*ELMİ TƏDQİQAT Beynəlxalq Elmi Jurnal.* İyun 2026, Cild: 6, Sayı: 6, s. 254–264
*SCIENTIFIC RESEARCH International Scientific Journal.* June 2026, Volume: 6, Issue: 6, pp. 254–264
DOI: https://doi.org/10.36719/2789-6919/60/254-264

# AI Defense Automation: The Ethics of Security through AI in Medium Enterprises

**Rafael Mehdiyev**

***Abstract.*** *With the growing adoption of AI in cybersecurity, the process has been made more efficient in terms of threat detection. On the downside, the integration of the technology into SME cybersecurity practices happens in conditions of resource shortages, which makes it difficult not only to implement but also to develop the ethics and governance needed for the appropriate use of the technology. Thus, this research assesses the appropriateness of existing models for guiding AI-enabled security practices among SMEs. For comparative content analysis, six well-known models have been chosen: ISO/IEC 27001:2022, ISO/IEC 27005:2018, NIST CSF, NIST AI RMF, EU AI Act (Regulation (EU) 2024/1689), and ENISA Threat Landscape 2022.*
*Two gaps in the frameworks' capacity for AI-enabled security practices are revealed. The first one is a domain-integration gap since the ISO/IEC standards are focused on information security and do not consider AI ethics; only the NIST AI RMF and EU AI Act mention issues such as audits of algorithmic bias, explainability, and human involvement. The second is an SME adaptation gap, meaning that no model has provisions specific to the peculiarities of SMEs and gradual implementation. The solution offered in this regard involves the implementation of a three-level governance structure consisting of strategic, operational, and ethics layers based on the ethics-by-design approach in line with ISO 27005 and NIST AI RMF.*

***Keywords:*** *cybersecurity governance, artificial intelligence, ethical AI, algorithmic bias, medium-sized enterprises, risk-based governance, explainable AI, data privacy, ethics-by-design*

---

Baku Higher Oil School, Master's student, Baku, Azerbaijan
E-mail: rafaelmehdiyev0@gmail.com
Received: 26 February 2026; Accepted: 29 March 2026; Published online: 30 June 2026


# Süni İntellekt müdafiə avtomatlaşdırılması: orta müəssisələrdə Süni İntellekt vasitəsilə təhlükəsizlik etikası

**Rafael Mehdiyev**

***Xülasə.*** *Kibertəhlükəsizlikdə süni intellektin artan tətbiqi ilə təhdidlərin aşkarlanması prosesi daha səmərəli hala gəlmişdir. Lakin texnologiyanın KOB-ların kibertəhlükəsizlik təcrübələrinə inteqrasiyası resurs çatışmazlığı şəraitində baş verir ki, bu da həm tətbiqi, həm də texnologiyanın məsuliyyətli istifadəsi üçün zəruri olan etika və idarəetmə mexanizmlərinin inkişaf etdirilməsini çətinləşdirir. Beləliklə, bu tədqiqat mövcud modellərin KOB-lar arasında süni intellektə əsaslanan təhlükəsizlik təcrübələrinə rəhbərlik etmək üçün uyğunluğunu qiymətləndirir.*

*Müqayisəli məzmun təhlili üçün altı tanınmış model seçilmişdir: ISO/IEC 27001:2022, ISO/IEC 27005:2018, NIST CSF, NIST AI RMF, AB-nin Süni İntellekt Aktı (Regulation (EU) 2024/1689) və ENISA Təhdid Mənzərəsi 2022.*
*Çərçivələrin süni intellektə əsaslanan təhlükəsizlik təcrübələri üçün imkanlarında iki boşluq aşkar edilmişdir. Birincisi, domenlərarası inteqrasiya boşluğudur: ISO/IEC standartları informasiya təhlükəsizliyinə yönəlmiş olub süni intellekt etikasını nəzərə almır; yalnız NIST AI RMF və AB-nin Süni İntellekt Aktı alqoritmik qərəzin auditi, izahlılıq və insan iştirakı kimi məsələlərə toxunur. İkincisi, KOB-lara uyğunlaşma boşluğudur: heç bir modeldə KOB-ların xüsusiyyətlərinə və mərhələli tətbiqə dair müddəalar mövcud deyil. Bu baxımdan təklif edilən həll ISO 27005 və NIST AI RMF ilə uyğun olan etika-dizayn yanaşmasına əsaslanan strateji, əməliyyat və etika təbəqələrindən ibarət üç səviyyəli idarəetmə strukturunun tətbiqini nəzərdə tutur.*

***Açar sözlər:*** *kibertəhlükəsizlik idarəetməsi, süni intellekt, etik süni intellekt, alqoritmik qərəz, orta ölçülü müəssisələr, riskə əsaslanan idarəetmə, izah edilə bilən süni intellekt, məlumat məxfiliyi, etika-dizayn*

---

Bakı Ali Neft Məktəbi, magistrant, Bakı, Azərbaycan
E-poçt: rafaelmehdiyev0@gmail.com
Daxil oldu: 26 Fevral 2026; Qəbul edildi: 29 Mart 2026; Onlayn dərc edildi: 30 İyun 2026


## Introduction

Almost every enterprise today, no matter what business sector it operates in, relies heavily on its digital infrastructure – to a degree unprecedented a decade ago. Modern cloud infrastructures, IoT deployments, big data analysis, and hybrid working models have broadened the scope of IT perimeter to the extent that a conventional rule-based security system is unable to cover the entire perimeter of operations and provide the necessary level of protection. This challenge is especially pronounced for SMEs. These enterprises possess enough data and infrastructure resources to be targeted by attackers, while lacking sufficient security-related resources compared to major corporations. According to ENISA (ENISA, 2022) and the Verizon Data Breach Investigations Report, ransomware attacks and data exfiltration campaigns cause relatively more damage for organizations which do not have continuous security monitoring capability, as the period between compromise and discovery in such cases is longer. The technological advances in the field of cybersecurity have provided SMEs with opportunities to employ machine-learning-based security systems in their digital infrastructure. The development of anomaly detection systems, user and entity behavior analytics, and Security Orchestration, Automation, and Response (SOAR) solutions now makes such solutions available for procurement in a form of managed services or cloud solutions at a reasonable cost without the need for significant additional personnel costs. Furthermore, the concepts such as a SOC have been traditionally available only for large financial and governmental institutions; however, nowadays, they can be simulated in the context of SMEs through a joint operation of MSSPs and AI-based security solutions.

However, the implementation of AI-based systems poses certain challenges in terms of ethics and governance. First, these solutions rely on processing sensitive data related to employees' actions; second, they make decisions which have direct impact on users and may affect their rights; finally, models used for decision making are often based on complex logic incomprehensible even to the developers of the software. Each of these features raises important questions which are still unaddressed by the current state of cybersecurity: Who is accountable for incorrect decisions made by an AI solution blocking an employee's account? When an anomaly detection solution provides

biased results with respect to particular user categories, how should an enterprise react to it? Under the scope of the GDPR or CCPA, what data is considered acceptable to collect and analyze for the purposes of security monitoring? These are questions of ethics, policy, and governance – not just of technology. However, in the course of SME digitalization, it has become clear that technical solutions precede appropriate governance mechanisms and ethical policies.

Hence, there is an organizational gap in which security operations heavily rely on AI solutions, while appropriate measures to ensure their governance and ethical implementation are absent or simply borrowed from different contexts. This study aims to bridge this gap. Specifically, the research question addressed by this project is as follows: how should ethics, policy, and governance frameworks for AI-enabled cybersecurity operations be structured in order to reflect the peculiarities of SMEs' resource constraints and risk profiles? The contributions of this study consist in developing a conceptual model for an SME governance readiness for AI-enabled security operations. The proposed model synthesizes governance recommendations from international standards related to artificial intelligence (NIST AI RMF (NIST, 2023)), cybersecurity (ISO/IEC 27001 (ISO/IEC, 2022); ISO/IEC 27005 (ISO/IEC, 2018)) and AI (EU AI Act (European Parliament and Council, 2024)) and adjusts them to the specific needs of 50-to-250-employees enterprises operating under the jurisdictions of Europe and the USA.

ML applications have addressed cybersecurity concerns for over twenty years already. However, the currently available generation of operational systems differs qualitatively from earlier approaches (Sarker, 2021). Modern AI-based SOC frameworks consist of several levels of functions. First, there is a stage of data acquisition, where network traffic log files, endpoints telemetry, application server events, access management logs, and information about user behavior is collected and integrated into SIEM solutions. Second, an ML model is used to recognize anomalies and assess risk factors (such ML models are commonly represented by unsupervised and semi-supervised methods, such as isolation forests, autoencoders, or LSTM networks). Finally, a response is automated using SOAR solutions and may include automatic isolation of an infected endpoint, deauthentication, or blocking of network communications. These advantages are quantifiable. Automating triage processes leads to lower MTTD and MTTR. Thus, detecting breaches within 200 days leads to significantly lower remediation costs in comparison with later detection. For SMEs lacking 24/7 coverage by humans, automation enables monitoring in those hours when analysts sleep, thereby adding actual value from the risk perspective (ENISA, 2022).

In the case of algorithmic bias, the root cause is again intrinsic to the nature of supervised and semi-supervised machine learning: the biases present in the historical dataset get transferred through the algorithmic model into its predictions (NIST, 2023; Mehrabi et al., 2021). If previous incidents disproportionately targeted certain employee cohorts, network zones, or applications, an anomaly detection algorithm based on this dataset will flag any future events in these cohorts/segments/zones/applications as anomalies, irrespective of actual risks. The bias here is caused by past human decisions about surveillance efforts and not the algorithm itself; however, the way machine learning works, such biases in the training set are reproduced in predictions, even unintentionally (Mehrabi et al., 2021). The problem is magnified for SMEs due to their small historical sample size; for such enterprises, vendor-supplied or community-supplied anomaly detectors, already trained on historical data, pose an additional risk, since the characteristics of that training set are unknown. Explainability is a separate but closely related issue. High-performance anomaly detectors tend to be black-boxes, for instance, a deep autoencoder that detects anomalies via high reconstruction errors cannot provide human-understandable explanations for why that is the case (NIST, 2023; Arrieta, 2020).

In consequence, when such automated decisions lead to automated action (e.g., blocking a user account), there is no opportunity for the target individual or the security operator taking charge of the situation to investigate the decision's reasoning. The EU AI Act (European Parliament and Council, 2024) recognizes this issue explicitly, mandating human oversight and transparency reporting mechanisms for high-risk AI applications. In the context of SMEs, where the individual handling the security event may not necessarily have the necessary knowledge to contest the AI's recommendation, the problem is exacerbated by the lack of the relevant background knowledge. The data privacy issue stems from the fundamental trade-off between effective detection and minimization principles of GDPR Article 5(1)(c) (European Parliament and Council, 2016); ISO/IEC 27001 Annex A.8 (ISO/IEC, 2022). Behavioral anomaly detection requires monitoring and processing of personal information about employees – keystrokes, applications, files and messages used, etc. – precisely the kind of data that must be minimized to comply with legal requirements. However, there is significant technical debate regarding what constitutes "necessity" in data collection and how much data is required for accurate behavior analysis (ISO/IEC, 2018). Therefore, there exists a regulatory conflict that cannot be resolved at a policy level: the most accurate models require the richest datasets available. Finally, there is the problem of human oversight erosion over time. Anomaly detection systems that evolve technologically are likely to increase automation of actions taken upon identifying a suspicious activity. The human operator's responsibilities thus move from taking the decision to investigating exceptions; however, even that task becomes increasingly difficult when alert fatigue leads to automated response mechanisms, documented in ENISA guidelines (ENISA, 2022). The problem is well recognized in the NIST AI RMF framework, which calls for special consideration when designing human oversight mechanisms for govern and manage activities.

There exist multiple international governance frameworks that cover certain elements of AI ethics and cybersecurity governance, yet there is none that covers both areas and that would be adapted for use in the SME context.

The NIST Cybersecurity Framework (CSF 1.1) (NIST, 2018) offers a five-functions structure for cybersecurity governance: Identify, Protect, Detect, Respond, Recover. These functions have established themselves as the standard structural basis for the practice of cybersecurity governance for enterprises in the United States, and the CSF continues to be widely adopted internationally as well. It is risk-based and sector-neutral, which makes it flexible but unable to accommodate the AI-specific ethical concerns. Algorithmic biases, the explainability of the algorithm's functioning, and automated decision systems in particular are beyond the scope of the framework.

On the other hand, the NIST AI Risk Management Framework (AI RMF 1.0) (NIST, 2023) offers an element of compensation. Its four governing functions – Govern, Map, Measure, Manage – address explicitly such issues of trustworthiness as explainability, algorithmic bias, and accountability. This is the only framework addressing specifically AI-related governance issues as discussed above. However, the AI RMF does not address cybersecurity in SMEs specifically: it offers generic AI governance solutions without taking into account security operations, response times, SOAR architecture considerations, and specific types of data generated by SOC architectures.

The ISO/IEC 27001:2022 (ISO/IEC, 2022) specifies requirements of the Information Security Management System (ISMS) that govern information security at the organization level, encompassing asset management, access control, supplier security management, and incident management practices. The new edition includes updates on cloud computing but does not offer any AI-specific guidance. However, ISO 27001 serves as the basis for cybersecurity ISMS, and this is where the relevant AI-based controls will have to be implemented.

The ISO/IEC 27005:2018 (ISO/IEC, 2018) specifies the methodology for the risk assessment and management practices required for ISMS operation under ISO 27001. Asset management, threat modeling, and prioritization of risks are key parts of this approach and serve as the logical basis for AI system risk assessment in an ISMS context. The framework does not specify how AI-specific risks should be assessed (algorithmic bias, explainability, automation malfunction), however the methodology provided is general enough to accommodate them with proper adaptations.

The EU AI Act (Regulation (EU) 2024/1689) (European Parliament and Council, 2024) offers the most comprehensive regulation of AI ethics currently in force. It classifies as high-risk any AI system used for critical infrastructure, law enforcement, and employment applications and mandates conformity assessment, bias testing, documentation of transparency, and human oversight prior to deployment. Any AI systems used for cybersecurity in an organization managing critical infrastructures will be classified as high-risk, while those that don't fall under this definition require nonetheless compliance with the minimum requirements of the act to be fully governed.

The ENISA Threat Landscape 2022 (ENISA, 2022) does not offer a cybersecurity governance framework but rather describes the current threat landscape in relation to which all cybersecurity governance approaches must be formulated. Findings concerning the increased frequency of ransomware attacks, supply chain attacks, and internal threats serve as a justification for the need to employ AI-based solutions to detect them while necessitating additional governance to ensure AI ethical performance.

**Methods**

Research approach. The study makes use of qualitative methodology, specifically framework synthesis, with systematic analysis of relevant documents on governance and regulation as the main source of information. In contrast to the previous study discussed above, this research question is normative in nature. The proposed question does not focus on practices implemented by organizations (which would be suitable for survey-based research), but rather on the most justifiable governance architecture taking into account the current regulatory context, risk assessment evidence, and other relevant factors. For normative questions like the one discussed in this paper, the best approach is framework synthesis and comparative analysis of documents, as they provide a unified conceptual model based on structural compatibility and gaps in existing authoritative frameworks, and not hypothesis verification against primary data (Arrieta, 2020).

*Data sources and analysis procedure*. The following six documents were selected due to their international legitimacy and specific relevance to the examined governance dimensions of AI-powered cyber security: NIST CSF 1.1 (NIST, 2018), NIST AI RMF 1.0 (NIST, 2023), ISO/IEC 27001:2022 (ISO/IEC, 2022), ISO/IEC 27005:2018 (ISO/IEC, 2018), the EU AI Act (Regulation (EU) 2024/1689) (European Parliament and Council, 2024), and ENISA Threat Landscape 2022 (ENISA, 2022). Secondary sources discussing ethics risks related to the application of AI technology in cybersecurity operations, as well as governance capacities for SMEs were selected as the theoretical framework (Von Solms; Van Niekerk, 2013; Sarker, 2021).

The analysis of each framework was done according to eight governance dimensions deemed important for governing AI-enabled cybersecurity operations: Risk Identification; Algorithmic Bias Auditing; Explainability and Transparency; Data Privacy and Minimization; Human Oversight Mechanisms; Incident Response; Guidance for SME Adaptation; and Implementation Phasing Support. The coverage level of each of these dimensions in the analyzed framework was evaluated as "full," "partial," or "absent." As a result, a comparison chart was made (see Table 1) identifying where the gaps in the framework coverage overlaps and whether there are gaps in SME-specific aspects

only. Single-rater coverage ratings were used to evaluate the coverage in the document; however, a lack of inter-rater reliability test results should be considered a limitation.

*Scopes and limitations*. There are three limits to the scope of the research. First, the focus organizationally will be on medium-sized companies under the EU SME definition (between 50 and 250 employees). These will be large enough to manage their cybersecurity through structured processes, yet small enough not to develop a complex governance structure like large enterprises. Second, the study will focus geographically on EU and US regulatory contexts, as the governance frameworks discussed originate there; other countries would require additional regulatory mapping work. Third, the time scope of the framework analysis will run until the latest version as of 2023.
First, empirical testing of governance models derived from frameworks is limited in this type of research; no such evidence exists for the governance model suggested in Chapter 4. It is still uncertain what effectiveness, costs, and barriers to adoption it would demonstrate in practice. Second, this approach does not allow for studying the process of threat actor adaptation to organizational defenses. If cyber attackers know the detection logic implemented by a company, they might adapt their actions accordingly; current frameworks do not consider this possibility.

**Results**

*The governance gap in SME ai-enabled cybersecurity*. The gaps that arise through the dimensional analysis performed above can be categorized into two major categories (as seen in Table 1). The first type is the integration gap. While AI-ethics control such as algorithmic bias auditing, explainability, and oversight are accounted for in the NIST AI RMF (NIST, 2023) and EU AI Act (European Parliament and Council, 2024), they are not present in ISO 27001 and ISO 27005, which are mostly used by SMEs as the underlying ISMS framework. On the other hand, the operational aspects of incident response are adequately covered in the NIST CSF (NIST, 2018) and ISO framework, but only partially discussed in AI-specific frameworks. The other type of gap is the adaptation gap, which pertains to the lack of guidance specifically tuned for SME constraints and implementation roadmap in all six frameworks reviewed.

The presence of the SME constraint causes this problem to manifest in a unique manner. While governance gaps can be mitigated by establishing certain roles such as the AI ethics officer and the DPO and procuring explainability and auditing systems in the case of large organizations, the same strategy cannot be applied to SMEs. In particular, since SMEs may rely on one security team to perform both monitoring of their security system (using AI-based technology), govern and adapt the company based on regulations, and provide response to incidents, the SME governance model will have to take this into consideration.

**Table 1**
*Dimensional coverage of major governance frameworks and the proposed model*

| Governance dimension | NIST CSF | NIST AI RMF | ISO 27001 | ISO 27005 | EU AI Act | Proposed framework |
|---|---|---|---|---|---|---|
| Risk identification | Full | Full | Full | Full | Partial | Full |
| Algorithmic bias auditing | — | Full | — | — | Full | Full |
| Explainability / transparency | — | Full | — | — | Full | Full |
| Data privacy / minimization | Partial | Partial | Full | Partial | Full | Full |
| Human oversight mechanisms | — | Full | — | — | Full | Full |
| Incident response | Full | Partial | Full | Partial | — | Full |
| SME-adapted guidance | — | — | Partial | — | — | Full |
| Phased implementation roadmap | — | — | — | — | — | Full |

The suggested framework aims to serve as a bridge for the two types of gaps, where the strategic and implementation levels of the framework are borrowed from NIST CSF (NIST, 2018) and ISO standards (ISO/IEC, 2022 ; ISO/IEC, 2018) as the cybersecurity foundation, and the ethical protection level takes into account the special characteristics of AI control mechanisms in NIST AI RMF (NIST, 2023) and the European Union's AI Act (European Parliament and Council, 2024).

*Framework architecture. Layer 1 – Strategic governance.* The layer covers risk-based policy-making, regulation alignment, and assignment of accountability. Its deliverables include the governance charter (what AI systems are deployed, who is accountable for their operation, and risk tolerance levels), the regulation compliance matrix (GDPR (European Parliament and Council, 2016), EU AI Act risk classifications (European Parliament and Council, 2024), ISO 27001 certification scope (ISO/IEC, 2022)), and accountability framework (whom to hold formally accountable for automated decisions made by the AI and whether any escalation process exists in case AI-produced outcomes are contested). For SMEs, the layer should be controlled by an AI-Security Governance role holder. Rather than establishing a whole new department, SMEs would appoint an existing person responsible for information security to fulfill the role (or even combine it with the Data Protection Officer [DPO] function).

*Layer 2 – Implementation.* This layer defines how AI-based systems will be deployed, monitored, and managed. The layer includes an AI inventory and classification in accordance with the EU AI Act risk categories, controls over configurations of AI models (including access controls), SOAR playbook review procedures, and governance of AI data management (what data is collected and deleted, and based on what legal ground). This layer is linked directly to the NIST CSF Protect, Detect, Respond, and Recover functions (NIST, 2018) and to ISO 27001 controls (ISO/IEC, 2022).

*Layer 3 – Ethical safeguards.* The layer covers the ethical risk controls listed in above. The deliverables include bias audit procedures (how frequently they need to be performed, what tools to use, and what steps are to be taken in the case of detection of biases), explanation of decisions documentations requirements for automated decision making affecting users, data minimization

procedures for behavioral data sets, and Human-in-the-Loop controls defining which SOAR playbook actions require human approval before execution. This layer maps to NIST AI RMF Measure and Manage functions (NIST, 2023) and to EU AI Act requirements for High-Risk AI (European Parliament and Council, 2024).

The implementation process of the recommended governance framework can be carried out over five phases. In Phase I (Months 1-3), the position of the AI-Security Governance is created; an inventory of AI is performed; risk classification of the most important AI implementations within the EU AI Act risk levels is carried out, while the governance charter is developed, which will use the approaches of NIST CSF Identify function (NIST, 2018) and Article 6 of EU AI Act (European Parliament and Council, 2024). At the completion point, the position is created; an inventory list is made; risk classification of AI is performed.

Phase II (Months 4-6) deals with the baseline assessment. The most crucial aspect includes algorithmic bias audit for the riskiest AI solutions; AI data flows mapping and explainability requirements for the automated decision-making related to the SOAR incidents will be addressed, leveraging NIST AI RMF Map and Measure functions (NIST, 2023) and ISO 27005 (ISO/IEC, 2018). In order to consider Phase II to be completed, the algorithmic bias audit report and data flows mapping with GDPR requirement annotation are needed.

Phase III (Months 7-9) will cover developing of a modular policy in respect to AI ethics; defining human-in-the-loop thresholds for SOAR incidents; and integrating the AI governance policy into the already established ISMS within ISO 27001. For this purpose, ISO 27001 Clauses 5 and 6 (ISO/IEC, 2022) and NIST AI RMF Govern function (NIST, 2023) can be used. Readiness is confirmed if the AI governance policy is approved; the human-in-the-loop thresholds are set; and integration with the ISMS is carried out.

Phase IV (Months 10-12) will relate to the operational phase. Ethical protections will be implemented in SOC configurations and playbooks; there will be applied access controls for AI models; and staff will be trained on the governance policies and procedures, considering NIST CSF Protect and Respond functions (NIST, 2018) and NIST AI RMF Manage function (NIST, 2023). To complete this phase, SOC playbooks will be updated and all staff will have been trained.

Phase V is continuous and should be conducted on a yearly minimum schedule. This phase will include reviewing the governance approach in the context of the evolving threat landscape and new regulations; bias audit methodology will be reviewed; and the AI governance best practices will be shared with supply chain and industry peers, using ENISA Threat Landscape (ENISA, 2022) and ISO 27001 Clause 10 (ISO/IEC, 2022). The annual AI governance report and policy version history will be published.

Ethics-by-design and security-by-design are structural components in this approach. Ethics-by-design involves making ethics a requirement in the design process through requirements such as bias testing, documentation of explanation of AI system decisions, and minimum data collection practices. Security-by-design refers to the inclusion of security features at the time of designing and implementing AI systems. Both aspects follow a risk-based methodology similar to that described in ISO 27005 (ISO/IEC, 2018). The Govern aspect of the NIST AI RMF, which includes governance as a design-time activity, also supports both aspects.

**Discussion and Conclusion**

*Mechanisms underlying ethical risks*. The baseline learned by the algorithm depends on both the underlying nature of the system (organizational or operational) and how much of the past history is devoted to observing this normal state (i.e., how frequently security issues occurred in this particular system). As such, an organization where some categories of users (or application segments) were observed to have a higher than usual security risk would develop, in an unsupervised anomaly detection algorithm, an expectation that all these users have a high probability of exhibiting malicious behavior, and, vice versa. If the anomaly detection algorithm is implemented using a vendor-pretrained model, the biases present in this model may not be known to the SME purchasing this service. Therefore, there are two critical governance points to cover here – one in the purchasing phase (obliging vendors to disclose their training data bias and testing results), and the other in the operational phase (periodical revaluation of the models' performance across various user segments).

As demonstrated in the literature, deep neural networks (DNN), ensembles, and attention models, although having superior performance, are among the hardest-to-explain algorithms (Sarker, 2021; Arrieta, 2020). Although recently published research suggests potential approaches to creating inherently interpretable ML models, this approach is not yet available to SMEs implementing SOAR programs. Thus, the ethical question arises whether opaque decisions of a DNN or an ensemble is ethically acceptable in a situation of a detected anomaly. NIST AI RMF explicitly states that the human-AI teaming process is at the heart of ethical use of AI (NIST, 2023); hence, the issue for governance is not about making the decision-making transparent to the end-users but rather ensuring transparency of the overall human-AI decision-making system. Therefore, the governance task, in the context of SOAR systems, is to identify the conditions for automation of response actions and the situations where human authorization would be mandatory despite the ML algorithm's certainty.

Deep data collection and storage is an inherent component of an anomaly detection problem, but the privacy issues can be mitigated via proportionality. That is, the governance process must establish the balance between the amount of privacy intrusion and the security gains provided by this intrusion. As mentioned above, data flow analysis must be conducted to map the security purposes for which data was collected (and legal grounds for processing it), as well as retention periods.

*Theoretical implications*. The governance framework proposed here further solidifies the claim about cybersecurity moving away from being a purely technical field towards becoming more organizational and even human-centered that was advanced in Von Solms and Van Niekerk (Von Solms; Van Niekerk, 2013). With the help of artificial intelligence, cybersecurity became a challenge at least partially delegated to algorithms. In this case, organizational and human-centered approaches should extend towards those very algorithms: their training origins, decision-making processes, error conditions, and the governance of any output produced.

There are several ways to relate the three-layer model proposed above with the NIST AI RMF structure. First, the Govern function of the AI RMF correlates with the strategic governance layer. The Map and Measure functions of the same framework coincide with the ethical safeguards layer since they include bias auditing and explainability requirements. Finally, the Manage function correlates with the operational implementation layer proposed here. Such correlation is necessary because adopting the framework proposed above is supposed to be much easier for organizations implementing the NIST AI RMF than adopting another approach altogether.

*Limitations.* There are several limitations in this research that may affect its findings. Firstly, this is a conceptual model derived from the analysis of primary regulatory and standards documents, rather than empirical observations of practical implementation of the discussed programs. The possibility

of practical implementation of the suggested three-level structure, order of phases and timelines is still only a hypothesis and should be further tested via empirical research, such as case studies and organizational surveys.

The geographical limitation in this study means that there might be no straightforward application to those countries or regions where the approaches towards personal data regulation or artificial intelligence are different from that of EU and USA. The structural principles of the proposed framework are general, however, the specifics of GDPR (European Parliament and Council, 2016) legal bases, risk categories of EU AI Act (European Parliament and Council, 2024), and SME definition require a substitution in other regions.

As for Table 1's content, the coverages were defined by one person who analyzed the texts and made qualitative judgments about the extent of compliance with GDPR (European Parliament and Council, 2016). These ratings show authors' interpretation of those texts; however, readers should consider them as suggestions rather than absolute quantitative measure.

In this study, the SME criterion includes organizations with 50-250 employees. While a small firm of 60 people and large company of 240 people with production lines interconnected with AI are significantly different when it comes to risks of malicious activities, the framework still treats them as one category. This is one more limitation of the paper.

Lastly, there is a significant challenge in the field, which is the use of AI-based detection and adversarial behavior as a response to it. As more and more companies deploy AI-based systems to detect anomalies in their internal networks, threat actors adapt their techniques accordingly by making sure that their activity appears regular and thus, normal for anomaly detection models. This issue mentioned in ENISA Threat Landscape (ENISA, 2022) report cannot be solved with frameworks focused on ethical operation of artificial intelligence.

This paper investigated the extent to which available governance frameworks are ready to be applied to guide responsible use of AI-based tools for cybersecurity operations in medium-sized enterprises. Framework synthesis revealed that neither the NIST AI RMF nor any other existing framework meets the discussed criteria since none covers all of the required governance dimensions and provides appropriate implementation guidance for resource-limited organizations.
The most important takeaway is that technological readiness, regulatory compliance, and capacity to govern algorithmic decision-making cannot be achieved separately. Development of one of these three components without the others results in either a technologically advanced yet unaccountable security process, an operationally inefficient yet compliant one, or a governance program that nobody in the organization has the skills to implement.

**References**

1. Arrieta, A. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*(1), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012
2. Bowen, G. (2009). Document analysis as a qualitative research method. *Qualitative Research Journal, 9*(2), 27–40. https://doi.org/10.3316/QRJ0902027
3. ENISA. (2022). *Threat Landscape 2022*. European Union Agency for Cybersecurity.
4. European Parliament and Council. (2016). Regulation (EU) 2016/679 on the protection of natural persons with regard to the processing of personal data (General Data Protection Regulation). *Official Journal of the European Union*.

5. European Parliament and Council. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union.*
6. ISO/IEC. (2018). *ISO/IEC 27005:2018 Information technology — Security techniques — Information security risk management*. International Organization for Standardization.
7. ISO/IEC. (2022). *ISO/IEC 27001:2022 Information security, cybersecurity and privacy protection — Information security management systems — Requirements*. International Organization for Standardization.
8. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys, 54*(6), 1–35. https://doi.org/10.1145/3457607
9. NIST. (2018). *Framework for Improving Critical Infrastructure Cybersecurity (Version 1.1).* National Institute of Standards and Technology. https://doi.org/10.6028/NIST.CSWP.04162018
10. NIST. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0).* National Institute of Standards and Technology. https://doi.org/10.6028/NIST.AI.100-1
11. Sarker, I. (2021). Machine learning: Algorithms, real-world applications and research directions. *SN Computer Science, 2*(3), 160. https://doi.org/10.1007/s42979-021-00592-x
12. Von Solms, R., & Van Niekerk, J. (2013). From information security to cybersecurity. *Computers & Security, 38*(1), 97–102. https://doi.org/10.1016/j.cose.2013.04.004