.org/10.6028/NIST.CSWP.29

National Security Agency Artificial Intelligence Security Center, Cybersecurity and Infrastructure Security Agency, Federal Bureau of Investigation, Australian Cyber Security Centre, New Zealand National Cyber Security Centre, & United Kingdom National Cyber Security Centre. (2025). *AI data security: Best practices for securing data used to train & operate AI systems* [Cybersecurity information sheet]. https://www.cisa.gov/news-events/alerts/2025/05/22/new-best-practices-guide-securing-ai-data-released

OWASP GenAI Security Project. (2025). *OWASP top 10 for LLM applications 2025*. https://genai.owasp.org/llm-top-10/

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. https://doi.org/10.1038/s42256-019-0048-x

Sarker, I. H. (2021). AI-driven cybersecurity: An overview, security intelligence modeling and research directions. *SN Computer Science, 2*, Article 173. https://doi.org/10.1007/s42979-021-00557-0

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. https://doi.org/10.1109/SP.2017.41

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In *2010 IEEE Symposium on Security and Privacy* (pp. 305–316). IEEE. https://doi.org/10.1109/SP.2010.25

Umair, M., Shaukat, J., & Bhatti, M. F. T. (2026). Artificial Intelligence in Cybersecurity: Threat Detection, Defense, and the Governance of a Dual-Use Technology. *Research Consortium Archive*, *4*(2), 4951-4969.

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). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-2e2025

Verizon. (2025). *2025 data breach investigations report*. https://www.verizon.com/business/resources/reports/dbir/

Verizon. (2026). *2026 data breach investigations report*. https://www.verizon.com/business/resources/reports/dbir/