© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1720158

Cybersecurity and Privacy Protection in AI-Driven Digital Environments

DARSHANKUMAR JAYSUKH DHANANI
D-Tech Studios LLC

Abstract- ***The rise of artificial intelligence (AI) has revolutionized the nature of cybersecurity, both in terms of its attack and defense capabilities, and also due to the new privacy threats that it poses. This paper reviews the literature systematically, and investigates the usage of AI technologies in the context of enhancing cyber-defence and safeguarding data privacy in AI-driven digital environments as well as their adverse usage and destabilisation by adversarial actors such as machine learning (ML), deep learning (DL), federated learning, and explainable AI (XAI). A total of 21 peer-reviewed and indexed sources have been analysed and thematically categorised into four areas: AI-based threat and intrusion detection, adversarial machine learning and model robustness, privacy-preserving AI techniques, and governance/regulatory compliance. The review reveals that, compared to signature-based approaches, AI can significantly improve accuracy of anomaly and intrusion detection, that federated learning and differential privacy provide mathematically rigorously, but utility limited, privacy assurance, and that adversarial perturbations continue to be a major vulnerability which can compromise even very accurate detection models. Escalating demands for trustworthy deployment of AI systems have come from regulatory frameworks like the General Data Protection Regulation (GDPR) and the European Union Artificial Intelligence Act, among other sources. The paper summarizes the elements of resilient AI-driven digital ecosystems and proposes a conceptual framework and research agenda for future research in this fast-moving space.***

Keywords: ***Artificial Intelligence, Cybersecurity, Data Privacy, Intrusion Detection, Adversarial Machine Learning, Federated Learning, Differential Privacy, Explainable AI, GDPR, AI Governance***

I. INTRODUCTION

Digital environments are changing their structure to become more and more permeated by the extension of artificial intelligence into networks, cloud platforms, the Internet of Things (IoT) and even the commonplace use of these applications. AI is integral to basic cybersecurity operations, such as intrusion detection, malware identification, fraud analysis, and automatic incident response, and is now dealing with more and more personal and behavioural data as well as increasing and unprecedented cyber security incidents, leading to high levels of privacy exposure (Achuthan et al., 2024; Chukwunweike et al., 2024). The double-edged sword aspect of AI (the shield and the sword) is a key element in today's cybersecurity studies.

Machine learning and deep learning models can identify suspicious network traffic which is difficult to detect with static and signature-based systems providing adaptive protection against zero-day and polymorphic attacks (Ahmad et al., 2021; Liu & Lang, 2019; Pinto et al., 2023). The same class of algorithms can be used by attackers: Generative models can be used to automate spear phishing, adversarial perturbation can be used to make the model harder to detect, and adversarial models can be used to speed up the reconnaissance process for an adversary, which is unaffordable for human attackers (Alotaibi & Rassam, 2023; Anthi et al., 2021; Mohammadiounotikandi & Babaeitarkami, 2024). The data-intensive nature of AI training pipelines, however, challenges essential data-protection concepts and grounds set out by instruments like the GDPR and the EU AI Act (Goncalves & Correia, 2026), including purpose limitation, data minimisation and the right to explanation.

Federated learning and differential privacy are among the key privacy-preserving computation techniques that have emerged as partial technical solutions, allowing the training of models without the need to centralize the raw data from various sources (Lim & Lee, 2025; Ranaweera et al., 2025; Shen et al., 2023). However, they come inherent with their own set of compromises in statistical utility versus formal privacy assurances and fail to address governance

issues of accountability, transparency, or human oversight (Chowdhary, 2025; Zhang et al., 2022).

In this context, this paper aims at answering the following three questions:
(1) How is AI currently applied to strengthen cybersecurity detection and response in digital environments?
(2) What privacy-preservation techniques are used to reconcile AI's data requirements with individual privacy rights, and what are their limitations?
(3) What adversarial and regulatory challenges constrain the trustworthy deployment of AI-driven security systems, and what governance architecture can address them?

This line of inquiry is by no means solely theoretical. The constant increase of the volume, velocity, and variety of cyber threats have been observed through industry incident reporting as an area where security analysts' human ability to perform manual triage has been overwhelmed, and where AI-driven detection would help address the challenge (Achuthan et al., 2024). Meanwhile, as regulators increasingly consider algorithmic opacity to be no longer an acceptable operating condition for organisations using algorithms to process personal data on a large scale, algorithmic transparency becomes a pressing compliance risk in the GDPR and the forthcoming EU AI Act compliance obligations, respectively (Goncalves & Correia, 2026). The challenge for organisations, then, is to make sure they are using AI in an aggressive enough way to stay one step ahead of automated, AI-powered attacks, and to make sure that their own AI systems are private, explainable and legally defendable. This conflict of urgency and responsibility is evident in several of the texts cited in Section 2 and is the basis for the conceptual merging that is suggested in Section 5.

The key categories of AI applications for cybersecurity and privacy preservation that are discussed in this review are summarised in Table 1, along with the primary function of the AI applications and exemplary studies.

Table 1. Key AI-Enabled Cybersecurity and Privacy-Preservation Techniques

Technique / Concept	Primary Function
Machine-learning / deep-learning Intrusion Detection Systems (IDS)	Classify network traffic as benign or malicious using learned patterns rather than fixed signatures
Adversarial machine learning	Crafting or defending against perturbed inputs designed to fool ML/DL classifiers
Federated learning	Trains shared models across distributed devices without centralising raw data
Differential privacy	Adds calibrated statistical noise to bound the privacy leakage of individual records
Explainable AI (XAI) & regulatory compliance	Produces human-interpretable rationales for automated decisions to support GDPR/AI Act accountability

II. LITERATURE REVIEW

2.1 AI and Machine Learning for Threat Detection and Intrusion Prevention

A large number of papers have been written on the evolution of rule based intrusion detection systems (IDS) to learning based architectures. The taxonomy of Liu and Lang (2019) is an early and popular classification of ML and DL based IDS by the type of data object analyzed (host-based, network-based and hybrid) and indicates that deep architectures like recurrent neural networks, and convolutional neural networks, tend to be superior in high-dimensional traffic data, but with the downside of lack of interpretability and high computation costs. Ahmad et al. (2021) add a systematic study of the differences between classical ML algorithms (such as support vector machines, random forests, k-nearest

neighbours) and deep learning algorithms and show that, in some benchmark datasets, combinations of these classifiers are often the best at achieving a tradeoff between detection accuracy and false alarm rate.

Pinto et al. (2023) focus on industrial control systems and critical infrastructures, and recommend that the convergence of operational technology (OT) with IT networks brought about by Industry 4.0 has increased the attack surface that has to be faced by supervisory control and data acquisition (SCADA) systems, which is why ML-based IDS are becoming indispensable, as attackers are now routinely evading traditional perimeter defences. Similarly, Heidari and Jamali (2023) discuss IDS for resource-constrained IoT devices, noting that lightweight ML models need to sacrifice the level of granularity of the detection in exchange for the low latency and energy budgets that are typical of edge deployments. Figueiredo et al. (2023) show that moving a deep-learning architecture from one type of network configuration to another is not an easy task, as models optimised for one traffic distribution tend to have low external validity for another; this is true for the external validity of lab-reported accuracy numbers.

One thing that is common to many of these studies is the reliance on performance metrics based on the quality, timeliness and representativeness of the training data. It is generally accepted that older datasets have been generated by attacks that are not well representative of the attacks of today, as this is the point in which more recent, specifically created releases of benchmarks, such as the CICIoT2023 (Neto et al., 2023), aim to fill the gap captured by large-scale, real-time attack traffic from the Internet of Things. The importance of benchmark realism directly in affecting the external validity of comparative claims between algorithms comes up repeatedly throughout the detection literature, and was a driving factor for the cautionary note of Figueiredo et al. (2023) on model transposition across datasets: a classifier that seems to perform very well on an outdated and well-studied dataset may perform abysmally on novel attack traffic. Another recurring theme in multiple studies is the cost of false positives: with traffic volumes common in many enterprise networks or critical infrastructure sites, a high false-positive rate can eventually become too much and lead to an “alert fatigue” that can actually become a security risk as analysts become immune to and so disregard automatically generated alerts.

Together, this body of research confirms that AI detection has become the dominant paradigm in network defence research, but also that the results of any AI fail to be comparable across studies without the use of some form of benchmarking. In particular, the results are highly dependent on the characteristics of the datasets used, and it is imperative to use benchmarking datasets like those released by Neto et al. (2023) in their CICIoT2023 release to make the results of any AI comparable across studies.

2.2. Adversarial machine learning and on model robustness

Another, related line of inquiry is the security of the AI models. The original paper by Goodfellow et al. (2015) showed that deliberately engineered perturbations to the input of an image classification network can result in high-confidence misclassifications, introducing the infamous fast gradient sign method (FGSM) to the arena of adversarial attacks and spawning over a decade of related attacks and defenses. On this basis, He et al. (2023) conduct a thorough review of adversarial machine learning in network intrusion detection, classify adversarial attacks into three categories (evasion, poisoning, and model-extraction) and demonstrate its susceptibility to evasion attacks, especially when the IDS is based on a static dataset and an attacker can iterate over its decision boundaries.

Alotaibi and Rassam (2023) also perform a survey of adversarial strategies against IDS, again separating out attacks that occur during training time (data poisoning) from those that occur during inference time (evasion) and summarizing the various strategies and methods proposed to defend against such attacks such as adversarial training, defensive distillation, and ensemble diversification, concluding that no single defence is fully robust across all types of attacks, and that defending in layers or defence-in-depth is the most defensible posture. The research by Jmila and Khedher (2022) is an empirical comparative study of adversarial robustness for different ML algorithms for

network intrusion detection; they argue that meaningful differences exist between model families, which makes algorithm choice a security-related design decision, rather than just one that is performance related. In operational technology (OT) spaces, as in IT spaces, adversarial attacks designed to compromise ML-based cybersecurity defenses can also lead to a degradation of a system's detection capability, potentially enough to cause an OT system to miss an attack, which has potentially serious safety consequences for critical infrastructure, as shown in Anthi et al. (2021).

The essence of this literature comes to the same conclusion: under ideal testing conditions, detection accuracy is significantly higher than the accuracy that would be observed in the field when AI security systems are faced with adaptive adversaries. This further suggests that adversarial evaluation should be a core part of AI security systems assurance.

2.3 Federated learning and differential privacy are two approaches to privacy-preserving AI

A third stream discusses the privacy aspects of AI systems that have to acquire knowledge from personal data, which is highly sensitive and regulated. Federated learning (FL), a primary architecture that enables the sharing of model updates instead of data between model clients and an aggregator, is the key solution that has emerged. Based on the FL-adjacent architectural principles, Lim and Lee (2025) propose a comprehensive logging system that integrates FL and the Analytic Hierarchy Process to prioritize security scenarios of AI transformation environments and show that the integrated log system through holistic and cross-system logging can significantly outperform single-system logging in detecting insider data-leakage attempts.

Although this is the case, a number of studies warn that FL does not necessarily provide privacy. While private information is still possible to be inferred from the model parameters uploaded during federated training, Ranaweera et al. (2025) argue that privately adding calibrated statistical noise to local updates is a useful approach to overcome the utility loss of standard DP mechanisms and propose a noise-injection scheme based on the wavelet transform to achieve this. Similarly, the authors of Shen et al. (2023) tackle the very related concern of privacy heterogeneity, where a single uniform privacy budget is not well suited to the realities of federations where the data held by each client may be very sensitive; their personalised local differential privacy (PLDP-FL) framework lets each client define their own privacy-utility trade-off. To overcome the statistical inefficiency from sending DP summary statistics, Liu et al. (2021) send compressed summary statistics, so as to guarantee heterogeneous privacy with lower communication costs.

In addition to structured or tabular data, Zhang et al. (2022) provide an overview of privacy threats and countermeasures to visual data in deep learning systems, which can include membership-inference, model-inversion, and property-inference attacks that can exploit trained image models to reveal or infer sensitive data elements, as well as reviewing techniques for defence, including differential privacy, adversarial obfuscation, and homomorphic encryption. Combined, this literature suggests that privacy-preserving AI is an emerging and evolving space where formal guarantees (e.g., mathematical bounds from DP) need to be considered alongside practical considerations of model utility, computational cost and client heterogeneity.

2.4 Governance, ethics and regulatory compliance

The last stream of literature places AI in the realm of changing regulations and regulations that address cybersecurity and privacy protection. Achuthan et al. (2024) perform a broad topic modelling analysis of the literature on cybersecurity and AI, and conclude that explainability, ethics and privacy preserving data mining are emerging, but relatively under-researched topics compared to the core detection topics, and that this gap needs to be addressed to ensure that the defence using AI can be trusted and auditable in industries with regulation. Goncalves and Correia (2026) address this gap directly by introducing a modular explainable-AI (XAI) engineering framework that sends information about model explanations and technical evidence to an audit-ready document, both according to GDPR and the EU AI Act, urging explainability to be considered a first-class engineering requirement, ‘compliance-by-design,' not an afterthought added to a deployed model.

As AI-driven automation of various services proliferates the number of identities (containers, service accounts, IoT devices, and others) and certificate-management and continuous-monitoring requirements mount, Chowdhary (2025) explores how the mushrooming of non-human identities has outstripped traditional identity-governance strategies and needs AI-assisted oversight instead of manual oversight. More broadly, Mohammadiounotikandi (2024) and Chukwunweike et al. (2024) both discuss the dual-use nature of AI technologies in cybersecurity, highlighting that the same features which can bolster security can also make it easier for bad actors, and propose a multi-stakeholder framework of technical standards, corporate policy, and international regulation to manage the dual use of AI responsibly.

Fragmentation of jurisdictions adds to the governance challenge. Although the GDPR and the EU AI Act form the basis of much of the reviewed literature that is focused on Europe, similar, but not identical, requirements are starting to appear in other parts of the world: the article on federated learning and the personal information protection law in China for example shows that privacy-by-design requirements are becoming a world expectation, if not a European one, and that the precise technical standards of compliance vary by jurisdiction. This fragmentation poses practical challenges for multinational organisations, as they have to develop AI security and privacy architectures, often to comply with the most stringent of the relevant regimes, rather than a single harmonised regime, in support of the arguments made by Chowdhary (2025) and Achuthan et al. (2024) for the need for a modular and mappable approach to governance rather than a hard-coding of a single jurisdiction's requirements.

The governance literature suggests there is a growing consensus that technical robustness and privacy-preservation mechanisms are required, but not enough. These need to be integrated into auditable, explainable and legally accountable governance bodies for AI-driven digital environments to be considered trustworthy.

2.5 Synthesis: From Fragmented Techniques to an Integrated Framework

The literature reviewed reveals that there is a complex dependency structure instead of four independent research problems, when traversed across the four thematic streams outlined above. Detection capability can only be as reliable as the system's demonstrated robustness against adversarial manipulation, methods for training the system that ensure privacy will determine which data an organisation will be allowed to use to create the detection system in the first place; and governance mechanisms will decide whether or not the resulting system can be operated lawfully and with accountability. This interdependency relationship is made clear by several authors. For example, while access-control and privacy-preserving data-mining techniques were formerly discussed separately from detection techniques in the same publications, they are now appearing together in the same publications (Achuthan et al., 2024); and Goncalves and Correia (2026) explicitly design their compliance framework to incorporate technical evidence from upstream detection and privacy-engineering pipelines. This convergence calls for the conceptual framework illustrated in Figure 1, which views the concept of AI-powered cybersecurity and privacy protection as a five-layer stack of technical elements and capabilities (digital environment and threat landscape, AI detection, adversarial-robustness processing, privacy preservation and automated response, governance and regulatory compliance, and a trustworthy digital environment outcome) that are interdependent and continuously fed back by incidents, audit outcomes and regulatory developments, which retrain and reconfigure the upstream elements.

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1720158

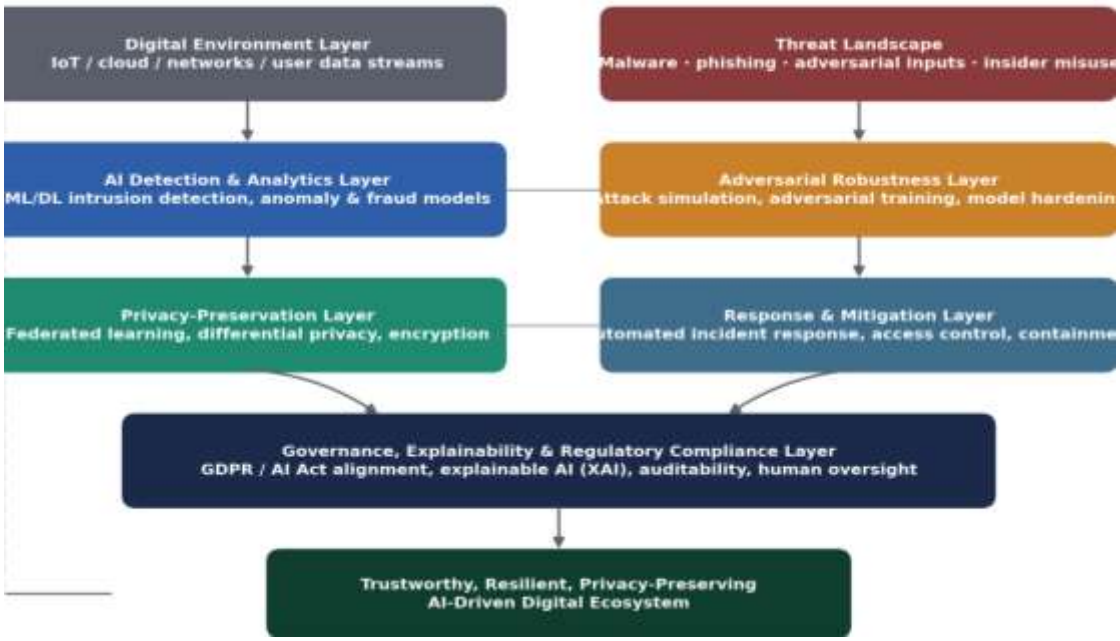

Figure 1: Conceptual frameworks between the digital environment, AI detection, adversarial robustness, privacy preservation and governance layers

III. METHODOLOGY

This study uses a systematic narrative literature review (SNL) methodology suitable for the synthesis of a conceptually diverse and a growing stream of research from the computer science, information systems and legal fields. The review process was conducted in four steps: identification, screening, thematic coding and synthesis.

Identification: A structured search was performed in Scopus-indexed articles, and in Directory of Open Access Journals (DOAJ) listed journals, such as MDPI journals (Sensors, Applied Sciences, Future Internet, Entropy, Journal of Cybersecurity and Privacy), IEEE Xplore, Springer (Artificial Intelligence Review), Elsevier (Journal of Information Security and Applications, Computer Networks), Frontiers in Big Data, and the arXiv preprint records for emerging research that has not yet been indexed. Some of the search terms included: AI/ML/deep-learning descriptors and cybersecurity and privacy descriptors (e.g., "machine learning intrusion detection," "adversarial machine learning survey," "federated learning differential privacy," "explainable AI GDPR compliance").

Screening: Records were added where they: (a) were published in the 2015-2026 time frame; (b) explained how AI/ML technologies relate to cybersecurity, adversarial robustness, privacy-preserving computation, or AI governance/regulatory compliance; and (c) had a DOI or other persistent identifier that enabled verification. No blogs or sources were included that were not peer-reviewed, or were purely promotional. Following the set of inclusion criteria, 21 sources were considered and make up the analytic corpus of this review (see References).

Thematic coding: All of the sources were read in full, and each was categorized into one of four mutually exclusive primary themes, each corresponding to one of the four main contributions of the article. AI-based threat detection and intrusion detection, adversarial AI and model robustness, privacy-preserving techniques for AI, and governance, ethics, and regulatory compliance. If a source covered more than one theme, it was coded to the theme that was its main contribution, based on the research goal/aim and abstract of the source.

© JUL 2026 | IRE Journals | Volume 10 Issue 1 | ISSN: 2456-8880
DOI: https://doi.org/10.64388/IREV10I1-1720158

Synthesis: The coded corpus was then analysed to find out: (a) the relative focus of the research for each of the four themes, (b) the technical trade-offs that were reported frequently across the studies, and (c) the common themes of convergent or divergent findings on the trustworthiness of the AI-based cybersecurity systems. This approach of thematic-frequency is a well documented, transparent way of characterising the structure of an emerging interdisciplinary literature and is reported quantitatively in Section 4 to aid in the reproducibility of the classification.

IV. RESULTS

The twenty-one reviewed sources can be divided into six sources (28.6%) focused on AI-based threat/ intrusion detection, five sources (23.8%) on adversarial machine learning and model robustness, five sources (23.8%) on privacy-preserving AI methods, and five (23.8%) on governance, ethics and regulatory compliance (Table 2; Figure 2). This near-even spread over the latter three themes reveals a marked shift within cybersecurity research from the traditional, detection-oriented focus, towards a much more diversified field, with adversarial robustness, privacy engineering and governance becoming research areas of comparable sizes as the core research on detection.

Table 2. Thematic Classification of Reviewed Literature (n = 21)

Thematic Category	Studies (n)	Share (%)
AI-based threat & intrusion detection	6	28.6
Adversarial machine learning & robustness	5	23.8
Privacy-preserving AI (FL / differential privacy)	5	23.8
Governance, ethics & regulatory compliance	5	23.8
Total	21	100.0

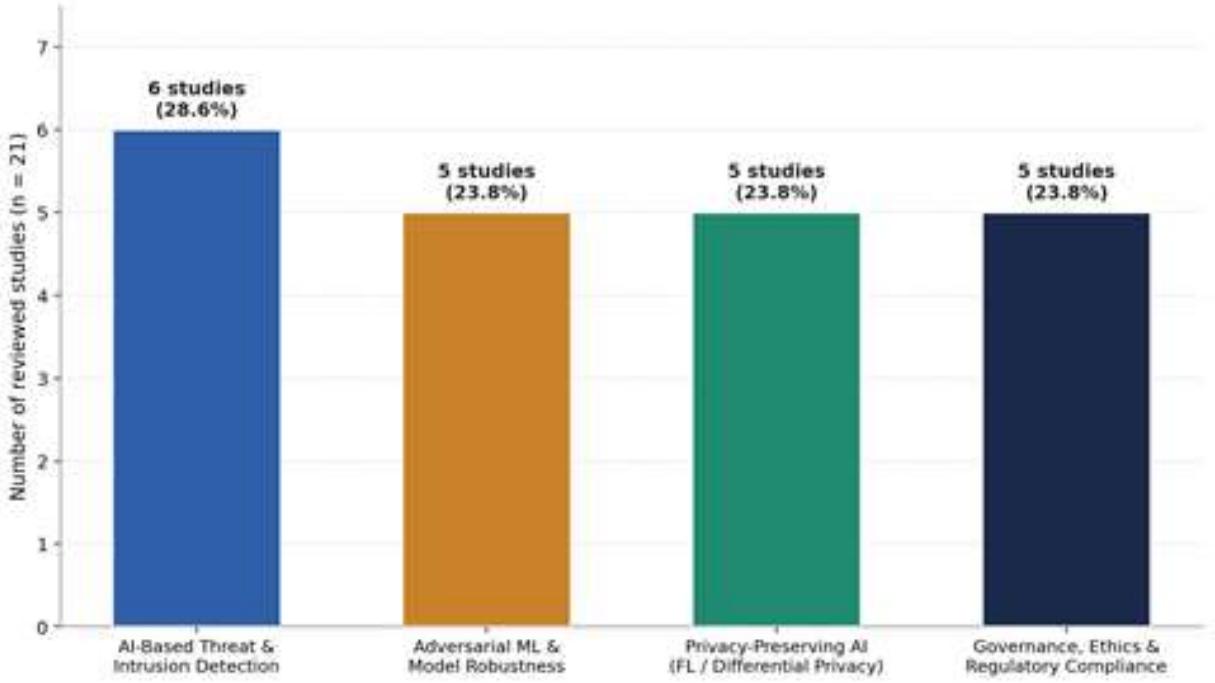

Figure 2: Literature review (n = 21 studies) on the distribution of the reviewed literature according to theme

In addition to the quantitative distribution of position, there were three patterns that were found throughout the corpus that were of a qualitative nature. First, in the detection literature, a deep-learning-based classifiers' superiority over classical ML baselines on intrusion datasets has been reported repeatedly in the past when tested on the detection benchmarks, whereas this superiority is either not as significant or

even reversed when evaluated in different deployment scenarios, that is, on out-of-distribution traffic or on resource constrained edge devices (Figueiredo et al., 2023; Heidari & Jamali, 2023), rendering headline accuracy figures more subject to dataset provenance and deployment context than absolute performance claims.

As for adversarial robustness in adversarial literature, all adversarial studies reviewed that evaluated adversarial conditions reported a decrease in the performance of their detectors when compared with their performance under benign conditions, while no single defense technique has been proven to be robust against all types of attacks tried: adversarial training, ensemble diversification and input sanitisation each proved successful against some attack types and others.

This is also reflected in the temporal pattern of the corpus: most of the sources related to the governance theme and privacy-preserving-AI (nine out of ten) are from the past few years (2023 and later), while the sources for the detection and adversarial-robustness themes are more evenly distributed across the years (including the foundational study on adversarial examples, Goodfellow et al. 2015). The distribution aligns with a view that the interpretation of adversarial robustness and of detection is a relatively mature line of research, while the interpretation of AI governance, explainability-by-design and of personalised privacy budgeting is a more recent research reaction to more stringent regulatory requirements, and not a long-standing subfield of its own.

Third, in the privacy-preservation literature, a clear utility – privacy trade-off was reported, with higher levels of differential privacy (lower privacy-budget epsilon values) leading to lower model utility in each study that quantified the trade-off, which led to the popularity of the personalised and utility-enhanced variants suggested by Shen et al. (2023) and Ranaweera et al. (2025). The governance literature, on the other hand, has now come to a consensus that explainability mechanisms are now considered not only as a usability tool but also as a regulatory requirement for legal and compliant automated decision-making, such as the GDPR and the EU AI Act (Goncalves & Correia, 2026).

V. DISCUSSION

The results presented in Section 4 have several implications for future research and practice with AI in a digital environment. First, the distribution of research focus between detection, adversarial robustness, privacy, and governance is approximately equal, indicating that cybersecurity research is no longer solely concerned with increasing the accuracy of the detection process, but rather, the growing number of researchers is considering the challenge of securing AI as a socio-technical issue. As noted in this review, the proportion of the literature that includes explainable, ethical, and privacy preserving methods continues to increase, though they are not as well represented as other methods in the literature despite being relatively small compared to detection research (which makes up the largest proportion of the literature, as argued by Achuthan et al., 2024).

Second, there is a clear practical implication from the consistent results that adversarial conditions lead to a performance reduction of the detector (Alotaibi & Rassam, 2023; Anthi et al., 2021; He et al., 2023): it is likely that organisations that test their AI supported security solutions under benign conditions only will overestimate their operational abilities. This is especially so for critical-infrastructure applications, where Pinto et al. (2023) and Anthi et al. (2021) both note that the stakes in invisible intrusions in ICS are safety-related. No single hardening technique was seen as the predominant approach in all the adversarial-robustness studies reviewed, suggesting a defence-in-depth approach of using adversarial training, ensemble methods, and continuous red-teaming.

Third, the privacy–utility trade-off that has been observed in the federated learning and differential privacy literature (Liu et al., 2021; Ranaweera et al., 2025; Shen et al., 2023) indicates that the design of privacy-preserving AI is an open space and not a solved engineering problem—one that must be calibrated to the specific context for which it is being developed. Organizations with strict data protection standards might have to trade off some performance of the model to ensure adherence to the data protection laws, and the personalization of privacy-budgeting proposed by Shen et al. (2023) is a potential way for

balancing heterogeneous data protection-risk requirements in a single federated system. Notably, federated learning and differential privacy have different threat models: FL aims at preventing raw-data centralisation, while DP aims at limiting the risk of inference from outputs of a model (Ranaweera et al., 2025); as a result, they are increasingly considered to be complementary techniques, rather than mutually exclusive ones.

Fourth, the literature of governance highlights explainability-by-design (Goncalves & Correia, 2026), suggesting a more general trend towards an increased alignment of technical practice of AI-security and legal compliance obligations. Historically, explainability was a feature of AI systems that was considered a usability or debugging tool, but it is now seen more as a structural requirement for process of lawful automated decision-making, especially when using 'high-risk' AI systems as per the EU AI Act. This regulatory push also leaves no escape for security engineering teams that try to add audit-logging and interpretability tooling onto the AI pipeline, as noted by Chowdhary (2025), who states that machine-identity and access-governance frameworks have not been able to catch up with automatic processes driven by AI.

Collectively, these results align with the notion that illustrated in Figure 1, that none of these layers alone can effectively provide cybersecurity and privacy protection. A strong adversarial robustness evaluation will not give the illusion of security; a privacy preserving training regime that doesn't actually explain its training methodology will not meet current regulatory accountability requirements; a governance policy that does not tie to technical implementation will not be effective. The literature reviewed here thus suggests the need for the most robust design paradigm for organisations that operate in AI-enabled digital environments, namely, integrated, cross-layer architectures as opposed to point solutions.

Another implication is not just with the technology, it's also about the capacity of the organization. Some of the reviewed studies also presume access to specialized knowledge not always available among organisations, specifically smaller operators of critical infrastructure, and susceptible to the threat landscape described by Pinto et al. (2023) and Heidari and Jamali (2023), but without having access to that knowledge. This indicates that what is documented in the academic literature, is sophisticated enough to rely on, compared to what is possible in realistic deployments, and suggests a need for standardised and less resource intensive tooling to be able to leverage the benefits identified in this review beyond well resourced organisations, such as pre-hardened detection models, off-the-shelf differential-privacy libraries, and template compliance-documentation pipelines. Finally, human oversight is mentioned in both the GDPR's obligations on automated decision-making and the EU AI Act's obligations on human oversight, and is also a recurring theme throughout the governance literature as a whole: it is assumed that, aside from the data scientists who designed the underlying model, there is a competent human party who can make sense of and, if needed, override an AI system's output.

5.1 Practical Recommendations

Based on the patterns that have been found within Section 4, four recommendations for practitioners are outlined below. First, organizations that implement an AI-based intrusion detection system must make adversarial evaluation (ee), including the selection of adversarial examples, as part of a regular process of validating their models, not simply as a research task.Second, adversarial evaluation (ee) should be done as a regular practice during model validation, not just as a research project, where adversarial examples are selected, such as the ones catalogued by He et al. (2023) and Jmila and Khedher (2022). Then, from the results of the different attack types, there was no single robust defensive technique, which further means that security teams should consider implementing defence-in-depth architectures, as recommended by Alotaibi and Rassam (2023), such as adversarial training, ensemble diversification and continuous monitoring.

Thirdly, organisations should explicitly describe the regulatory threat model that they are facing before deciding on the choice of privacy preserving training architectures, as they cater to different types of threats – FL is best suited for addressing raw data centralisation risk, while DP can be used to bound inferential leakage from model outputs (Liu et al., 2021; Ranaweera et al., 2025) – and a wrong choice could lead to a false sense of security. Fourth, to be

structurally more compatible to the evidentiary needs of frameworks like the GDPR (Article 22) and the EU AI Act (high-risk system obligations), as well as being more costly, explainability tooling and audit-logging infrastructure must be designed into AI security systems from the beginning, not after a compliance inquiry or a breach notification obligation.

Not all of the limitations of narrative syntheses apply to this review. Although the corpus covers a variety of databases and disciplines, it is still a purposive sample and in instances where studies had multiple themes, there was some interpretive judgement involved in the thematic classification. Qualitative synthesis is useful as a first step, but will only be as precise as the larger, PRISMA-compliant corpus that might be used in future systematic reviews using formal meta-analytic methods to quantify effect sizes, such as the degree of accuracy loss in the adversarial setting or the empirical privacy–utility frontier of differential privacy.

VI. CONCLUSION

Twenty-one peer-reviewed and indexed sources have been analysed in this paper, focusing on the impact of AI on cybersecurity defence and privacy protection in the era of AI-driven digital environments. It demonstrates the superior performance of an AI-based intrusion detection under benign scenarios when compared to signature based methods, mathematically principled but useful privacy guarantees in federated learning and differential privacy, and the increasing requirements for explainability and auditability as conditions for trustworthy AI use in regulation. No single technology is capable of delivering detection and robustness, privacy engineering and governance, it is the combination of them, as illustrated in the conceptual architecture of this paper, which is capable of delivering these various aspects of it.

In short, the following three developments were identified in the literature reviewed that will impact future research and practice. Second, privacy-preservation and detection architectures developed today should be aware of the fact that they may need to be adapted to post-quantum primitives in the future, due to the increased interest in quantum-resilient cryptography and quantum machine learning, pointed out by Achuthan et al. (2024). Second, with the growing number of non-human, AI-driven identities that Chowdhary (2025) has documented, it means that identity and access governance systems should not be considered a stand-alone workstream but should evolve alongside identity detection and privacy methods. Third, the evidence-routing method to compliance proposed by Goncalves and Correia (2026) suggests that, in the future, explainability infrastructure will not be a product of compliance, but rather a source of data to support retraining of detection-models, as well as adversarial-robustness testing, thus extending the compliance-security-technology-to-compliance feedback loop shown in Fig. 1 and further blurring the lines between technical security engineering and regulatory compliance.

The results suggest that adversarial evaluations and explainability tools are not "nice to have" features, but should be integrated as a core part of AI security engineering, and should be chosen and tuned based on the context and the threat model of the deployment, instead of being universal. For researchers, the relatively even coverage of the reviewed literature across detection, robustness, privacy and governance areas indicates rich areas for interdisciplinary research that directly links these areas together, such as work that quantifies the effect of explainability mechanisms on the adversarial robustness of an algorithm, or on the accuracy of detection in a federated intrusion-detection system, depending on the size of the personalised differential-privacy budget. Ongoing and robust focus on these four dimensions will be critical to creating cybersecurity and privacy protections that are both technically effective and legally responsible and socially trustworthy as AI becomes increasingly entrenched in the digital infrastructure. The power of AI-powered digital environments will not be reliant on any single breakthrough algorithm, but rather on the consistent discipline and ongoing implementation of detection, robustness, privacy and governance as a unified, organisational best practice.

REFERENCES

[1] Achuthan, K., Ramanathan, S., Srinivas, S., & Raman, R. (2024). Advancing cybersecurity and privacy with artificial intelligence: Current trends and future research directions. *Frontiers*

in Big Data, 7, Article 1497535. https://doi.org/10.3389/fdata.2024.1497535

[2] Ahmad, Z., Khan, A. S., Shiang, C. W., Abdullah, J., & Ahmad, F. (2021). Network intrusion detection system: A systematic study of machine learning and deep learning approaches. *Transactions on Emerging Telecommunications Technologies, 32*(1), Article e4150. https://doi.org/10.1002/ett.4150

[3] Alotaibi, A., & Rassam, M. A. (2023). Adversarial machine learning attacks against intrusion detection systems: A survey on strategies and defense. *Future Internet, 15*(2), 62. https://doi.org/10.3390/fi15020062

[4] Anthi, E., Williams, L., Rhode, M., Burnap, P., & Wedgbury, A. (2021). Adversarial attacks on machine learning cybersecurity defences in industrial control systems. *Journal of Information Security and Applications, 58*, Article 102717. https://doi.org/10.1016/j.jisa.2020.102717

[5] Chowdhary, S. (2025). Protecting the digital ecosystem: AI's dual role in machine identity security. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 11*(1), 1986–1996. https://doi.org/10.32628/CSEIT251112200

[6] Chukwunweike, J. N., Ayodele, P. A., Yussuf, M., & Atata, B. B. (2024). The intersection of artificial intelligence and cybersecurity: Safeguarding data privacy and information integrity in the digital age. *International Journal of Computer Applications Technology and Research, 13*(9), 14–26. https://doi.org/10.7753/IJCATR1309.1002

[7] Figueiredo, J., Serrão, C., & de Almeida, A. M. (2023). Deep learning model transposition for network intrusion detection systems. *Electronics, 12*(2), 293. https://doi.org/10.3390/electronics12020293

[8] Goncalves, A., & Correia, A. (2026). Engineering explainable AI systems for GDPR-aligned decision transparency: A modular framework for continuous compliance. *Journal of Cybersecurity and Privacy, 6*(1), 7. https://doi.org/10.3390/jcp6010007

[9] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples* (arXiv:1412.6572). arXiv. https://doi.org/10.48550/arXiv.1412.6572

[10] He, K., Kim, D. D., & Asghar, M. R. (2023). Adversarial machine learning for network intrusion detection systems: A comprehensive survey. *IEEE Communications Surveys & Tutorials, 25*(1), 538–566. https://doi.org/10.1109/COMST.2022.3233793

[11] Heidari, A., & Jamali, M. A. J. (2023). Internet of Things intrusion detection systems: A comprehensive review and future directions. *Cluster Computing, 26*(6), 3753–3780. https://doi.org/10.1007/s10586-022-03776-z

[12] Jmila, H., & Khedher, M. I. (2022). Adversarial machine learning for network intrusion detection: A comparative study. *Computer Networks, 214*, Article 109073. https://doi.org/10.1016/j.comnet.2022.109073

[13] Lim, D.-S., & Lee, S.-J. (2025). Privacy protection in AI transformation environments: Focusing on integrated log system and AHP scenario prioritization. *Sensors, 25*(16), 5181. https://doi.org/10.3390/s25165181

[14] Liu, H., & Lang, B. (2019). Machine learning and deep learning methods for intrusion detection systems: A survey. *Applied Sciences, 9*(20), 4396. https://doi.org/10.3390/app9204396

[15] Liu, J., Lou, J., Xiong, L., Liu, J., & Meng, X. (2021). Projected federated averaging with heterogeneous differential privacy. *Proceedings of the VLDB Endowment, 15*(4), 828–840. https://doi.org/10.14778/3503585.3503592

[16] Mohammadiounotikandi, A., & Babaeitarkami, S. (2024). Cybersecurity in the age of AI: Protecting our data and privacy in a digital world. *Australian Journal of Engineering and Innovative Technology, 6*(4), 86–92. https://doi.org/10.34104/ajeit.024.086092

[17] Neto, E. C. P., Dadkhah, S., Ferreira, R., Zohourian, A., Lu, R., & Ghorbani, A. A. (2023). CICIoT2023: A real-time dataset and benchmark for large-scale attacks in IoT environment. *Sensors, 23*(13), 5941. https://doi.org/10.3390/s23135941

[18] Pinto, A., Herrera, L.-C., Donoso, Y., & Gutierrez, J. A. (2023). Survey on intrusion detection systems based on machine learning techniques for the protection of critical infrastructure. *Sensors, 23*(5), 2415. https://doi.org/10.3390/s23052415

[19] Ranaweera, K., Nguyen, D. C., Pathirana, P. N., Smith, D., Ding, M., Rakotoarivelo, T., & Seneviratne, A. (2025). *Federated learning with differential privacy: An utility-enhanced approach* (arXiv:2503.21154). arXiv. https://doi.org/10.48550/arXiv.2503.21154

[20] Shen, X., Jiang, H., Chen, Y., Wang, B., & Gao, L. (2023). PLDP-FL: Federated learning with personalized local differential privacy. *Entropy, 25*(3), 485. https://doi.org/10.3390/e25030485

[21] Zhang, G., Liu, B., Zhu, T., Zhou, A., & Zhou, W. (2022). Visual privacy attacks and defenses in deep learning: A survey. *Artificial Intelligence Review, 55*(6), 4347–4401. https://doi.org/10.1007/s10462-021-10123-y