Ai Agents and Full Automation in Cybersecurity: Technical, Business, Ethical and Legal Considerations

Moufid El-Khoury
Sultan Qaboos University
Ferdinand Kpieleh
University of Cincinnati
Sabine Khalil
Illinois State University
Jacques Bou Abdo
bouabdjs@ucmail.uc.edu

University of Cincinnati

Research Article

Keywords: AI Agents, AI in Cybersecurity, Technical Issues, Legal Issues, Technical Issues, Ethical Issues

Posted Date: August 5th, 2025

DOI: https://doi.org/10.21203/rs.3.rs-7180712/v1

Additional Declarations: No competing interests reported.

AI AGENTS AND FULL AUTOMATION IN CYBERSECURITY: TECHNICAL, BUSINESS, ETHICAL AND LEGAL CONSIDERATIONS

Abstract

The paper explores the multifaceted impacts of the integration of the increasingly autonomous AI agents into the field of cybersecurity. The paper attempts to address the technical, business, ethical, and legal considerations associated with the increasing autonomy of security tools and solutions. The discussion starts by shedding light on the legal qualification of AI agents, and the paper proposes to distinguish between automated AI and AI agents. The basis of this distinction is legal, especially regarding liability in case of fully autonomous AI agent. A systematic review protocol in accordance with PRISMA guidelines was undertaken to elaborate a comprehensive analysis of the effects of increasing autonomy in AI security agents. Key technical limitations were identified, business considerations were discussed, and ethical and legal challenges were analyzed. Despite awaiting various challenges, literature suggests that full autonomy of AI agents is becoming progressively feasible and offers a competitive advantage to organizations in specialized cybersecurity fields. Finally, the paper strongly advocates for the study of future research gaps to ensure that the integration of agentic AI systems into the field of cybersecurity leads to robust, safe, and effective security solutions.

Keywords: AI Agents, AI in Cybersecurity, Technical Issues, Legal Issues, Technical Issues, Ethical Issues

1 Introduction

The rapid deployment of AI systems poses various economic, ethical, and social concerns. Establishing guardrails for the development of AI becomes much more pressing [1], especially in light of the agentic shift that AI systems are witnessing. The recent emergence of agentic AI systems brings, at the same time, powerful capabilities and complex challenges since AI agents not only generate content, but can take actions on behalf of users and execute tasks autonomously. In this context, the integration of autonomous AI agents into cybersecurity might be well suited since security tools and frameworks might require the instantaneous analysis of vast amounts of data, especially given the increasing rate of cyber incidents [2, 3]. Additionally, offensive AI

agents are likely to enable faster and more sophisticated cyberattacks, making cybersecurity automation a logical choice. However, the integration of fully autonomous AI agents into cybersecurity requires a careful balance between innovative efficiency and robust safety. The appropriate level of automation in organization's cybersecurity function is determined by multiple factors. Such factors include:

- **Technical**: the availability, support and maturity of automation technology are essential technical components in the adoption decision [4].
- **Business**: the opportunity cost, insurance and acquired competitive advantage are essential business components in the adoption decision.
- **Legal**: liability and compliance are essential legal components in the adoption decision.
- **Ethical**: bias and social responsibility are essential ethics components in the adoption decision [5].

This paper explores the technical, business, ethical, and legal considerations that full automation and the integration of autonomous AI agents pose in the field of cybersecurity. It specifically attempts to systematically answer the following Research Questions (RQ):

- **RQ1**: What technical capabilities and architectural components are required for fully autonomous AI agents to independently detect, analyze, and respond to cybersecurity threats without human intervention?
- **RQ2**: What are the ethical and legal challenges related to the use of autonomous AI Agents and the pursuit of full-automation in the field of cybersecurity?
- **RQ3**: Are fully autonomous AI agents in cybersecurity attainable, financially feasible, and a source of competitive advantage?

The remainder of this paper is organized as follows: section 2 sets the background for the paper and discusses to what extent the appellation of AI agents is accurate from a legal perspective. Section 2 also addresses the liability problem of increasing autonomy of AI agents. Section 3 discusses the scope of the paper by distinguishing between automated AI and AI agents, and argues that the autonomy attributed to automated AI systems is different in nature from the autonomy conferred to emerging AI agents. Section 4 defines the methodology used for the systematic review protocol and the selection process of relevant literature. Section 5 presents the results of the comprehensive systematic review. Section 6 discusses technical, business, ethical, and legal considerations of the increasing full automation in cybersecurity. Section 7 concludes the paper.

2 BACKGROUND

2.1 Are AI Agents Agents under U.S. Agency Law?

According to the Third U.S. Restatement of Agency, agency is defined in law as "the fiduciary relationship that arises when one person (a "principal") manifests assent to another person (an "agent") that the agent shall act on the principal's behalf and subject to the principal's control, and the agent manifests assent or otherwise consents

so to act" [6]. Thus, the agency relationship is a contract based on the consent of two parties: a principal and an agent. In the same vein, an AI agent is usually perceived as a bilateral relationship between two parties: the programmer as a sort of principal and the software program as the AI agent. Furthermore, from a legal perspective, the principles of agency are defined based on the interaction between three actors, not two: the principal, the agent, and the third party. In addition to the bilateral agency agreement concluded between the principal and the agent, the agency relationship is tripartite, since the law of agency is also concerned with the effects of the agency on third parties when dealing with the agent of the principal [7]. If both computer science and the law employ the same term "agent," it is significant to examine in this context whether we can qualify AI agents as agents under agency law.

Electronic agents cannot reciprocate consent under agency law:

From the outset, an AI agent cannot legally be qualified as an agent under agency law since a software is not yet a legal person that can provide mutual consent. If the user-principal can manifest its assent to allow a software-agent to act on the principal's behalf and subject to the principal's control, a software-agent cannot reciprocate with consent to a fiduciary relationship. For the purposes of the common law of agency, Section 1.04 of [6] requires that a 'person' must be one who is directly the object of liabilities and the holder of rights. From a legal standpoint, a 'person' can be a natural person or a juridical person which refers to an organization "that has legal capacity to possess rights and incur obligations," (3) a governmental entity, or (4) any other entity that "has legal capacity to possess rights and incur obligations" [8]. Furthermore, the Third U.S. Restatement of Agency explicitly affirms that computer programs qualified as electronic agents are not capable of acting as principals or agents. This position is founded on a formal and substantive legal basis: From a formalist perspective, computer programs cannot be qualified as legal juristic persons since they are not granted legal personhood by law. From a substantive perspective, electronic agents are not capable of acting as agents because they lack independent volition, owe no duties, and are designed as tools for their users. Under American agency law, computer programs are characterized as "consequential instrumentalities" similar to animals or inanimate objects that a person uses to alter her, his, or its own legal rights and obligations [8].

AI agents and the scope of authority granted by the principal under agency law:

In the agency agreement, the principal defines the scope of authority granted to the agent. The agent must act only within the scope of the authority granted by the principal and must abide by all the lawful instructions given by the principal. The scope of authority under agency law poses significant challenges because the instructions given by the principal are often, if not always, ambiguous and subject to interpretation [9]. The common law of agency established the concept of actual authority granted to the agent to recognize the need for constant discretion and interpretation in the performance of agency. The actual authority granted by the principal may be an explicit authority given, for instance, by explicit written instructions. In contrast,

the actual authority could also consist of an implied authority. Since the principal does not, and essentially cannot, detail specifically how to perform and achieve the desired result, the law of agency recognizes the existence of an actual implied authority based on reasonableness. In this regard, the actual authority also includes what the agent reasonably understands to be included in their actual authority at the time of performance [7]. Thus, the scope of the mandate granted to the agent entails a dynamic process that requires a process of interpretation, discernment, and deliberate discretion. In this context, imagine a principal-user prompts its AI agent with the following instruction: 'Deploy 200,000 American dollars in initial capital to generate 1.5 million American dollars in profit within a period of six months by trading digital assets on centralized and decentralized exchanges in the crypto market of the United States'. The previous instruction, which gives little indication of how the AI agent should be generating profit in the crypto market, fleshes out the difficulties related to the definition of the scope of authority given to an AI agent. The emergence of the concept of superalignment, which is based on the direct embedment of explicit long-term goal representations into the decision-making process of an AI agent, can contribute to the technical delimitation of the scope of the actual authority granted to AI agents [10].

AI agents and the duty of loyalty under agency law:

In general, the term "fiduciary" under the Third U.S. Restatement of Agency refers to the duty of loyalty incumbent on the agent when acting on behalf of the principal. Loyalty in this context does not only require mere fairness and honesty. The duty of loyalty obliges the agent to act and further the principal's best interests. Practically, the agent cannot use property or confidential information belonging to the principal for its own benefits or those of a third party. In addition, the agent should be loyal to the principal and cannot serve the interests of an adverse party in a transaction related to the agency, compete directly with the principal, or assist the principal's competitors. Furthermore, the agent cannot pursue or obtain material benefits resulting from transactions related to the agency [9]. In this context, one can investigate to what extent AI agents can be self-interested or disloyal to their principal-user. Some growing research is starting to show that AI systems could be capable of deceptive or fake alignment during training [11], or frontier models are capable of in-context scheming [12]. One of the rationales underlying the fiduciary obligations provided by agency law is to prevent potential conflicts of interest that might be exploited by the agent to the detriment of the principal. In this context, alignment strategies could aim at embedding in AI agents a positive duty of disclosure that would lead to the communication to the principal-user of facts that might lead to conflicted situations prohibited under agency law. Furthermore, the design of AI agents should be tailored to support, for example, a negative duty of confidentiality that would prohibit AI agents from disclosing or sharing personal information against the user's best interests [9].

2.2 Can AI Agents be Held Accountable and Liable for their Actions?

The increasing autonomy of AI agents and the liability problem:

The responsibility related to agentic AI presents novel challenges in comparison with previous technological innovations. In general terms, the risks associated with the technology preceding agentic AI were reasonably anticipated, managed, and decreased with time. The effects that could result from previous technology were always associated with a 'proximate responsibility of individual humans and human institutions.' The current increasing autonomy of AI agents makes less proximate the individual human responsibility and less effective the policy measures based on human intervention and human control. AI agents pose a central legal challenge coined by legal research as 'the liability problem,' which consists of investigating how to continue holding individual humans and human institutions legally liable for the actions of increasingly autonomous AI agents. This legal challenge is intertwined with another ethical one: How to ensure that human agents continue to be morally responsible for the technology they develop and deploy as these technologies become increasingly autonomous [13].

Piercing the 'agentic veil' of autonomous AI agents:

According to [7], the real legal challenge related to the liability problem is to ensure that the companies behind AI agents are responsible for the products and services they develop and offer. The increasing autonomy of AI agents might exacerbate the liability problem. In practice, the more AI actors confer independent agency and technical autonomy on their AI agents, the more AI companies will tend to shield themselves from personal and proximate liability and invoke the separate liability of their autonomous and independent AI agents [7]. A recent case related to the deployment of an AI-based chatbot solution illustrates well this tendency. In fact, Jake Moffatt booked a flight with Air Canada after the death of his grandmother. While searching for tickets, Mr. Moffat conversed with a chatbot on Air Canada's website. The chatbot suggested that Mr. Moffat could apply for the bereavement rate retroactively after traveling, as long as the applicant files a Ticket Refund Application Form within 90 days starting from the ticket issue date. Mr. Moffat relied on this assertion and booked two separate one-way tickets to arrange travel from Vancouver to Toronto. Later, Air Canada employees told Jake that the airline did not allow retroactive applications. Air Canada rejected the retroactive bereavement application claiming that Mr. Moffat did not comply with the proper procedure and did not abide by the contractual provisions provided under its domestic tariff terms and conditions despite the filing of the application well within the 90 days required by the chatbot. One Air Canada representative admitted in one of the exchanged emails that the chatbot provided 'misleading words,' and affirmed that Air Canada had noted the problem and will update the chatbot. The Civil Resolution Tribunal framed the main allegation of the plaintiff as a negligent misrepresentation claim and examined whether Air Canada negligently misrepresented the procedure for claiming bereavement rates when dealing with the plaintiff. The main defense of Air Canada alleged that 'it cannot be held liable for

information provided by one of its agents, servants, or representatives - including a chatbot.' The airline further suggested that 'the chatbot is a separate legal entity that is responsible for its own actions.' *In fine*, the court found that Air Canada did not take reasonable care to ensure the communication of accurate information to the plaintiff, concluded that all the elements of the claim of negligent misrepresentation were established, and awarded damages to Mr. Moffat [14]. Despite being a small claims case, this decision is significant because it presages a risk for AI companies to attempt and 'hide' behind the autonomous liability shield of the AI agents they develop and deploy.

3 SCOPE

Automation was defined by referring to the term "machine agent," where automation would be the execution of some tasks by a machine agent that were previously carried out by a human. Intrusion detection systems and vulnerability assessment tools are depicted as examples of machine agents [15]. More recently, the scientific literature observed a transition from 'simple automation' to automated AI systems [16]. This transition is characterized by two variants that are evolving in opposite directions: An increased autonomy of AI systems that is concomitant with a continuous decrease in direct human control exerted in less controlled environments [13, 17, 18].

3.1 A Spectrum Approach to Agentic AI

The definition of AI agents is often addressed by resorting to a spectrum approach. Instead of proposing a binary definition of AI agents, some research defines an AI agent as a system with a relatively high degree of agency and refers in this context to increasing agentic systems [19, 20]. Chan et al. identified 4 key characteristics that would define the degree of agency in an algorithmic system. First, the AI agent is capable of 'underspecification,' which is the capacity of an AI system to achieve a goal without precise or concrete specifications, that are given by operators or designers, on how the goal should be achieved. The second characteristic contrasts directly with the concept of human in the loop and fleshes out the autonomy of an AI agent. In this regard, "Directness of impact" refers to the degree to which actions taken by an AI system would impact or affect the world "without mediation or intervention by a human." Third, 'Goal directedness' consists of the achievement of a particular quantifiable objective. Lastly, an AI agent is capable of 'long-term planning,' which refers to the ability to make 'temporally' short-term decisions that lead to long-term predictions or goals [19]. Jovanović and Campbell addressed AI agent autonomy by identifying 5 levels of autonomy: the lowest level of autonomy would refer to "Agent Assistance" and consist of AI agents that guide and correct humans during the performance of a task (e.g., assistance in learning a programming language). The second level of autonomy would be a "Task-Specific Autonomy" where AI agents perform specific tasks autonomously under the supervision of humans (e.g., search for information online). The third level of autonomy would refer to 'Supervised Autonomy' where an AI agent would plan and perform actions requiring human approval (e.g., Planning and booking travel online). The fourth level of autonomy would consist of an

"Unsupervised Autonomy" where the AI agent executes autonomously and humans monitor and intervene if needed (e.g., software development and testing). Lastly, an AI agent with "Full Autonomy" executes autonomously under the active monitoring of humans (e.g., runtime infrastructure maintenance) [21].

Resorting to a spectrum approach in defining AI agents would confine the difference between automated AI and AI agents to a difference in degree and intensity. When we apply a spectrum approach to the emergence of AI agents, an AI agent would be perceived as an augmented automated AI system. However, the difference between automated AI and AI agents is not merely variations in the levels of autonomy that would exist on the same spectrum.

3.2 A Difference in Nature: Automated AI vs. AI Agents

Since the nature of autonomy developed in autonomous AI agents is fundamentally different from the nature of autonomy implemented in the context of automated AI, the difference between automated AI and AI agents could be perceived as a difference in nature. In this context, AI agents, unlike automated AI, are capable of reactivity, adaptiveness, collaboration, proactiveness, and, above all, are unpredictable.

AI Agents are capable of reactivity and adaptiveness to unexpected events occurring in an unstructured environment:

AI agents are capable of reactivity to unanticipated stimuli in an open and unpredictable environment. If both automated AI systems and AI Agents are capable of performing complex tasks, automated AI systems execute complex functions in 'structured environments' and have limited ability to learn and adapt to 'unexpected events'. On the other hand, AI agents are characterized by their reactivity to unanticipated stimuli, operating in 'uncertain and unstructured physical and/or information environments, and managing unexpected external or internal events' [8].

AI Agents are proactive, collaborative, and unpredictable:

AI agents are portrayed as autonomous software systems that make decisions and pursue goals with 'creativity and flexibility'. AI agents are 'proactive as they interact with other agents and changes to the world around the agent on behalf of the user' [7]. Such features contrast fundamentally with automated AI systems that are usually restrained to close and structured pre-defined environments. AI agents have also been depicted as 'rebel agents' that are capable of rejection, protestation or developing attitudes of 'reluctance or opposition to goals or courses of action assigned to them by other agents, or the general behavior or attitudes of other agents' [22]. In addition to rebel agents, AI agents have the ability to circumvent security oversight, raising significant multi-agent security challenges such as unintended information sharing or undesirable coordination coined under the term of multi-agent deception [23]. The proactiveness and collaborative ability of AI agents can lead to unpredictable behaviors [17]. The unpredictability of autonomous systems is manifested in their unintended actions and general consequences on society and would result in a liability gap that would 'occur beyond the scope of human responsibility' [13].

This noticeable shift from automated AI into autonomous AI agents will undoubtedly impact the field of cybersecurity. If the training and validation of AI models in cybersecurity were performed in structured environments under traditional automated AI-driven security solutions, the development and deployment of autonomous AI agents capable of detecting, classifying and responding autonomously to security threats in real time pose unexplored challenges.

4 METHODOLOGY

4.1 Systematic Review Protocol

We conducted this systematic review following the PRISMA guidelines. Our methodology details our research sources, selection criteria, data extraction process, and synthesis methods. This approach ensures methodological rigor and accuracy of results.

4.2 Search Strategy

We searched Scopus and ACM Digital Libraries and created search queries that would particularly find research about AI, cybersecurity, autonomous systems, full automation, AI agents, fully autonomous AI agents, cybersecurity automation, AI-driven cybersecurity, Financial feasibility of AI in cybersecurity, Competitive advantage of AI in cybersecurity, Limitations of AI in cybersecurity, and similar topics. In addition, Boolean operators and search filters were employed to refine the results further and ensure that only the most relevant information was returned. The queries encompass the three research questions and meet our criteria. The used queries are listed below in Table 1 based on the format used in each digital library. Following the execution of the queries, we had 240 publications from Scopus DL and 461 publications from ACM DL. Two (2) were identified as duplicates and removed, and 536 were excluded from our research since they did not meet our inclusion criteria. After screening and reading the title, abstract, and keywords of 163 publications, 74 publications were assessed for eligibility and were fully read. Finally, 74 studies were included in our review. After applying these queries respectively in each library we obtained the following results, which are listed in Table 2.

4.3 Study Selection Process

After filtering to exclude papers that do not meet our inclusions, we run another check for paper duplication using Rayyan and Microsoft Excel. Rayyan is a web-based application specifically designed to support systematic review workflows. Developed to assist researchers in managing and screening large volumes of references, Rayyan provides functionalities for blinded, collaborative screening of titles and abstracts, the consistent application of eligibility criteria, and structured conflict resolution among reviewers [24]. Using Rayyan, we imported and organized all retrieved citations enabling multiple reviewers to independently assess each record based on predefined inclusion and exclusion criteria, thereby minimizing bias from other reviewers' decisions. The application's filtering and tagging capabilities facilitate the categorization of

records allowing for the efficient identification of studies that meet our criteria. When discrepancies arose in reviewer decisions, Rayyan's conflict resolution feature provided a structured interface for discussion and adjudication. Through this approach, Rayyan enhanced the transparency, rigor and reproducibility of the study selection process, ensuring that only studies directly aligned with the research objectives were advanced to full-text screening and qualitative synthesis. The final results are listed in Table 3.

Table 1 Search Queries Used in Digital Libraries

Digital Library	Query
Scopus DL	TITLE-ABS-KEY (("AI" AND "Cybersecurity" AND "Autonomous Systems") OR ("AI Agents" AND "Threat Mitigation") OR ("Cybersecurity" AND "Full Automation"))
ACM DL	"query": { ("AI" AND "Cybersecurity" AND "Autonomous Systems") OR ("AI Agents" AND "Threat Mitigation") OR ("Cybersecurity" AND "Full Automation") } "filter": { ACM Content: DL }

Table 2 Initial Search Results from Digital Libraries

Digital Library	Number of Publications
Scopus DL	61
ACM DL	461

Table 3 Filtered Publications after Applying Inclusion Criteria

Digital Library	Number of Filtered Publications
Scopus DL	34
ACM DL	444

We establish explicit guidelines concerning the types of studies considered acceptable. We assess based on criteria such as title and abstract content to establish suitability. We search for technical reports, conference proceedings, and peer-reviewed articles published in the last specific time frame. The papers that satisfy the inclusion criteria detailed in Table 4 are considered for further investigation. In contrast, publications not meeting these criteria are excluded from our analysis by the exclusion criteria outlined in Table 4.

We further conducted detailed screening to ensure the relationship between the primary studies and the three (3) research questions that guided this study. The following Table 5 was obtained as the result of the comparison.

Table 4 Inclusion and Exclusion Criteria

Inclusions	**Exclusions**
Fully Automated AI for Cybersecurity	Partial Automated AI for Cybersecurity
Fully Autonomous AI Systems for Cybersecurity	Partial Autonomous AI Systems for Cybersecurity
Papers written in English	Papers written in languages other than English
Papers published between 2019 and 2025	Papers written before 2019
Journals and Conference Proceedings	Books, Magazines, Reports, and Reviews

Table 5 Distribution of Publications by Research Question

Research Question	**Number of Publications**
RQ 1	65 [21, 25–89]
RQ 2	27 [27, 36, 39–41, 50, 52, 56–58, 61, 66, 68, 70–73, 75, 83, 87–94]
RQ 3	55 [21, 25–34, 36–38, 41–55, 59–69, 72–74, 76–82, 84–89, 95, 96]

The above steps are summarized in the below PRISMA flow diagram, which is listed in Figure 1, shows the steps that we have followed, and the number of publications at each step.

4.4 Data Extraction and Analysis

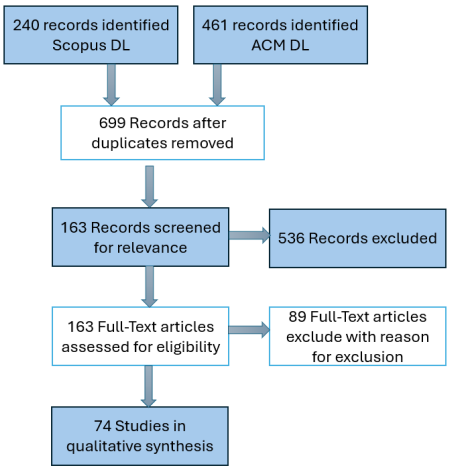

Fig. 1 PRISMA flow diagram representing the selection process of primary studies.

5 RESULTS

This section presents the results of a comprehensive systematic literature review conducted on 74 peer-reviewed publications concerning the development and deployment of fully autonomous AI agents in cybersecurity. The analysis employed a combination of thematic synthesis, open coding, and bibliometric visualization to explore the technical, ethical, and feasibility aspects of cybersecurity automation. The results are structured to provide in-depth responses to the three guiding research questions, with findings organized into logically progressive subsections that trace key themes, gaps, and emergent trends across the corpus.

5.1 Technical Limitations and Required Innovations (RQ1)

The literature indicates that while progress has been made in AI-driven cybersecurity, achieving full autonomy remains a complex and unresolved challenge [55, 83]. The most prominent limitation across the reviewed studies is the incomplete implementation of autonomous capabilities in threat detection, analysis, and response. Many systems today continue to rely heavily on human oversight, particularly in the final decision-making phases. Although models based on supervised learning, reinforcement learning, and deep neural networks have shown strong performance in detecting known threats, they consistently struggle to generalize to novel attack vectors, particularly zero-day exploits [60]. Several studies identified that technical failures are frequently linked to overfitting and lack of contextual awareness in AI models [62, 97]. For example, autonomous systems trained on specific datasets often perform poorly when confronted with unfamiliar environments. Moreover, real-time responsiveness a critical

requirement in cybersecurity remains inadequate due to high computational latency and limitations in current hardware configurations. These challenges are exacerbated in dynamic environments such as cloud computing, smart cities, and industrial IoT, where decision latency can have catastrophic consequences [87]. Architectural shortcomings further compound these technical constraints as the current literature reveals a marked absence of standardized and modular architectures that support seamless orchestration across distributed environments [85]. Most existing frameworks are vertically integrated and difficult to scale or adapt to evolving security needs. The lack of interoperability between AI modules, data sources, and legacy systems significantly hampers system resilience and adaptability. The thematic co-occurrence network graph in Figure 2 illustrates that while terms such as "orchestration," "learning models," and "threat detection" are frequently co-occurring, the studies rarely converge on a unified design paradigm. To overcome the limitations the literature proposes some innovations. First, federated learning architectures are recommended to enable collaborative defense mechanisms without compromising data privacy [65]. These systems can adapt to distributed environments by continuously updating models across decentralized nodes. Second, explainable AI (XAI) is identified as a necessary complement to autonomy, enhancing system transparency and aiding in post-incident forensics [70]. Third, the integration of real-time, continuously evolving knowledge graphs is posited as a means to enhance situational awareness and reduce false positives. These innovations while still largely theoretical, are increasingly being referenced in the recent corpus indicating a shift in research priorities. Figure 3 presents the top 20 keywords by frequency, highlighting the dominance of terms such as "threat detection," "automation," "risk" and "compliance." The prominence of these terms underscores the literature's concentrated effort on technical resilience and systemic adaptability. Even though the technical autonomy remains aspirational the reviewed studies collectively chart a roadmap of innovations needed to close the gap between current capabilities and future requirements.

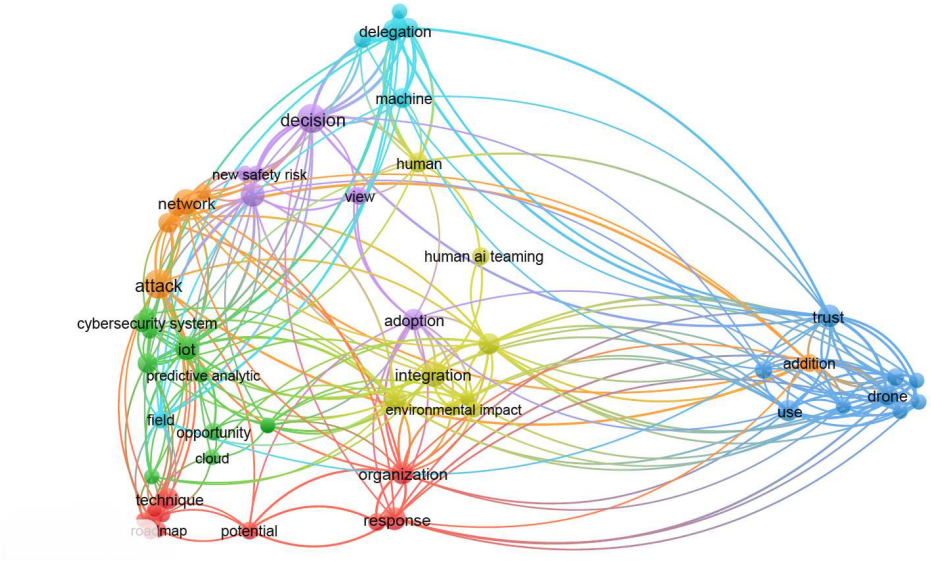

Fig. 2 Keyword Co-occurrence Network Graph

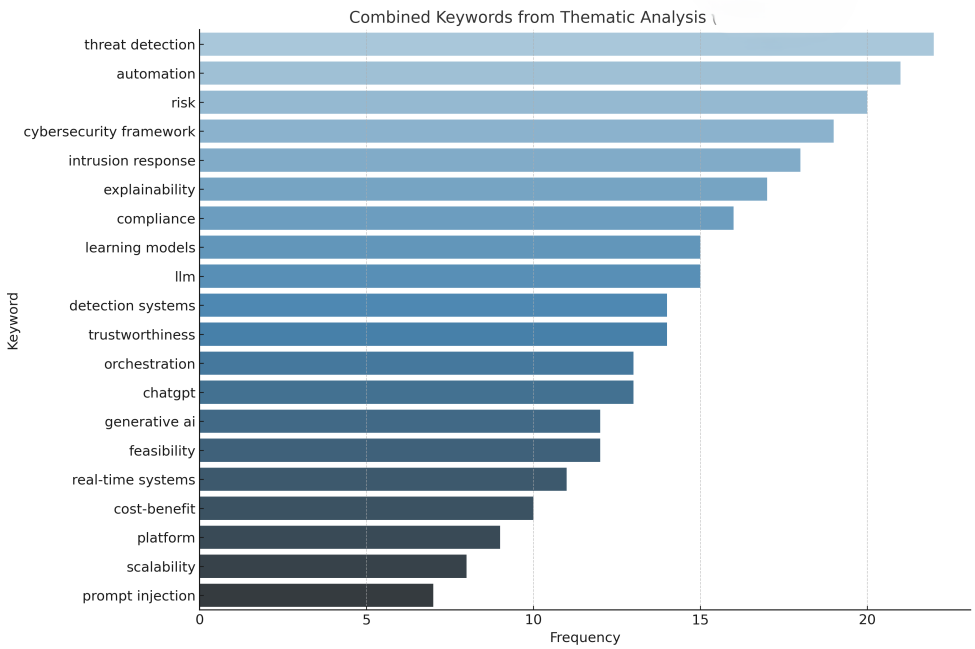

Fig. 3 Combined Keywords from Thematic Analysis

5.2 Ethical and Legal Challenges (RQ2)

Beyond technical limitations the pursuit of fully autonomous AI agents in cybersecurity is accompanied by substantial ethical and legal challenges. A recurring theme in the literature is the ambiguity surrounding responsibility and accountability in autonomous decision-making [57, 95]. In systems where AI agents operate without direct human oversight, the assignment of fault in cases of error, harm, or breach becomes legally and ethically contentious. This is particularly relevant in sectors such as finance, defense and healthcare where cybersecurity failures can have far-reaching consequences. Several studies have raised concerns about the opacity of AI decision-making processes particularly when deep learning models are involved [70, 75]. The lack of explainability in AI outputs complicates efforts to audit decisions, trace the root cause of errors, and establish legal liability. Although explainable AI (XAI) methods are being developed, many fall short of producing human-interpretable justifications that meet regulatory standards. This gap between technological capability and legal requirements presents a formidable barrier to widespread adoption. Bias and fairness also feature prominently in the reviewed literature. Autonomous systems trained on historical cybersecurity incident data may inadvertently replicate human biases or systemic flaws, leading to discriminatory outcomes [94]. For instance, certain network behaviors associated with benign user activity may be misclassified as malicious if the training data lacks diversity. These issues are not merely theoretical; real-world cases of algorithmic bias in cybersecurity decision-making are beginning to emerge, prompting calls for more inclusive and representative training datasets [82]. On the legal front the review uncovered a fragmented regulatory landscape. Jurisdictional differences in data privacy laws, AI governance and cybersecurity policies complicate the deployment of fully autonomous agents, particularly in cross-border scenarios [85]. The absence of harmonized standards inhibits the development of universally accepted compliance protocols. Terms such as "trustworthiness," "liability" and "compliance," appear frequently in the thematic keyword visualization in Figure 3 highlighting a preoccupation in the literature with these regulatory and ethical issues. The studies underscore the need for developing comprehensive legal frameworks that can address the unique challenges posed by autonomous agents. Such frameworks should address not only accountability and transparency but also ethical norms concerning user consent, surveillance, and the potential misuse of AI in offensive cyber operations. Until such frameworks are established the deployment of fully autonomous AI systems in critical infrastructures will likely remain constrained.

5.3 Feasibility, Financial Viability, and Competitive Advantage (RQ3)

Despite the challenges outlined above, the literature suggests that fully autonomous AI agents are gradually becoming feasible, particularly in specialized cybersecurity domains [21, 76]. The publication trend presented in Figure 4 shows a notable increase in research outputs from 2022 onward with a peak in 2024. The surge aligns with the proliferation of large language models (LLMs), reinforcement learning systems and

adaptive threat intelligence platforms signaling a growing confidence in the attainability of autonomy. Multiple studies report partial deployments of autonomous agents in areas such as cloud workload security, malware classification, and automated patch generation [26, 32]. These systems, while not entirely human-free, operate with minimal supervision and have demonstrated measurable gains in efficiency and detection accuracy. For instance, organizations deploying AI-assisted Security Operations Centers (SOCs) report reduced mean time to detect (MTTD) and mean time to respond (MTTR), key metrics in cybersecurity performance. However, the cost of achieving full autonomy remains a significant barrier. High-performance computational infrastructure, model training, and updating costs, as well as the complexity of maintaining secure and interpretable models, contribute to substantial capital and operational expenses [95]. The reviewed literature notes that while larger enterprises can absorb these costs and reap long-term benefits, smaller organizations often lack the necessary resources to do so. This financial disparity may contribute to an uneven distribution of AI benefits across the industry. From a strategic perspective several papers argue that autonomous AI systems confer a competitive advantage. Organizations that successfully integrate such systems often experience enhanced operational agility, improved compliance readiness and differentiation in cybersecurity capability [60, 83]. These advantages are particularly pronounced in sectors with high regulatory burdens and real-time operational demands. The geographic distribution of publications, illustrated in Figure 5, indicates that research activity is concentrated in technologically advanced countries, including the United States, China, the United Kingdom, and Germany. This suggests that national investment in AI and cybersecurity infrastructure plays a critical role in determining feasibility and pace of adoption. Overall, the literature reflects cautious optimism regarding the future of fully autonomous AI agents in cybersecurity. While challenges remain, particularly in ensuring affordability and fairness the trajectory of current research suggests that autonomy is both attainable and strategically advantageous under the right conditions.

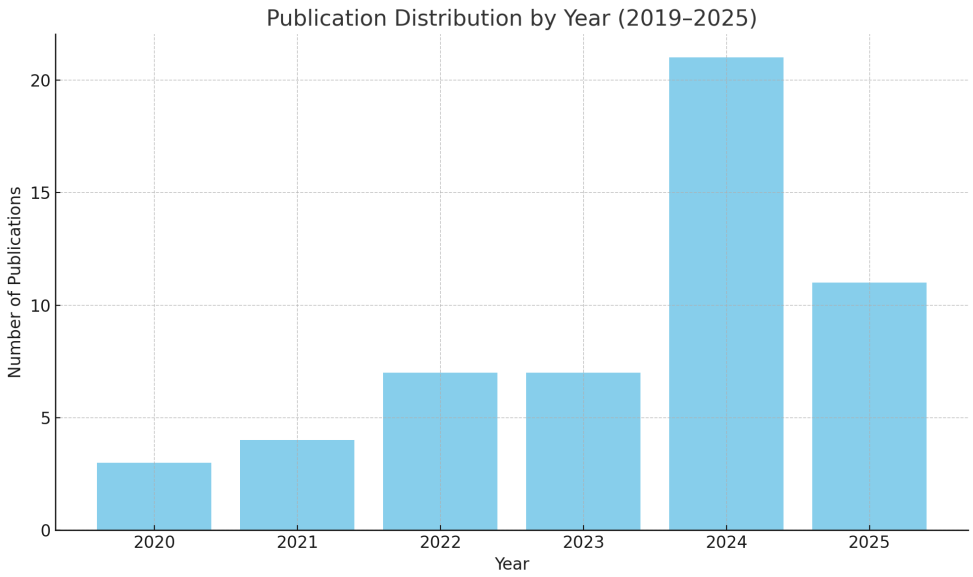

Fig. 4 Publication Distribution by Year (2019-2025)

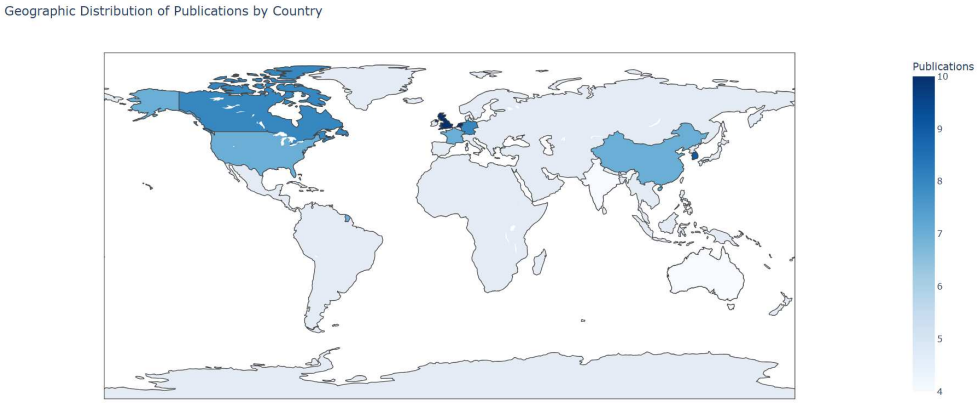

Fig. 5 Geographic Distribution of Publications by Country

6 ANALYSIS AND DISCUSSIONS

6.1 Technical

The integration of autonomous AI agents into cybersecurity is no longer a theoretical aspiration but a rapidly evolving reality. Across the 74 included studies, several themes emerged regarding how these systems are being implemented, the challenges they face and the benefits they bring. This section synthesizes findings on architectures, methodologies, automation techniques, technical benefits, and the key barriers to adoption.

6.1.1 Architectures and Automation Frameworks

A number of papers proposed comprehensive frameworks or architectures for integrating fully autonomous AI into cybersecurity operations. [25] presented a modular architecture capable of full automation in threat detection, information collection, and dynamic response adjustments. Their work emphasized adaptive security levels based on real-time conditions across distributed systems. In [26], the authors delivered a realistic environment for reinforcement learning agents in pentesting, replacing unrealistic simulation-based environments and enabling agents to autonomously learn and exploit network weaknesses. The A2C Framework introduced by [28] stands out for enabling dynamic transitions between automated, augmented, and collaborative decision-making in SOCs. It demonstrated a 59% improvement in intrusion detection when collaborative human-AI exploration was added to automation. These systems embody the principles of autonomous cybersecurity by minimizing the need for human oversight and allowing real-time adaptive defense strategies.

6.1.2 Technical Benefits of Fully Autonomous AI Agents

Across the studies multiple technical advantages of automation were observed. Speed and scalability emerged as critical benefits, with [34] demonstrating that AI agents in cloud environments could detect and mitigate threats 6x faster than traditional methods, supporting SRE-aligned metrics like Mean Time to Detect (MTTD) and Resolve (MTTR). Enhanced detection and coverage were evident in tools like PenHeal [47] and Acquirer [64], which improved vulnerability coverage and detection rates using LLMs and selective path exploration, revealing 11 previously unknown vulnerabilities. Resource efficiency was another advantage, with time-series-based distributed AI models for IoT [31] reducing overhead by detecting and mitigating threats locally and autonomously, thus preserving bandwidth and power across devices. Adaptability to complex environments was demonstrated by smart fuzzing [65] and DRL-based privilege escalation models [62], which showcased how AI agents could autonomously test system weaknesses and adapt behavior in unpredictable, real-world cyber-physical systems.

6.1.3 Technical Challenges and Limitations

Despite these promising outcomes several limitations and obstacles were consistently identified. The complexity of environments presented significant challenges with many

autonomous agents struggling to handle dynamic, multi-layered environments such as IoBT and autonomous vehicles. Intricacies in protocols and hybrid architectures [91] [27] introduce challenges in generalizability. Data constraints constitute a major bottleneck, particularly access to high-quality, labeled data for training. Some environments lack realistic datasets, as noted by [26], which can reduce the real-world effectiveness of trained agents. Lack of explainability and trust remains problematic; while autonomous systems can act rapidly they often lack transparency in decision-making processes. This impedes trust and adoption particularly in sensitive domains like critical infrastructure and the military. Interoperability and integration challenges also persist, with the incorporation of autonomous agents into legacy systems or across distributed platforms presenting non-trivial issues related to standardization, security, and compatibility [37].

6.1.4 Evaluation Practices

A diverse range of datasets and evaluation strategies were adopted across studies. Simulated environments including NSL-KDD, KDDCup, and Caldera were commonly used for reinforcement learning and threat detection. Real-world testbeds such as PenGym and smart fuzzing frameworks employed physical cyber-physical testbeds to test agent performance in operational settings. Hybrid metrics were frequently employed, with many papers moving beyond accuracy to use robustness, speed (TTD, TTM), coverage, and adaptability as critical evaluation criteria. This highlights an increasing sophistication in how AI-based cybersecurity systems are tested, though standardized benchmarks are still needed.

6.1.5 Research Gaps and Future Directions

Several under-explored areas were highlighted in the literature. Human-AI interaction remains limited in terms of trust calibration and conflict resolution between humans and autonomous agents, with A2C representing a promising exception. Cross-domain intelligence sharing, as studied by [46], points to privacy-preserving collaborative frameworks as an essential but largely untapped domain. Self-healing and long-term autonomy represent another gap; while self-healing systems were proposed [37], few works operationalize lifelong learning in AI agents for cybersecurity. Benchmarking autonomy levels presents another challenge as not all studies clearly delineate whether the AI systems are fully autonomous or simply augmented tools. There is a need for a classification framework to guide future evaluation and policy.

6.2 Business Perspective

6.2.1 Operational Efficiency:

Fully autonomous AI agents are improving operational efficiency, by eliminating delays, reducing the dependency on manual processes, and enabling round-the-clock task execution [68]. In sectors like cybersecurity, these agents can detect and respond to threats in real-time, significantly reducing incident response time compared to human analysts [71]. For instance, AI-powered threat detection systems can process large-scale

network data in real time, identify threats like advanced persistent threats (APTs) and zero-day vulnerabilities with her accuracy than traditional systems [67]. These fully autonomous agents execute defensive responses without any human input, drastically reducing response time and minimizing system downtime. Furthermore, [70] emphasize how digital twins supported by AI agents enable proactive system monitoring and predictive maintenance, improving system uptime, and reducing costs. Furthermore, by integrating AI into digital ecosystems, businesses can automate data collection, pattern recognition, and decision making [73], which will increase throughput and reduce waste in production or information workflows [71]. This level of automation allows firms to scale operations without proportionally increasing labor costs, achieving more with less, and improving operational margins.

6.2.2 Strategic Agility:

Strategic agility, or the ability to rapidly adapt to changing environments, is highly improved by fully autonomous AI agents' ability to make rapid, data-driven decisions in dynamically changing environments [21]. These agents can constantly analyze environmental variables, consumer behavior, and competitor strategies to suggest or implement adjustments in real time. For example, in cybersecurity, fully automated AI agents enable proactive defense strategies through predictive analytics and pattern recognition. This anticipates potential threats before they manifest, allowing organizations to shift from reactive to preventive strategies [69]. In business scenarios, these agents continuously scan the market, competitors, and operational signals, enabling dynamic strategy realignment [72]. Moreover, in broader business contexts, agents can autonomously adjust marketing strategies, supply chain configurations, or customer support flows without waiting for executive human intervention. This agile behavior enables companies to seize emerging opportunities faster and mitigate risks before they escalate, offering a sustained competitive edge in volatile markets. [21] explains that self-directing AI agents adapt to new goals based on environmental cues without explicit reprogramming, which allows firms to pivot rapidly during crises or when seizing emerging opportunities. This real-time responsiveness is crucial for firms operating in fast-paced digital markets and volatile geopolitical environments.

6.2.3 Innovation:

Fully autonomous AI agents act as innovation accelerators by uncovering novel solutions, automating experimentation, and exploring large-scale simulation scenarios that humans alone would struggle to manage [21]. They are unlocking new frontiers of innovation by autonomously exploring solutions spaces beyond the human imagination. According to [21], self-directing agents exhibit emergent behaviors, generating new workflows, adapting to novel conditions, and independently designing or modifying systems. In cybersecurity specifically, fully automated AI agents simulate cyberattack scenarios, learn from adversarial behavior, and suggest defensive strategies that human experts may not conceive [69]. Across most sectors, agents are utilized in creative tasks such as content generation, product design, and data exploration. By integrating domain-specific knowledge with reinforcement learning, agents not only automate known processes, but also discover unconventional approaches and optimize

them through feedback [95]. This continuous learning cycle transforms agents from tools into co-innovators.

6.2.4 Risks and Control:

Nevertheless, despite their benefits, fully autonomous AI agents present complex risks that require advanced governance frameworks [66]. Risks range from technical errors and unintended actions to ethical issues such as bias or privacy violations. Plus, the potential for adversarial attacks, where AI systems are misled through manipulated inputs, is a serious concern according to many authors ([67], [72]). These vulnerabilities could allow malicious actors to bypass or disable defense mechanisms. Privacy issues are also paramount, especially with agents accessing sensitive personal or organizational data. Explainability, or the ability to understand and justify agent decisions, becomes vital in high-stakes domains such as finance, law, or healthcare [66], [98]. In addition, as AI agents assume more decision-making roles, the questions of accountability, legal liability, and regulatory compliance intensify. Companies must design control mechanisms that allow human oversight, even in environments where agents operate independently. This includes traceability logs, intervention points, and built-in ethical constraints. Failing to implement such safeguards not only exposes firms to legal consequences and risks to companies' reputations, but also undermines stakeholder confidence in AI-driven systems.

6.2.5 Barriers to Entry:

The deployment of fully autonomous AI systems sets challenging barriers to entry for smaller companies as it requires significant investments in computing infrastructure, high-quality data pipelines, and skilled personnel [99]. [67] note that effective AI integration in cybersecurity depends on the ability to collect and analyze vast and high-velocity datasets. [72] also highlight the high computational and energy demands of AI systems, particularly those involving deep learning or large models. This infrastructure burden means that large corporations or early adopters with established AI operations gain compounding advantages over time, monopolizing access to quality data, refining agent behaviors faster, and benefiting from economies of scale. Smaller companies often lack the resources to implement or sustain such systems, potentially widening the digital divide.

6.3 Ethical Perspective

6.3.1 Privacy and Bias:

In general, AI enables organizations to reduce security breach losses and to better manage and deal with data breaches [96]. More specifically, generative AI facilitates the automation of security tasks that would reduce the time for recovery and limit the damage caused by successful attacks. In contrast, AI allows attackers to create new forms of malware, phishing, and ransomware at the same time, which could result in further privacy exposure [78]. [61] notes in this context that the advent of AI tools reinforces the asymmetry in cybersecurity efforts. For instance, the ability to generate

phishing context is increased at an unprecedented scale while adversaries need only one single breach to entail tremendous reputational and financial harm. Furthermore, generative AI training in the field of cybersecurity requires the use of large datasets. *Ex-ante*, the collection and processing of personal data for training purposes should be lawful and compliant with the applicable data protection and privacy laws. *Ex post*, data biases in the datasets could lead to biased findings, leading to security measures that might show bias or injustice [78].

6.3.2 Sustainability and Social Responsibility:

[76] proposes a model for the integration of AI and Robotic Process Automation emphasizing not only on operational efficiency but also on social and environmental responsibility. With the increase of the use of autonomous AI in cybersecurity on a large scale, an ethical challenge would be centered on sustainability and the need to balance the deployment of autonomous AI agents with social and environmental responsibility.

6.3.3 The Automation of Cybersecurity Roles and the Impact on Security Job Market:

[15] identifies 4 variables that would determine the extent to which cybersecurity positions and roles can be fully automated. The lower the requirements for creativity, social interaction, physical work and the higher the quality of a sufficient quantity of existing statistical data, the easier it would be to fully automate cybersecurity roles. The criterion of creativity would be significant when the security task requires an alternative course of action that no one has previously identified and applied. Second, it would be difficult to automate a security task that would require situation-adapted or dynamic interpersonal communication with nuances. Third, it would be easy to automate a security task that can be performed completely within a computer and that does not require physical work. Fourth, if it is difficult to produce security data to be used for machine learning training, it would be difficult to automate a security task since there would be very few cases to learn from. Furthermore, [15] found that it would be easier to automate technical or more practically oriented security roles such as database administrators, data analysts, and network operators. In contrast, security positions that would require more formal responsibility would be more difficult to automate, such as executive directors or legal advisors. In the same vein, referring to the NICE categories, tasks related to the 'Investigate' category would be more suitable for automation than security roles that would involve executive leadership or judgmental considerations. In this context, [15] points out that, despite the possible technical automation of cybersecurity tasks, ethical and legal considerations should not allow such automation. Even if the majority of cybersecurity cases are likely not to raise ethical and legal issues (e.g., automated systems scanning employees' emails for malware), some cases would probably not be fully automated. For instance, ethical and legal considerations would not allow the full automation of a security investigation after the abusive use of an organization's IT infrastructure and resources. In the same vein, ethical and legal considerations would not allow the autonomous and full delegation

to a machine of the decision-making process related to filing formal criminal charges or firing an employee. [15]also speculates that ethical challenges would be less critical in automating security tasks related to the construction of new systems (e.g., systems development) compared to tasks that would require the processing of sensitive data found in operational systems (e.g. collection operations). According to [15], the impact on the security job market would be *prima facie* limited since full automation is likely to impact specific tasks within a security role rather than the entire role in itself.

6.3.4 Ethical Challenges of Autonomous Cybersecurity Data Extraction Tools:

According to [92], the proliferation of development security operations organizations that securitize databases, applications, and administrative systems by relying on the extractive power and speed of automated algorithmic programs, which run independently and continuously, poses new ethical challenges. [92] refers to the term "authoritarian conundrum" to encapsulate the moral dilemmas cybersecurity experts encounter as they loosen the leeway of ethical reckoning and argues that the automation of data extraction in Open-Source Intelligence and the Development Security Operations via automated tools will exacerbate the ethically dubious methods used by security practitioners to justify data extraction. The use of autonomous data extraction AI systems would blur the distinctions between those who abuse data and those claiming to have increased security online. Ethical decision-making is often suspended when security operatives are working on keeping their organization safe. In this context, the "expert superiority" of security practitioners makes recommendations on threat prevention into authoritative assertions, and offensive security methods are implemented by security analysts from a position of power, which, in consequence, entail little ethical accountability because actions are rarely audited or scrutinized by superiors. Employing automated AI systems to survey and monitor the web wider and longer would minimize the need for real-time auditing and ethical decision-making. The ubiquitous use of autonomous monitoring tools of the web would allow all actors, including the good and well-intentioned, to repurpose the data extracted autonomously without having to adhere to clear, binding ethical frameworks [92].

6.4 Legal Perspective

6.4.1 The Important Role of Law in Cybersecurity Automation:

[15] notes that the policy orientations that underlie laws and ethics can significantly impact the considerations for the automation of cybersecurity roles. Laws can regulate the availability of cybersecurity data, systems development and the alignment of automated AI solutions. For instance, when the law requires the mandatory reporting and the disclosure of cybersecurity related incidents, such legal requirements would enhance threat intelligence and enable the extraction of precious knowledge to enhance cybersecurity efforts. Conversely, [15] claims that more restrictive laws related to personal integrity would make the collection and the processing of large datasets significantly harder, making more challenging machine learning for security purposes.

6.4.2 Compliance and Cybersecurity Automation:

[76] stresses the importance of compliance with data privacy regulations with the increasing digitalization of processes within organizations. When assessing risk management in robotic process automation in general, [80] suggests regular compliance audits and robust documentation as means to mitigate regulatory and compliance risks. A robust documentation consists of documenting the automated processes and data management, which would strengthen and facilitate the proof of compliance and decrease legal risks [80]. Autonomous AI systems enable the automation of cybersecurity compliance. In this context, [82] proposes an Expert-System Automated Security Compliance Framework, which allows autonomous security compliance in retesting and automates repeated segments in Security Compliance. This automated framework would optimize penetrating testing while being self-sufficient and capable of self-learning. It safely replaces human expert intervention and, more significantly, makes security compliance practice accessible to non-experts [82]. To ensure compliance with laws and regulations, [83] states that security log analysis can help identify potential threats and track users' behaviors. Generative AI can be employed to automate burdensome log analysis and be used to identify anomalous patterns in log data. In this context, the AI automation of log analysis contributes directly to strengthen compliance with security management and requirements. In contrast, security logs data can contain information that would qualify under data protection and privacy laws as personal or sensitive data. Thus, processing log data through automated AI systems would require lawful compliance with privacy laws [83]. Cybersecurity regulations in a specific industry can be implemented by resorting to the enactment of either binding laws or soft law guiding instruments. [87] investigates cybersecurity regulations affecting connected automated vehicles (CAV), and more specifically automated city shuttles (ACS). [87] differentiates between hard law instruments such as the European NIS Directive 1 (2016) and (2020), which broadly apply to all IT fields, and soft law instruments such as the guidelines issued jointly by ENISA and the JRC addressing cybersecurity challenges in the uptake of AI in autonomous driving (2021). This dichotomy between hard law and soft law can also be applied to the field of security regulation of AI autonomous agents. Future research directions would assess the best legal instrument that would effectively regulate the security of autonomous AI agents while considering the pervasive, versatile, and constantly adaptive nature of AI agents' cybersecurity challenges.

6.4.3 Future Research Directions in Cybersecurity Law:

Significant research gaps exist in the legal literature. Future research directions can explore the various effects autonomous AI agents would have on cybersecurity law. For instance, it is significant to study how AI agents will affect the prosecution and the litigation of cybercrimes. Future research can explore the impact of full autonomy on data tampering, chain of custody, and the admissibility of electronic evidence. The basic elements of cybercrimes would be impacted by the deployment of agentic AI. In this regard, if law enforcement succeeded in evidencing the material criminal act committed by an AI security agent (i.e., actus reus), the guilty state of mind (i.e.,

mens rea) should be redefined when an AI agent acts autonomously and on its own motion to commit a cyber act reprehended by law. In this context, the traditional criminal intent departs significantly and does not fit with the new reality of agentic AI. In addition, the application of the requirement of causation might be challenging when law enforcement is looking to impede human actors who developed and deployed the AI security agent that executed the crime.

Future legal research can also explore further the impact of full autonomy on cyber torts and suggest solutions to solve the liability gaps at individual and organizational levels. In case of a failure in full automation, who would be legally liable? The question of legal liability is crucial in the context of cybersecurity since the financial, reputational, and moral harm can be significant. In this regard, a future direction of research is the evolving nature of cyber risk and its insurance. Furthermore, the implementation of full automation in certain domains of cybersecurity will undoubtedly impact the security frameworks related to Governance, Risk, and Compliance. Future research can explore how the different security frameworks of cybersecurity are impacted by the deployment of autonomous AI agents and how these frameworks can serve as effective safeguards to the inherent risks of full autonomy.

7 CONCLUSION

The comprehensive systematic review of the literature showed that achieving full autonomy in AI-driven cybersecurity remains a complex and evolving challenge. Although significant progress has been made in automated cybersecurity solutions and have been deployed, the implementation of fully autonomous capabilities in threat detection, analysis, and response remains to be seen. In addition to the technical impediments of the pursuit of full autonomy of AI security agents, ethical and legal considerations can build a strong case against the full autonomy of AI agents in cybersecurity where human oversight remains fundamental. Yet, some reviewed literature suggests that full autonomy in cybersecurity solutions is progressively becoming feasible and possible, especially in specialized cybersecurity domains. This complete autonomy would confer a competitive advantage to the organization that successfully integrates autonomous AI security agents. Even if full autonomy could be attainable and desirable, the road to full autonomy of AI security agents is undoubtedly paved with considerable technical, business, ethical, and legal challenges that capable organizations cannot willfully ignore.

The agentic shift in AI requires a continuous need to deploy security solutions that address the constantly changing nature of cyber threats in evolving and unpredictable digital environments. Future research should attempt to close the current identified gaps. Furthermore, future technical, business, ethical and legal research should constantly reassess the development and deployment of increasingly autonomous AI security agents to ensure that AI-driven security tools offer robust, safe and effective solutions capable of upholding high security standards. These security standards are fundamental to maintaining, cultivating, and reinforcing public trust in the emerging agentic AI systems.

Funding Statement

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

[1] Mulahuwaish, A., El-Khoury, M., Qolomany, B., Abdo, J.B., Zeadally, S.: Does ai need guardrails? International Journal of Pervasive Computing and Communications **21**(2), 177–186 (2025)

[2] Falowo, O.I., Bou Abdo, J.: 2019–2023 in review: Projecting ddos threats with arima and ets forecasting techniques. IEEE Access **12**, 26759–26772 (2024)

[3] Falowo, O.I., Ozer, M., Li, C., Bou Abdo, J.: Evolving malware and ddos attacks: Decadal longitudinal study. IEEE Access **12**, 39221–39237 (2024)

[4] Ray, G., Muhanna, W.A., Barney, J.B.: Competing with it: The role of shared it-business understanding. Communications of the ACM **50**(12), 87–91 (2007)

[5] Nehme, E., El Sibai, R., Bou Abdo, J., Taylor, A.R., Demerjian, J.: Converged ai, iot, and blockchain technologies: a conceptual ethics framework. AI and Ethics **2**(1), 129–143 (2022)

[6] Restatement (Third) of Agency: Restatement (Third) of Agency. § 1.01 (Am. Law Inst. 2006) (2006)

[7] Desai, D.R., Riedl, M.: Responsible ai agents. Research Paper 5147666, Georgia Tech Scheller College of Business (February 2025)

[8] Powell, D.: Autonomous systems as legal agents: Directly by the recognition of personhood or indirectly by the alchemy of algorithmic entities. Duke Law & Technology Review **18**, 307–331 (2019-2020)

[9] Kolt, N.: Governing ai agents. Notre Dame Law Review **101** (2025). Forthcoming

[10] Liu, B., Li, X., Zhang, J., Wang, J., He, T., Hong, S., Liu, H., Zhang, S., Song, K., Zhu, K., Cheng, Y., Wang, S., Wang, X., Luo, Y., Jin, H., Zhang, P., Liu, O., Chen, J., Zhang, H., Yu, Z., Shi, H., Li, B., Wu, D., Teng, F., Jia, X., Xu, J., Xiang, J., Lin, Y., Liu, T., Liu, T., Su, Y., Sun, H., Berseth, G., Nie, J., Foster, I., Ward, L., Wu, Q., Gu, Y., Zhuge, M., Tang, X., Wang, H., You, J., Wang, C., Pei, J., Yang, Q., Qi, X., Wu, C.: Advances and Challenges in Foundation Agents: From Brain-Inspired Intelligence to Evolutionary, Collaborative, and Safe Systems (2025). https://arxiv.org/abs/2504.01990

[11] Greenblatt, R., Denison, C., Wright, B., Roger, F., MacDiarmid, M., Marks, S., Treutlein, J., Belonax, T., Chen, J., Duvenaud, D., Khan, A., Michael, J.,

Mindermann, S., Perez, E., Petrini, L., Uesato, J., Kaplan, J., Shlegeris, B., Bowman, S.R., Hubinger, E.: Alignment faking in large language models (2024). https://arxiv.org/abs/2412.14093

[12] Meinke, A., Schoen, B., Scheurer, J., Balesni, M., Shah, R., Hobbhahn, M.: Frontier Models are Capable of In-context Scheming (2025). https://arxiv.org/abs/2412.04984

[13] Asaro, P.M.: The liability problem for autonomous artificial agents. In: Papers from the 2016 AAAI Spring Symposium: Ethical and Moral Considerations in Non-Human Agents. AAAI Press, Palo Alto, California (2016). https://www.aaai.org/Library/Symposia/Spring/ss16-04.php

[14] Civil Resolution Tribunal: Moffatt v. Air Canada, 2024 BCCRT 149, §§ 2–4, 11, 15, 27, 32, 44 (Civil Resolution Tribunal, Feb. 14, 2024). Retrieved from https://decisions.civilresolutionbc.ca/crt/crtd/en/525448/1/document.do (2024)

[15] Varga, S., Sommestad, T., Brynielsson, J.: Automation of cybersecurity work. In: Sipola, T., Kokkonen, T., Karjalainen, M. (eds.) Artificial Intelligence and Cybersecurity. Springer, ??? (2023)

[16] Bharti, M.: Ai agents: A systematic review of architectures, components, and evolutionary trajectories in autonomous digital systems. International Journal of Computer Engineering and Technology (IJCET) **15**(6), 809–820 (2025)

[17] List, C.: Group agency and artificial intelligence. Philosophy and Technology (4), 1–30 (2021)

[18] Wellman, M.P., Rajan, U.: Ethical issues for autonomous trading agents. Minds Mach. **27**(4), 609–624 (2017)

[19] Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., Weller, A., Krueger, D., Maharaj, T.: Harms from increasingly agentic algorithmic systems. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. FAccT '23, pp. 651–666. Association for Computing Machinery, New York, NY, USA (2023)

[20] Kapoor, S., Stroebl, B., Siegel, Z.S., Nadgir, N., Narayanan, A.: AI Agents That Matter (2024). https://arxiv.org/abs/2407.01502

[21] Jovanović, M., Campbell, M.: Self-directing ai: The road to fully autonomous ai agents. Computer **58**(2), 71–77 (2025)

[22] Coman, A., Aha, D.W.: Ai rebel agents. AI Mag. **39**(3), 16–26 (2018)

[23] Motwani, S.R., Baranchuk, M., Strohmeier, M., Bolina, V., Torr, P.H.S., Hammond, L., Witt, C.S.: Secret Collusion among Generative AI Agents: Multi-Agent Deception via Steganography (2025). https://arxiv.org/abs/2402.07510

[24] Ouzzani, M., Hammady, H., Fedorowicz, Z., Elmagarmid, A.: Rayyan—a web and mobile app for systematic reviews. Systematic reviews **5**(1), 210 (2016)

[25] Repetto, M., Striccoli, D., Piro, G., Carrega, A., Boggia, G., Bolla, R.: An autonomous cybersecurity framework for next-generation digital service chains. Journal of Network and Systems Management **29**(4), 37 (2021)

[26] Nguyen, H.P.T., Hasegawa, K., Fukushima, K., Beuran, R.: Pengym: Realistic training environment for reinforcement learning pentesting agents. Computers & Security **148**, 104140 (2025)

[27] Albshaier, L., Almarri, S., Albuali, A.: Federated learning for cloud and edge security: A systematic review of challenges and ai opportunities. Electronics **14**(5), 1019 (2025)

[28] Tariq, S., Chhetri, M.B., Nepal, S., Paris, C.: A2c: A modular multi-stage collaborative decision framework for human–ai teams. Expert Systems with Applications **282**, 127318 (2025)

[29] Shivhare, M., Dhir, S., Shrivastava, S.K.: Next-generation intrusion detection: Navigating challenges and embracing ai. In: 2025 International Conference on Pervasive Computational Technologies (ICPCT), pp. 788–792 (2025). IEEE

[30] Santoso, F., Finn, A.: An in-depth examination of artificial intelligence-enhanced cybersecurity in robotics, autonomous systems, and critical infrastructures. IEEE Transactions on Services Computing **17**(3), 1293–1310 (2023)

[31] Protogerou, A., Kopsacheilis, E.V., Mpatziakas, A., Papachristou, K., Theodorou, T.I., Papadopoulos, S., Drosou, A., Tzovaras, D.: Time series network data enabling distributed intelligence—a holistic iot security platform solution. Electronics **11**(4), 529 (2022)

[32] Kaiser, F.K., Andris, L.J., Tennig, T.F., Iser, J.M., Wiens, M., Schultmann, F.: Cyber threat intelligence enabled automated attack incident response. In: 2022 3rd International Conference on Next Generation Computing Applications (NextComp), pp. 1–6 (2022). IEEE

[33] Kumar, N., Berwal, K., Verma, R., Vishvakarma, S.: Ai-driven strategies for securing internet of battlefield things (iobt). In: 2024 IEEE 4th International Conference on ICT in Business Industry & Government (ICTBIG), pp. 1–6 (2024). IEEE

[34] Chigurupati, M., Malviya, R.K., Toorpu, A.R., Anand, K.: Ai agents for cloud

reliability: Autonomous threat detection and mitigation aligned with site reliability engineering principles. In: 2025 IEEE 4th International Conference on AI in Cybersecurity (ICAIC), pp. 1–4 (2025). IEEE

[35] Nisha, S.S., Patil, H., Bag, A., Singh, A., Kumar, Y., Kumar, J.S.: Critical information framework against cyber-attacks using artificial intelligence and big data analytics. In: 2022 2nd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE), pp. 533–537 (2022). IEEE

[36] Brankovic, B., Ebster, M., Polanec, K., Binder, C., Neureiter, C.: Towards an automated security-by-design approach in automotive system-of-systems architectures. In: 2023 8th International Conference on Smart and Sustainable Technologies (SpliTech), pp. 1–4 (2023). IEEE

[37] Rouff, C., Watkins, L., Sterritt, R., Hariri, S.: Sok: Autonomic cybersecurity-securing future disruptive technologies. In: 2021 IEEE International Conference on Cyber Security and Resilience (CSR), pp. 66–72 (2021). IEEE

[38] Andreoni, M., Lunardi, W.T., Lawton, G., Thakkar, S.: Enhancing autonomous system security and resilience with generative ai: A comprehensive survey. IEEE Access (2024)

[39] Alqahtani, H., Kumar, G.: Cybersecurity in electric and flying vehicles: Threats, challenges, ai solutions & future directions. ACM Computing Surveys **57**(4), 1–34 (2024)

[40] Dehghantanha, A., Parizi, R.M., Epiphaniou, G.: Autonomouscyber'24–workshop on autonomous cybersecurity. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, pp. 4911–4913 (2024)

[41] Dehghantanha, A., Yazdinejad, A., Parizi, R.M.: Autonomous cybersecurity: Evolving challenges, emerging opportunities, and future research trajectories. In: Proceedings of the Workshop on Autonomous Cybersecurity, pp. 1–10 (2023)

[42] Gong, N., Li, Q., Zhang, X.: Aacd'24: 11th acm workshop on adaptive and autonomous cyber defense. In: Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security, pp. 4884–4885 (2024)

[43] Ruman, Á., Drašar, M., Sadlek, L., Yang, S.J., Celeda, P.: Adversary tactic driven scenario and terrain generation with partial infrastructure specification. In: Proceedings of the 19th International Conference on Availability, Reliability and Security, pp. 1–11 (2024)

[44] Bar-Haim, R., Eden, L., Kantor, Y., Agarwal, V., Devereux, M., Gupta, N., Kumar, A., Orbach, M., Zan, M.: Towards automated assessment of organizational cybersecurity posture in cloud. In: Proceedings of the 6th Joint International Conference on Data Science & Management of Data (10th ACM

IKDD CODS and 28th COMAD), pp. 167–175 (2023)

[45] Kaloudi, N., Li, J.: The ai-based cyber threat landscape: A survey. ACM Computing Surveys (CSUR) **53**(1), 1–34 (2020)

[46] Nitz, L., Akbari Gurabi, M., Cermak, M., Zadnik, M., Karpuk, D., Drichel, A., Schäfer, S., Holmes, B., Mandal, A.: On collaboration and automation in the context of threat detection and response with privacy-preserving features. Digital Threats: Research and Practice **6**(1), 1–36 (2025)

[47] Huang, J., Zhu, Q.: Penheal: A two-stage llm framework for automated pentesting and optimal remediation. In: Proceedings of the Workshop on Autonomous Cybersecurity, pp. 11–22 (2023)

[48] Wichtlhuber, M., Strehle, E., Kopp, D., Prepens, L., Stegmueller, S., Rubina, A., Dietzel, C., Hohlfeld, O.: Ixp scrubber: learning from blackholing traffic for ml-driven ddos detection at scale. In: Proceedings of the ACM SIGCOMM 2022 Conference, pp. 707–722 (2022)

[49] Chetwyn, R.A., Eian, M., Jøsang, A.: Modelling indicators of behaviour for cyber threat hunting via sysmon. In: Proceedings of the 2024 European Interdisciplinary Cybersecurity Conference, pp. 95–104 (2024)

[50] Lopes Antunes, D., Llopis Sanchez, S.: The age of fighting machines: the use of cyber deception for adversarial artificial intelligence in cyber defence. In: Proceedings of the 18th International Conference on Availability, Reliability and Security, pp. 1–6 (2023)

[51] Li, S., Zhang, C., Yang, Z.: Research on the military application and development suggestions of artificial intelligence. In: Proceedings of the 2024 2nd International Conference on Artificial Intelligence, Systems and Network Security, pp. 8–12 (2024)

[52] Khadka, A., Sthapit, S., Epiphaniou, G., Maple, C.: Resilient machine learning: Advancement, barriers, and opportunities in the nuclear industry. ACM Computing Surveys **56**(9), 1–29 (2024)

[53] Teixeira De Castro, H., Hussain, A., Blanc, G., El Hachem, J., Blouin, D., Leneutre, J., Papadimitratos, P.: A model-based approach for assessing the security of cyber-physical systems. In: Proceedings of the 19th International Conference on Availability, Reliability and Security, pp. 1–10 (2024)

[54] Baruwal Chhetri, M., Tariq, S., Singh, R., Jalalvand, F., Paris, C., Nepal, S.: Towards human-ai teaming to mitigate alert fatigue in security operations centres. ACM Transactions on Internet Technology **24**(3), 1–22 (2024)

[55] Ekle, O.A., Eberle, W.: Anomaly detection in dynamic graphs: A comprehensive

survey. ACM Transactions on Knowledge Discovery from Data **18**(8), 1–44 (2024)

[56] Welzel, A., Wohlrab, R., Mohamad, M.: Increasing the confidence in security assurance cases using game theory. In: Proceedings of the 19th International Conference on Availability, Reliability and Security, pp. 1–13 (2024)

[57] Greshake, K., Abdelnabi, S., Mishra, S., Endres, C., Holz, T., Fritz, M.: 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 (2023)

[58] Uddin, M.P., Xiang, Y., Hasan, M., Bai, J., Zhao, Y., Gao, L.: A systematic literature review of robust federated learning: Issues, solutions, and future research directions. ACM Computing Surveys **57**(10), 1–62 (2025)

[59] Mundt, M., Baier, H.: Threat-based simulation of data exfiltration toward mitigating multiple ransomware extortions. Digital Threats: Research and Practice **4**(4), 1–23 (2023)

[60] Ortiz, J., Sanchez-Iborra, R., Bernabe, J.B., Skarmeta, A., Benzaid, C., Taleb, T., Alemany, P., Muñoz, R., Vilalta, R., Gaber, C., *et al.*: Inspire-5gplus: Intelligent security and pervasive trust for 5g and beyond networks. In: Proceedings of the 15th International Conference on Availability, Reliability and Security, pp. 1–10 (2020)

[61] Zhang, D., Jain, K., Singh, P.: Guarding against chatgpt threats: Identifying and addressing vulnerabilities. In: 2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR), pp. 612–615. IEEE, ??? (2024)

[62] Kujanpää, K., Victor, W., Ilin, A.: Automating privilege escalation with deep reinforcement learning. In: Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security, pp. 157–168 (2021)

[63] Huyghebaert, S., Keuchel, S., De Roover, C., Devriese, D.: Formalizing, verifying and applying isa security guarantees as universal contracts. In: Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security, pp. 2083–2097 (2023)

[64] Liu, Y., Meng, W.: Acquirer: A hybrid approach to detecting algorithmic complexity vulnerabilities. In: Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, pp. 2071–2084 (2022)

[65] Chen, Y., Poskitt, C.M., Sun, J., Adepu, S., Zhang, F.: Learning-guided network fuzzing for testing cyber-physical system defences. In: 2019 34th IEEE/ACM International Conference on Automated Software Engineering (ASE), pp. 962–973 (2019). IEEE

[66] Candrian, C., Scherer, A.: Rise of the machines: Delegating decisions to autonomous ai. Computers in Human Behavior **134**, 107308 (2022)

[67] Ali, S., Wang, J., Leung, V.C.M.: Ai-driven fusion with cybersecurity: Exploring current trends, advanced techniques, future directions, and policy implications for evolving paradigms–a comprehensive review. Information Fusion, 102922 (2025)

[68] Kiranbabu, M., Viji, A.J., Chandanan, A.K., Birchha, V., Pandey, T.K., Sar, S.K.: The challenge of adversarial attacks on ai-driven cybersecurity systems. Journal of Cybersecurity & Information Management **15**(1) (2025)

[69] Zeng, H., Yunis, M., Khalil, A., Mirza, N.: Towards a conceptual framework for ai-driven anomaly detection in smart city iot networks for enhanced cybersecurity. Journal of Innovation & Knowledge **9**(4), 100601 (2024)

[70] Sarker, I.H., Janicke, H., Mohsin, A., Gill, A., Maglaras, L.: Explainable ai for cybersecurity automation, intelligence and trustworthiness in digital twin: Methods, taxonomy, challenges and prospects. ICT Express (2024)

[71] Desai, B., Patil, K., Mehta, I., Patil, A.: Explainable ai in cybersecurity: A comprehensive framework for enhancing transparency, trust, and human-ai collaboration. In: 2024 International Seminar on Application for Technology of Information and Communication (iSemantic), pp. 135–150 (2024). IEEE

[72] Achuthan, K., Ramanathan, S., Srinivas, S., Raman, R.: Advancing cybersecurity and privacy with artificial intelligence: current trends and future research directions. Frontiers in Big Data **7**, 1497535 (2024)

[73] Ilieva, R., Stoilova, G.: Challenges of ai-driven cybersecurity. In: 2024 XXXIII International Scientific Conference Electronics (ET), pp. 1–4 (2024). IEEE

[74] Radhakrishnan, G.V., Gabhane, M., Jha, S., Al Said, N., Murthy, B.S.R.: Ai and iot in supply chains: Creating intelligent and autonomous business operations. Journal of Information Systems Engineering and Management **10**(11s) (2025)

[75] Sarker, I.H.: Llm potentiality and awareness: a position paper from the perspective of trustworthy and responsible ai modeling. Discover Artificial Intelligence **4**(1), 40 (2024)

[76] Patrício, L., Varela, L., Silveira, Z.: Integration of artificial intelligence and robotic process automation: Literature review and proposal for a sustainable model. Applied Sciences **14**(21), 9648 (2024)

[77] Tereszkiewicz, P., Cichowicz, E.: Findings from the polish insurtech market as a roadmap for regulators. Computer Law & Security Review **52**, 105948 (2024)

[78] Nagarajan, S., Kamalbabu, K.: Analysis of how chatgpt and gemini help in the

generation of cyber attacks. In: 2024 8th International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 663–667. IEEE, ??? (2024)

[79] Rane, N., Qureshi, A.: Comparative analysis of automated scanning and manual penetration testing for enhanced cybersecurity. In: 2024 12th International Symposium on Digital Forensics and Security (ISDFS), pp. 1–6 (2024). IEEE

[80] Yendluri, D.K., Thatikonda, R., Thatikonda, R., Ponnala, J., Kempanna, M., Bhuvanesh, A.: Risk management in robotic process automation. In: 2024 IEEE International Conference on Interdisciplinary Approaches in Technology and Management for Social Innovation (IATMSI), vol. 2, pp. 1–6 (2024)

[81] Heiding, F., Süren, E., Olegård, J., Lagerström, R.: Penetration testing of connected households. Computers & Security **126**, 103067 (2023)

[82] Ghanem, M.C., Chen, T.M., Ferrag, M.A., Kettouche, M.E.: ESASCF: Expertise Extraction, Generalization and Reply Framework for Optimized Automation of Network Security Compliance. IEEE Access **11**, 129840–129853 (2023)

[83] Gupta, M., Akiri, C., Aryal, K., Parker, E., Praharaj, L.: From ChatGPT to ThreatGPT: Impact of Generative AI in Cybersecurity and Privacy. IEEE Access **11**, 80218–80245 (2023)

[84] Dasawat, S.S., Sharma, S.: Cyber security integration with smart new age sustainable startup business, risk management, automation and scaling system for entrepreneurs: An artificial intelligence approach. In: 2023 7th International Conference on Intelligent Computing and Control Systems (ICICCS), pp. 1357–1363 (2023). IEEE

[85] Chalaris, M.: In the fourth industrial revolution era, security, safety, and health. Journal of Engineering Science & Technology Review **16**(2) (2023)

[86] Șcheau, M.C., Achim, M.V., Găbudeanu, L., Văidean, V.L., Vîlcea, A.L., Apetri, L.: Proposals of processes and organizational preventive measures against malfunctioning of drones and user negligence. Drones **7**(1), 64 (2023)

[87] Benyahya, M., Collen, A., Kechagia, S., Nijdam, N.A.: Automated city shuttles: Mapping the key challenges in cybersecurity, privacy and standards to future developments. Computers & Security **122**, 102904 (2022)

[88] Santofimia, M.J., Villanueva, F.J., Litvinov, E., Viksnin, I., Fernandes, A., Lopez, J.C.: Cybersecurity in active and healthy ageing era. Procedia Computer Science **192**, 2068–2076 (2021)

[89] Gkotsopoulou, O., Charalambous, E., Limniotis, K., Quinn, P., Kavallieros, D.,

Sargsyan, G., Shiaeles, S., Kolokotronis, N.: Data protection by design for cybersecurity systems in a smart home environment. In: 2019 IEEE Conference on Network Softwarization (NetSoft), pp. 101–109 (2019). IEEE

[90] Xie, Y., Gardi, A., Sabatini, R.: Cybersecurity risks and threats in avionics and autonomous systems. In: 2023 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech), pp. 0814–0819 (2023). IEEE

[91] Siddiqui, F., Khan, R., Sezer, S.: Bird's-eye view on the automotive cybersecurity landscape & challenges in adopting ai/ml. In: 2021 Sixth International Conference on Fog and Mobile Edge Computing (FMEC), pp. 1–6 (2021). IEEE

[92] Shapiro, M.: The extractive power of automated surveillance. Public Anthropologist **6**, 267–291 (2024). Online publication date: 11 Dec 2024

[93] Swain, R., Truelove, V., Rakotonirainy, A., Kaye, S.-A.: A comparison of the views of experts and the public on automated vehicles technologies and societal implications. Technology in Society **74**, 102288 (2023)

[94] Lim, H.S.M., Taeihagh, A.: Algorithmic decision-making in avs: Understanding ethical and technical concerns for smart cities. Sustainability **11**(20), 5791 (2019)

[95] Krakowski, S., Luger, J., Raisch, S.: Artificial intelligence and the changing sources of competitive advantage. Strategic Management Journal **44**(6), 1425–1452 (2023)

[96] Islam, M.M., Hosen, M.S., Khan, S.I., Vayyala, R., Oyebode, D., Mia, M.S.: Data analyst with expertise in artificial intelligence and machine learning algorithms. Journal of Information Systems Engineering and Management **10**(14s) (2025) 2468-4376. Research Article

[97] Zhang, C., Zheng, Y., Bai, M., Li, Y., Ma, W., Xie, X., Li, Y., Sun, L., Liu, Y.: How effective are they? exploring large language model based fuzz driver generation. In: Proceedings of the 33rd ACM SIGSOFT International Symposium on Software Testing and Analysis, pp. 1223–1235 (2024)

[98] Bornet, P., Wirtz, J., Davenport, T.H., De Cremer, D., Evergreen, B., Fersht, P., Gohel, R., Khiyara, S., Sund, P., Mullakara, N.: Agentic Artificial Intelligence: Harnessing AI Agents to Reinvent Business, Work and Life. Irreplaceable Publishing, ??? (2025)

[99] Kemp, A.: Competitive advantage through artificial intelligence: Toward a theory of situated ai. Academy of Management Review **49**(3), 618–635 (2024)