

# A cybersecurity AI agent selection and decision support framework

Masike Malatji

*Graduate School of Business Leadership (SBL), University of South Africa (UNISA)*

*Midrand, Johannesburg, South Africa, PO Box 392, Unisa, 0003*

*[malatmi1@unisa.ac.za](mailto:malatmi1@unisa.ac.za)*

*<https://orcid.org/0000-0002-9893-9598>*

## ABSTRACT

This paper presents a novel, structured decision support framework that systematically aligns diverse artificial intelligence (AI) agent architectures—reactive, cognitive, hybrid, and learning—with the comprehensive National Institute of Standards and Technology (NIST) Cybersecurity Framework (CSF) 2.0. By integrating agent theory with industry guidelines, this framework provides a transparent and stepwise methodology for selecting and deploying AI solutions to address contemporary cyber threats. Employing a granular decomposition of NIST CSF 2.0 functions into specific tasks, the study links essential AI agent properties such as autonomy, adaptive learning, and real-time responsiveness to each subcategory's security requirements. In addition, it outlines graduated levels of autonomy (assisted, augmented, and fully autonomous) to accommodate organisations at varying stages of cybersecurity maturity. This holistic approach transcends isolated AI applications, providing a unified detection, incident response, and governance strategy. Through conceptual validation, the framework demonstrates how tailored AI agent deployments can align with real-world constraints and risk profiles, enhancing situational awareness, accelerating response times, and fortifying long-term resilience via adaptive risk management. Ultimately, this research bridges the gap between theoretical AI constructs and operational cybersecurity demands, establishing a foundation for robust, empirically validated multi-agent systems that adhere to industry standards.

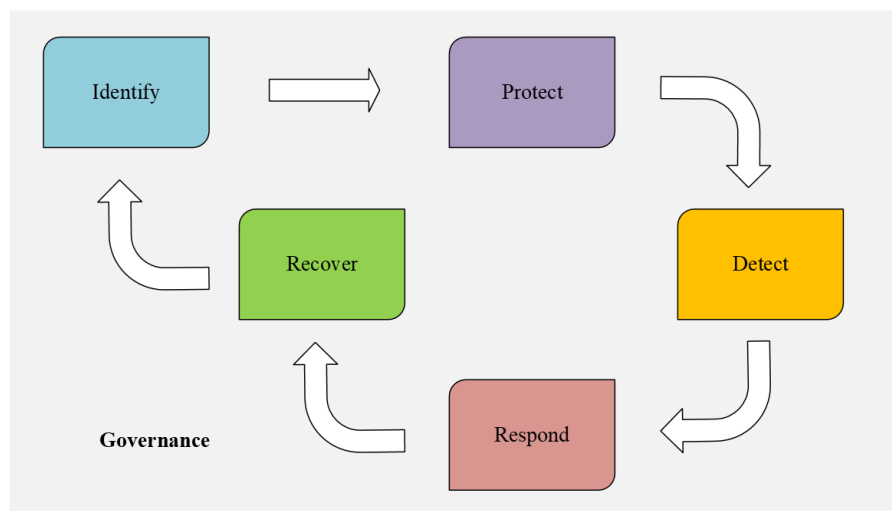
**Keywords:** Agents; Autonomous; Cybersecurity; Detection; Framework; GenAI; Governance

## 1. Introduction

The integration of artificial intelligence (AI) has fundamentally reshaped the contemporary cybersecurity landscape, emerging as a critical driver in the ongoing battle against increasingly sophisticated cyber threats. While traditional security paradigms, reliant on rule-based or signature-based detection, struggle to keep pace with the dynamic and evolving nature of modern attacks – characterised by the proliferation of novel exploits, polymorphic malware, and zero-day vulnerabilities [1], [2] – AI offers a paradigm shift. Leveraging the power of machine learning (ML) and deep learning (DL) algorithms, AI demonstrates unparalleled proficiency in real-time analysis of vast datasets, enabling the identification of subtle and often elusive indicators of compromise [3], [4]. This advancement empowers AI-driven systems with potent techniques such as pattern recognition, anomaly detection, and predictive analytics, automating core cybersecurity functions like intrusion detection and proactive threat hunting [5], [6]. This automation minimises the reliance on manual intervention. It addresses the critical shortage of skilled cybersecurity professionals by providing invaluable support to Security Operations Centres (SOCs) in maintaining continuous surveillance and managing high-volume incident streams [7], [8].

However, the attributes that position AI as a formidable defensive tool – its inherent automation, rapid learning capabilities, and capacity for generating novel insights – also present new avenues for exploitation by malicious actors. This necessitates a proactive and adaptive approach to cybersecurity. The evolving landscape demands the strategic adoption of AI-based defensive measures and the development of resilient, explainable AI (XAI) models capable of dynamically adapting to evolving attack vectors while preserving transparency and fostering user trust [9], [10].

The true power of AI in cybersecurity lies in its synergistic combination with human expertise, enabling a proactive, scalable, and intelligent approach to threat management [11]. However, realising this potential necessitates a sound and multifaceted security strategy. This strategy must extend beyond mere technological implementation to encompass good governance frameworks, address critical ethical considerations, and prioritise thorough validation of AI model performance [12], [13]. A cornerstone of responsible AI integration in cybersecurity is its alignment with established industry frameworks, standards, and guidelines. The National Institute of Standards and Technology (NIST) Cybersecurity Framework (CSF), particularly its updated version 2.0 [14], provides a critical foundation for this alignment. Fig. 1 illustrates the NIST CSF 2.0, a versatile framework designed to empower organisations across diverse sectors, including government agencies, in effectively managing their cybersecurity risks [14], [15].



**Fig. 1** NIST CSF 2.0: Illustrates the high-level structure of the NIST Cybersecurity Framework 2.0, highlighting its six core functions and how they interrelate

The NIST CSF 2.0 offers a comprehensive taxonomy of high-level cybersecurity outcomes, serving as a valuable tool for organisations of all sizes, sectors, and maturity levels. It facilitates a more precise understanding, assessment, prioritisation, and communication of cybersecurity efforts [14]. Notably, the NIST CSF 2.0 is not prescriptive in its implementation; it serves as a guiding resource, complemented by supplementary guidance on potential practices and controls that can be leveraged to achieve the desired security outcomes [14]. As visually represented in Fig. 1, the NIST CSF 2.0 is organised around a cohesive set of six core functions that underpin a resilient cybersecurity posture: Identify, Protect, Detect, Respond, Recover, and Govern. These interconnected

functions provide a comprehensive framework for managing and mitigating organisational cybersecurity risks. According to the NIST [14] guidelines, the core functions of the NIST CSF 2.0 are defined as follows:

- *Govern*: This overarching function establishes the importance of an organisation's commitment to defining, communicating, and consistently monitoring its cybersecurity risk management strategy, expectations, and policies.
- *Identify*: This foundational function emphasises the critical need for an organisation to understand its current cybersecurity risks thoroughly.
- *Protect*: This function focuses on implementing adequate safeguards to manage and mitigate identified cybersecurity risks.
- *Detect*: This involves the timely identification and analysis of potential cybersecurity attacks and compromises.
- *Respond*: This function outlines the necessary actions to address and minimise the impact of detected cybersecurity incidents.
- *Recover*: This focuses on the swift restoration of assets and operations affected by a cybersecurity incident to a functional state.

The practical implementation of the NIST CSF 2.0, however, can be significantly enhanced by leveraging the capabilities of advanced technologies. As mentioned earlier, AI has the potential to offer powerful tools to operationalise and strengthen these core cybersecurity functions. Indeed, as the sophistication of cyber threats continues to escalate [16], the need for comprehensive, adaptive, and scalable solutions across heterogeneous digital ecosystems becomes increasingly critical [17], [18]. With their inherent capacity for autonomous reasoning, real-time decision-making, and collaborative learning [19], [20], AI agents have emerged as compelling candidates to meet these evolving demands. AI agents are computational entities, often comprising software and sometimes hardware components, designed to exhibit autonomy, social interaction, responsiveness, and proactive behaviour [21], [22], [23], [24]. These agents can perceive their environment and execute actions to achieve predefined objectives [25]. While definitions may vary across disciplines, a common thread emphasises their ability to function independently, adapt to dynamic conditions, and interact with other agents or systems [24], [26], [27]. Within the cybersecurity domain, AI agents hold significant potential for automating critical tasks, such as threat monitoring, detection, and incident response, by effectively leveraging both reactive and proactive decision-making strategies.

From an architectural standpoint, the landscape of AI agents encompasses a diverse range of designs, each tailored to specific functionalities and environments. Key categories include virtual agents, which operate within digital environments; embodied agents, which possess a physical presence and interact with the physical world; reactive agents, characterised by their immediate response to environmental stimuli; hybrid agents, integrating multiple architectural paradigms; learning agents, which possess the capacity to adapt and improve their performance over time; and cognitive or deliberative agents, characterised by higher-level functions such as planning, reasoning, and learning [21], [28], [29], [30], [31], [32]. Categorising AI agents provides a framework for researchers and practitioners to strategically align agent capabilities and design limitations with the specific requirements of diverse application domains. This is particularly useful in modern cybersecurity, where layered defence and dynamic threat response strategies demand tailored technological solutions. To this end, an appropriate, theory-driven mapping of AI agent architectures to the NIST CSF 2.0 is required. Such a mapping ensures that the technical strengths of various agent types are effectively leveraged to support the core functions that underpin an organisation's cybersecurity posture.

However, the optimal AI agent architecture is not a one-size-fits-all solution. For instance, while reactive agents demonstrate exceptional proficiency in rapid, event-driven responses [29], they may lack the strategic foresight essential for proactive threat hunting or optimal resource allocation. Conversely, learning and cognitive agents offer sound planning and adaptive capabilities [28], but often come with increased computational and maintenance complexities [33]. Thus, research gaps persist in the current literature regarding the systematic and theoretically grounded integration of AI agents within established cybersecurity frameworks. In other words, while the potential of AI agents in bolstering cyber defences has been widely acknowledged, a cohesive and standardised framework for their implementation remains largely absent. A study by Kott [34] envisioned a future in which intelligent, autonomous cyber defence agents would play a pivotal role in threat mitigation. Subsequent research, including contributions from Théron and Kott [35] and Kott et al. [36], has further advanced this vision by proposing architectural models for autonomous intelligent cyber defence agents capable of outperforming human response speed and agility. Expanding the scope, Truong et al. [37] provided a broader perspective on the applications of AI agents across both offensive and defensive cybersecurity domains. Ligo et al. [38] delved into the intricacies of measuring cyber-resilience in systems that incorporate autonomous agents, highlighting the complexities of evaluating their effectiveness. Furthermore, Naik et al. [39] offered a review of AI techniques, including the role of AI agents in analysing, detecting, and mitigating diverse cyber threats. More recently, Sharma and Jindal [40]

explored the broader implications of AI, underscoring its potential to advance intelligent agent technologies across various domains, including cybersecurity. Despite these valuable contributions, a unified, framework-driven approach to integrating and evaluating AI agents within established cybersecurity standards remains an area for further investigation.

In the evolving landscape of AI in cybersecurity, this paper aims to develop a framework for selecting, designing, and deploying AI agents to address the NIST CSF 2.0 cybersecurity requirements. The contribution of this framework is that it provides a structured lens through which to understand, for instance, how the adaptive learning capabilities of learning agents can enhance the Detect and Respond functions of the NIST CSF 2.0 by proactively adapting to emerging threat vectors [14], [41], [42]. Similarly, it clarifies how the strategic planning inherent in cognitive/deliberative agents can better facilitate the Governance function by effectively balancing compliance requirements with actionable threat intelligence [14], [43]. Without a clearly defined mapping, organisations risk deploying AI agents ill-suited to their specific security needs or failing to fully capitalise on the unique strengths of agents optimised for particular operational contexts. Therefore, rigorous conceptual mapping is vital for informing decision-makers and system architects, guiding them towards the most appropriate agent solutions tailored to their cybersecurity objectives. This alignment between conceptual rigour and practical applicability paves the way for empirical validation and targeted tool implementations, ultimately fostering effective and sustainable AI-driven security operations.

This study is grounded in foundational concepts from agent-based AI, particularly the classification and behavioural properties of AI agents as applied to cybersecurity [21], [22], [24]. The framework builds upon four core agent architectures, reactive, cognitive, hybrid, and learning agents, each defined by its level of autonomy, decision-making mechanisms, and interaction with dynamic environments [21], [28], [29], [30], [31], [32]. As mentioned earlier, the paper also draws from recognised cybersecurity governance models, primarily the NIST CSF 2.0, which serves as the anchor for mapping agent capabilities to structured cybersecurity functions. Additionally, the concept of graduated levels of autonomy is introduced, informing the selection logic that guides agent deployment across the Identify, Protect, Detect, Respond, Recover, and Govern functions of the NIST CSF 2.0. This layered autonomy spectrum, ranging from assisted [44], [45], augmented [46], [47], to fully autonomous intelligence [48], [49], [50], enables a capability-based interpretation of agent suitability aligned with organisational maturity, task complexity, and risk appetite. These theoretical constructs form the foundational lens through which the AI Agent Taxonomy and Decision Framework (AIATDF) is developed and validated in this paper.

The methodology for developing the framework employs a matrix-based approach to map specific agent properties (e.g., autonomy, adaptiveness, reactivity) to the subcategories and tasks defined by the NIST CSF 2.0. This mapping forms the basis for identifying which agent architectures (e.g., reactive, hybrid, learning) are best suited to different cybersecurity functions (e.g., Detect, Respond, Recover) [14], [28], [29]. In a nutshell, the proposed approach is grounded in a structured, matrix-based conceptual mapping of agent capabilities to cybersecurity requirements. Below is a concise overview of the *main contributions* presented in this paper:

- *AI Agent Taxonomy and Decision Framework (AIATDF):*  
Presents a structured model for classifying AI agents and systematically mapping their properties to the NIST CSF 2.0 functions (Fig. 6).
- *Conceptual mapping matrix:*  
Provides a detailed mapping matrix, linking specific AI agent types and capabilities to NIST CSF 2.0 subcategories, facilitating optimal architecture selection for real-world cybersecurity tasks (Table 6).
- *Graduated levels of autonomy:*  
Establishes a preliminary capability maturity model (CMM), enabling organisations with varying maturity levels to incrementally adopt AI-driven solutions, progressing from assisted to augmented and autonomous intelligence (Fig. 3).
- *Holistic NIST CSF 2.0 alignment:*  
Demonstrates the integrated deployment of AI agents across all NIST CSF 2.0 functions, providing an end-to-end cybersecurity management approach (Table 6).
- *Rigorous methodology and applied utility:*  
Combines hierarchical task analysis with AI agent theory, yielding a transparent and reproducible framework suitable for academic research and applied SOC deployments (Steps beneath Fig. 6).

The theoretical framework presented here is based on several core assumptions. Firstly, I assume that the NIST CSF 2.0 provides sufficiently granular and actionable functions and subcategories for mapping AI agent tasks in operational settings. Secondly, I contend that the inherent properties of AI agents, such as autonomy, reactivity, and learning capabilities, can be sufficiently characterised and conceptually aligned with established cybersecurity requirements. Thirdly, the proposed framework assumes a baseline level of digital infrastructure and security maturity within deploying organisations. Despite these foundational assumptions, the study's primary limitation lies

in its conceptual genesis; the AIATDF remains to be validated through empirical testing in live cybersecurity environments. Furthermore, while drawing extensively from scholarly and industry literature, my classification of AI agent architectures and properties may inherently simplify dynamically evolving, complex hybrid systems. These limitations are elaborated upon in Section 6.3.

Building upon the foundational understanding established in the preceding content, this paper proceeds as follows: Section 2 provides a detailed review of the relevant literature, explicitly exploring the capabilities of various AI agent architectures, the tenets of the NIST CSF 2.0, and the rationale for systematically mapping these agent capabilities to the framework's functional requirements. Subsequently, Section 3 outlines the methodological approach employed in this study to develop this mapping framework. The framework is presented in Section 4 and conceptually validated in Section 5. Discussions of the mapping framework, its implications, and inherent limitations are provided in Section 6. Finally, Section 7 concludes the paper by offering key takeaways and practical recommendations for future cybersecurity practices and suggesting promising avenues for further research in this evolving domain.

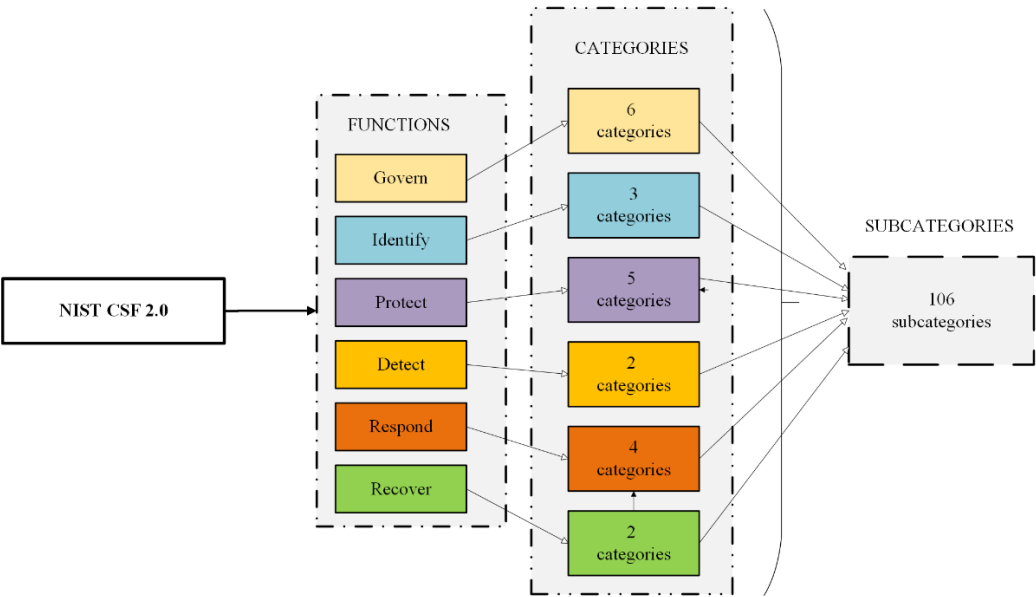
## 2. Related Works

### 2.1 NIST CSF 2.0

The NIST CSF 2.0 is an updated and comprehensive set of guidelines, standards, and best practices developed by the NIST to help organisations manage and reduce cybersecurity risks [14]. NIST is a non-regulatory federal agency that provides the scientific and technical foundation for innovation, economic growth, and public safety in the United States of America. The NIST CSF 2.0 is designed to be flexible and adaptable for organisations of all sizes, sectors, and cybersecurity maturity levels [14].

#### 2.1.1 The evolution of the NIST CSF

The NIST CSF has undergone a significant and necessary evolution from its initial release as version 1.0 in 2014 to the more comprehensive and globally relevant version 2.0. This progression mirrors the escalating complexity and scale of contemporary cybersecurity threats, necessitating more resilient and adaptable security strategies. While the foundational framework in version 1.0 established five core functions – Identify, Protect, Detect, Respond, and Recover [51], [52], [53] – version 2.0 represents a notable enhancement. A key advancement in version 2.0 is the introduction of a dedicated "Govern" function, which explicitly elevates the importance of cybersecurity within an organisation's overall governance structure [14], [53]. Therefore, the NIST CSF 2.0 is structured around six core functions, which are further elaborated through twenty-two categories representing specific cybersecurity outcomes and one hundred and six subcategories detailing more granular technical and management activities [14]. Fig. 2 provides a visual representation of this layered structure.



**Fig. 2** NIST CSF 2.0 framework core: Depicts the layered composition of the NIST Cybersecurity Framework 2.0, showing how the 22 categories and 106 subcategories are organised under the six core functions of the framework

The twenty-two NIST CSF 2.0 categories are summarised in Table 1 [14].

**Table 1** NIST CSF 2.0 categories: The 22 categories within the updated NIST Cybersecurity Framework 2.0 are listed, grouping them by the six core functions (Govern, Identify, Protect, Detect, Respond, Recover)

Function	Category
Govern	○ Organisational context
	○ Risk management strategy
	○ Roles, responsibilities, and authorities
	○ Policy
	○ Oversight
Identify	○ Cybersecurity supply chain risk management
	○ Asset management
	○ Risk assessment
	○ Improvement
Protect	○ Identity management, authentication, and access control
	○ Awareness and training
	○ Data security
	○ Platform security
	○ Technology infrastructure resilience
Detect	○ Continuous monitoring
	○ Adverse event analysis
Respond	○ Incident management
	○ Incident analysis
	○ Incident response reporting and communication
	○ Incident mitigation
Recover	○ Incident Recovery Plan Execution
	○ Incident Recovery Communication

Furthermore, the updated framework broadens its scope with a global perspective, ensuring its applicability and utility for worldwide organisations [52], [53]. Version 2.0 provides more detailed guidance and actionable recommendations, offering practical strategies and steps for cybersecurity professionals at all levels [53]. This enhanced version also explicitly addresses the rapid advancements in technology, including the Internet of Things, cloud computing, and big data, offering specific guidelines to manage the associated cybersecurity risks [14], [54], [55]. Moreover, the updated framework also emphasises a risk-based approach, equipping organisations with more solid methods for identifying, assessing, and effectively managing potential threats [14], [53]. Ultimately, version 2.0 empowers organisations to develop more comprehensive and threat-informed cybersecurity strategies tailored to the specific needs of diverse sectors and industries [52], [53]. The expanded functions and detailed guidance in version 2.0 provide cybersecurity professionals with a more precise and actionable roadmap for implementing effective security measures, including hundreds of immediately applicable recommendations [14], [53]. This enhanced international applicability positions version 2.0 as a valuable tool for standardising cybersecurity practices and fostering greater global cybersecurity resilience [14], [53].

However, because the functions, categories and subcategories are designed to be high-level and adaptable to various organisations, they only outline *what* cybersecurity outcomes should be achieved, but not necessarily *how*. ‘Informative references’ of the NIST CSF 2.0 provide the “how” by pointing to specific standards, guidelines, and best practices that offer detailed implementation guidance [14]. This is the level at which more granular cybersecurity controls are found and implemented. Thus, a clearly defined mapping framework is required to systematically examine the alignment between distinct AI agent capabilities and the functional pillars of the NIST CSF 2.0.

### 2.1.2 Rationale for mapping AI agents' capabilities to the NIST CSF 2.0

Reflecting the rapidly evolving cybersecurity landscape, NIST has proactively initiated an examination into how existing frameworks, such as the NIST CSF 2.0, can assist organisations in navigating emerging and expanding risks. As a testament to this, on February 14, 2025, NIST [56] published a concept paper for a cybersecurity and AI workshop, seeking public input on the critical challenge of addressing cybersecurity risks associated with the development and deployment of AI. The subsequent workshop, held on April 3, 2025, further underscored this focus. In essence, the NIST [56] concept paper centres on identifying and mitigating AI-related sources of cybersecurity risk that can significantly impact an organisation's operational risk profile. The concept paper categorises these risks into three key areas: (i) cybersecurity of AI systems, (ii) AI-enabled cyber attacks, and (iii) AI-enabled cyber defence, which align with the categories of adversarial AI, offensive AI, and defensive AI, respectively, previously described by Malatji and Tolah [57]. To quote NIST [56] directly, “*there is no consistent taxonomy or agreement on how AI advances inform organisations’ strategies for cybersecurity risk management.*”

They propose a “Cyber AI Profile” for guiding organisations deploying AI technologies and/or defending against AI-enabled attacks. At the time of writing, NIST is collaborating with its partners and stakeholders on the “Cyber AI Profile.”

Against this backdrop, this paper's rationale for mapping AI agents to the NIST CSF 2.0 is that the strategic integration of these agents with the NIST CSF 2.0 offers a significant (autonomous) opportunity to fortify an organisation's cybersecurity posture across its core functions. This involves strategically leveraging AI's inherent capabilities to enhance asset identification, bolster threat detection, strengthen protective measures, streamline incident response, and accelerate recovery processes [4], [58], [59]. By embedding AI agents within these and other cybersecurity functions, organisations can anticipate even more improvements in response times and ensure more resilient, continuous protection against the ever-evolving threat landscape. Therefore, a structured mapping and decision-making tool is essential for effectively aligning the technical capabilities of diverse AI agent architectures with the specific security tasks that define an organisation's overall cybersecurity posture. Ultimately, the absence of such a structured mapping can directly contribute to security vulnerabilities and potentially lead to SOC underperformance due to the deployment of ill-suited AI agent solutions. In the next section, I review existing literature on the generic application of AI on the NIST CSF 2.0.

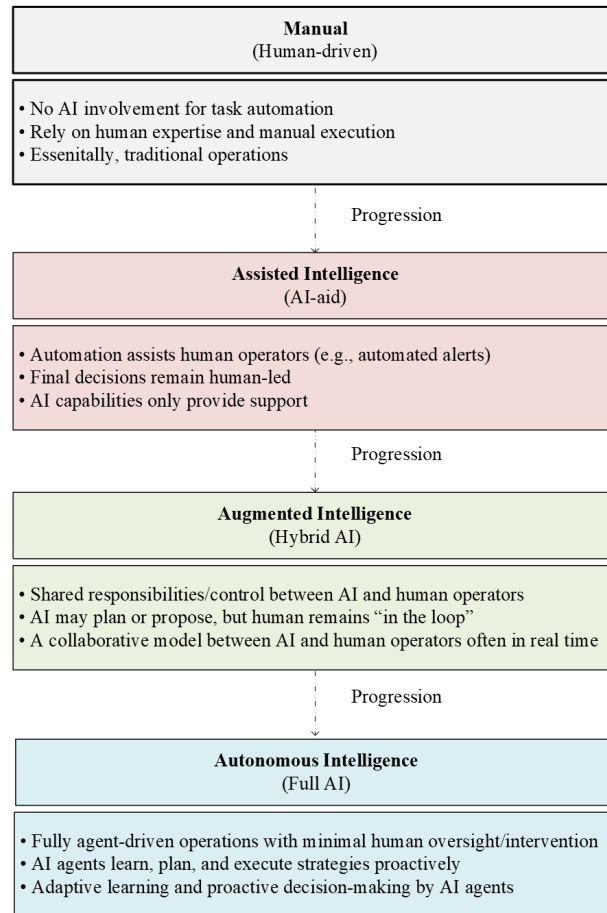
### 2.1.3 Application of AI on NIST CSF

This paper distinguishes between automation and autonomy within the context of AI/ML. From this perspective, the core differentiation hinges on the degree of intelligence embedded within the system and the extent of human involvement in the decision-making process [60], [61]. In this context, automation refers to AI-enhanced systems that execute predefined rules or algorithms to accomplish tasks while maintaining a significant degree of human control over the system's actions [62], [63]. These systems, even when employing sophisticated ML for optimisation, operate within the constraints of their initial programming. A typical example is a spam filter, which utilises algorithms and learned patterns to categorise emails based on pre-set parameters [64], demonstrating adaptability but not fundamentally altering its operational logic.

Conversely, autonomy describes AI systems characterised by their capacity for independent decision-making in dynamic and unpredictable environments [48], [49], [50]. These systems exhibit a greater degree of "agency" [65], enabling them to perceive their surroundings, interpret information, formulate plans, and execute actions to achieve objectives without explicit, step-by-step instructions for every possible scenario [24], [25], [26], [27]. A prime illustration is a self-driving car, which must navigate complex traffic scenarios, adapt to unexpected obstacles, and make real-time decisions regarding speed, direction, and braking [59], [66], [67]. While reliant on algorithms and pre-trained models, a self-driving car's ability to respond to novel and unforeseen situations distinguishes it from a purely automated system. Therefore, the fundamental distinction between automation and autonomy lies in the system's inherent capacity for adaptive behaviour, independent problem-solving, and the level of human intervention [48], [49], [50]. Automated systems excel in optimising predictable processes, whereas autonomous systems are designed to tackle unpredictable ones, necessitating continuous learning, adaptation, and real-time judgment [68]. This critical distinction underpins understanding the varying potential and limitations of different AI applications, a theme explored throughout this paper.

According to Morovat and Panda [69] and Dehghantanha et al. [68], cybersecurity paradigms evolve from traditional to AI-enhanced to autonomous. Traditional cybersecurity involves signature-based or rule-based defences, characterised by their reactive nature and limited capacity for dynamic analysis of novel threats [70]. This approach primarily responds to known threat patterns, lacking the agility to adapt to the ever-evolving adversary. AI-driven cybersecurity represents a significant advancement, integrating automated responses and leveraging ML for more flexible threat detection [10]. This shift transitions from static, rule-driven methodologies to sophisticated pattern-based analytics and accelerated incident triage. This paper focuses on the most advanced paradigm: autonomous cybersecurity. This paradigm leverages AI agents capable of reinforcement learning (RL), continuous self-improvement, and predictive orchestration across complex, distributed systems [71], [72], [73]. A core emphasis of this approach lies in developing XAI and implementing highly personalised threat defences that can rapidly adapt to the dynamic and unpredictable nature of modern attack landscapes [74], [75]. Furthermore, as AI agents are increasingly embedded in cybersecurity infrastructures, concerns about transparency, fairness, and privacy are becoming central to their design and deployment. As highlighted by Radanliev [76], ethical AI development demands proactive strategies that address data bias, algorithmic opacity, and differential privacy. These ethical safeguards are critical not only for user trust but also for regulatory compliance in sensitive domains like cybersecurity [9], [10], where decision-making must be both explainable and justifiable under audit conditions [77]. Incorporating such perspectives ensures that AI agents are not merely technically capable but also socially responsible and accountable.

Building upon the works of [Dehghantanha et al. \[68\]](#) and [Morovat and Panda \[69\]](#), this paper distinguishes four primary modes of AI system intelligence in cybersecurity. Fig. 3 illustrates the graduated levels of autonomy, ranging from purely manual (human-driven) to assisted intelligence, augmented intelligence, and ultimately, complete autonomy.



**Fig. 3** Graduated levels of autonomy in cybersecurity: Visualises a CMM’s progression from manual, human-driven modes to assisted, augmented, and fully autonomous intelligence within cybersecurity contexts

The *manual* or human-driven mode in Fig. 3 represents a state without automation or AI integration, relying entirely on human expertise. *Assisted intelligence* enhances human capabilities and cognitive functions by providing information, automating specific tasks, or offering insights [44], [45]. Within this mode, AI systems act as tools to enhance human abilities for efficiency and effectiveness, with ultimate responsibility for outcomes remaining with human operators. A prime example is healthcare, where assistive AI aids clinicians in making more informed diagnostic or management decisions, as seen in applications ranging from virtual informatics systems for health management to physical robotic-assisted surgeries [78]. The previously mentioned cybersecurity spam filter also exemplifies this category, as the human operator retains responsibility for email filtering, positioning assisted intelligence as a form of AI/ML-driven automation. *Augmented intelligence* goes a step further than assisted intelligence. It is about a collaborative partnership between human intelligence and AI to amplify human capabilities by leveraging the unique strengths of both while mitigating their inherent weaknesses [46], [47].

While assisted intelligence inherently operates within the human-in-the-loop (HITL) human-machine interaction (HMI) paradigm, augmented intelligence possesses the capability to function across both HITL and the more advanced human-out-of-the-loop (full autonomy) HMI paradigm [79], [80], [81], [82], [83]. Therefore, augmented intelligence exhibits characteristics of automation and autonomy, effectively representing a hybrid mode that bridges assisted and autonomous intelligence. In cybersecurity, augmented intelligence enhances threat detection and response by combining human expertise with the analytical power of AI/ML, leading to more accurate and efficient handling of complex threats [84]. Finally, *autonomous intelligence*, as discussed at the beginning of this section, encompasses systems capable of independent task execution without direct human intervention, relying on advanced AI/ML techniques [48], [49], [50]. These systems are increasingly prevalent across diverse domains,



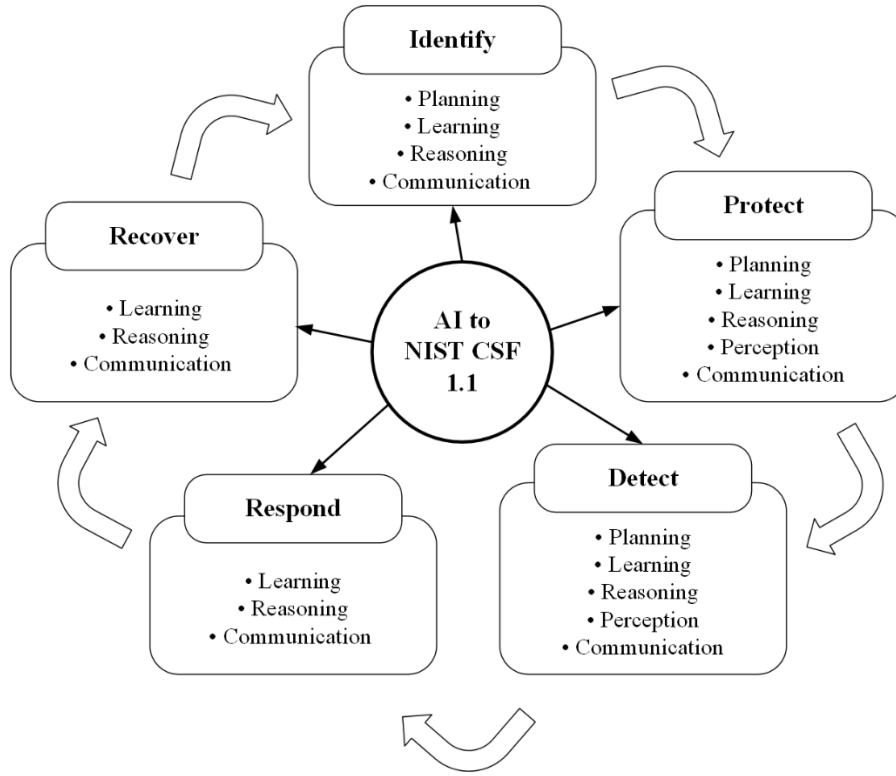
including healthcare, military, transportation, and cybersecurity [83], [85], [86]. This paper focuses explicitly on AI agents operating as autonomous intelligent entities within the cybersecurity domain, while acknowledging their potential deployment in augmented intelligence mode. To the best of my knowledge, only Salesforce [87] released a similar AI agent CMM with graduated levels of autonomy, as shown in Fig. 3, on April 10, 2025. In other words, there is a shortage of scholarly, well-defined CMMs for the adoption of AI agents. This avenue forms part of my future research as it is beyond the scope of this paper.

Beyond the current paradigm of autonomous AI agents in cybersecurity, emerging research heralds a future of transformative capabilities. Frontier research explores quantum-safe AI algorithms to fortify defences against future quantum attacks [88] and privacy-preserving ML, including homomorphic encryption and differential privacy [89], to ensure data security without compromising agent efficacy. Federated learning is also emerging [90], promising decentralised AI pipelines that enable collaborative learning while preserving data locality and regulatory compliance. Though currently in development, these cutting-edge technologies hold the potential to redefine agent architectures and significantly enhance cybersecurity resilience. In this paper, thirty-five units of analysis were obtained from a systematic literature review (SLR) process through Scopus (12 papers), IEEE (5 papers), and Google Scholar (18 papers) databases. The key phrase “*AI agents in cybersecurity*” was utilised to search for the relevant literature. The notable outcomes of the SLR are summarised in Table 2.

**Table 2** Literature review of the application of AI on NIST CSF: Summarises existing studies that intersect cybersecurity, AI agents, and NIST CSF, indicating whether each study addressed the NIST CSF and its functions

Author	Purpose of the study	Focus on cybersecurity	Focus on AI agents	Focus on NIST CSF
NIST [56]	Evaluate how an updated NIST CSF 2.0 can assist organisations facing new or expanded AI-related risks.	✓	✗	Potentially
Cronin [72]	Reviews generic AI agents and those that use Generative AI.	✗	✓	✗
Yu et al. [91]	Explore the potential for blockchain to be a foundational infrastructure for AI agents in the metaverse.	✗	✓	✗
Han et al. [92]	Propose a multi-agent system (MAS) to enhance financial investment research.	✗	✓	✗
Chan et al. [71]	Evaluate metrics to increase visibility into AI agents.	✗	✓	✗
Chen et al. [93]	Explore the role of 6G in realising AI agents' potential.	✗	✓	✗
Bovo et al. [94]	Explore large language models (LLM) agents in extended reality environments.	✗	✓	✗
Baabdullah [95]	Explores the impact of decision-making efficiency facilitated by generative conversational AI agents.	✗	✓	✗
Huang et al. [96]	Explore embodied AI agent systems.	✗	✓	✗
Agashe et al. [97]	Propose agent S, an open agentic framework.	✗	✓	✗
Kim and Saad [98]	Propose a novel continual learning algorithm for AI agents.	✗	✓	✗
Pleshakova et al. [99]	Create a neural network architecture for multi-agent tasks.	✓	✓	✓
Sharma and Jindal [40]	Explore the role of AI agents in various domains.	✗	✓	✗
Dehghantanha et al. [68]	Provide an overview of the current state of autonomous cybersecurity, highlighting the key challenges and opportunities.	✓	✓	✗
Hauptman et al. [100]	Explore how a team's work cycle could guide an AI agent's changing level of autonomy.	✓	✓	✓
Kott et al. [36]	Advance the reference architecture work on the AICA.	✓	✓	✗
Kaur et al. [101]	Analysed 236 AI use cases for cybersecurity provisioning against the NIST CSF 1.1.	✓	✓	✓
Cañas [102]	Studies human and AI agents' responsibility when interacting.	✗	✓	✗
Naik [39]	Review the application of AI techniques in fighting various cyberattacks.	✓	✗	✗
Roy [103]	Propose a MAS that detects and neutralises unseen cyber anomalies.	✓	✓	✗
Li et al. [104]	Train AI agents using the Cyber Gym for Intelligent Learning.	✗	✓	✗
Ligo et al. [38]	Examine approaches to measuring cyber-resilience systems with autonomous agents.	✓	✓	✗
Prasad et al. [77]	Evaluate AI agents' decisions using the Testing with Concept Activation Vectors XAI technique.	✓	✓	✗
Ashktorab et al. [105]	Investigate the social perceptions of AI agents.	✗	✓	✗
Zolotukhin et al. [106]	Explore attack mitigation techniques in software-defined networking (SDN) environments.	✓	✓	✗
Cao et al. [107]	Address fifth-generation (5G) SDN challenges with AI agents.	✗	✓	✗
Franco et al. [108]	Introduce a cybersecurity-driven conversational agent.	✓	✓	✗
Morovat and Panda [69]	Review the impact of AI on cybersecurity.	✓	✗	✗
Truong et al. [37]	Provide an overview of how AI can be used in cybersecurity for offensive and defensive AI.	✓	✗	✗
Kott and Théron [109]	Advance work on AICA and introduce its high-level reference architecture.	✓	✓	✗
Théron and Kott [35]	Advance work on AICA.	✓	✓	✗
Théron et al. [110]	Introduce the concept and architecture of an autonomous intelligent cyber-defence agent (AICA).	✓	✓	✗
Grzonka et al. [111]	Present a MAS cloud monitoring model.	✗	✓	✗
Yampolskiy [112]	Explains AI safety vs cybersecurity.	✗	✓	✗
Petrović [113]	Explores AI agents in virtual worlds.	✗	✓	✗

A significant research gap is evident in the literature, as only three studies, indicated in Table 2, have examined the convergence of cybersecurity, AI agents, and the NIST CSF. [Pleshakova et al. \[99\]](#) briefly model attackers by neural networks, highlighting the capabilities of decentralised LLMs, or autonomous LLM agent swarms, into distinct cyber operations categories aligned with four functions of the NIST CSF 1.1 (Identify, Protect, Detect, Respond). Notwithstanding that the current paper concerns the NIST CSF 2.0, [Pleshakova et al. \[99\]](#) do not map any AI agent capabilities to the NIST CSF functions. [Hauptman et al. \[100\]](#) further narrowed the scope, concentrating solely on the Response function of the NIST CSF 1.1. Notably, [Kaur et al. \[101\]](#) recognised the importance of governance in cybersecurity preceding the release of NIST CSF 2.0. They astutely argued for AI's role in policy enforcement and risk monitoring, aligning with the framework's eventual inclusion of the 'Govern' function. However, as shown in Fig. 4, [Kaur et al. \[101\]](#) specification of AI-driven techniques was ultimately grounded in the outdated NIST CSF 1.1, missing critical updates and changes in NIST CSF 2.0.



**Fig. 4** AI domain per the NIST CSF 1.1 function: Shows how AI-based techniques map to five original NIST CSF 1.1 functions (Identify, Protect, Detect, Respond, Recover) before the addition of the Govern function

A further limitation in Kaur et al. [101] approach is their reliance on terms like 'AI-based,' 'AI-powered,' and 'AI can automate,' suggesting a focus on automation or assisted intelligence (refer to Fig. 3). This contrasts with the exploration of AI agents (autonomous intelligence), a critical aspect of this paper. The phrase 'AI agent' is used only once in their work. While the AI automation techniques in Fig. 4, such as planning, reasoning and learning, exhibit some overlap with basic AI agent properties discussed in the introduction section and elaborated upon further in Section 2.2.3, the study, like Pleshakova et al. [99] and Hauptman et al. [100], did not address the NIST CSF 2.0. Therefore, the systematic mapping of AI agents to the current NIST CSF 2.0 remains a conspicuous research gap that requires further investigation.

## 2.2 Foundations of AI agents

### 2.2.1 Large-language models

Driven by DL architectures and massive pre-training, LLMs have emerged as a transformative force in natural language processing [114]. With transformer-based architectures like OpenAI's Generative Pre-trained Transformer (GPT) series, Google's Gemini series, and Anthropic's Claude series, LLMs showcase unparalleled abilities in capturing and generating human-like text [92], [115]. However, this general-purpose proficiency masks critical limitations. LLMs, without customisation, struggle with domain-specific knowledge and proprietary data, making them unsuitable for many real-world applications [114]. The prohibitive cost and resource demands of training LLMs from scratch further necessitate the development of tailored solutions [116]. As a result, the field has witnessed an explosion of customisation strategies, broadly falling into two distinct categories, designed to adapt LLMs to specific application contexts:

- *Parameter-efficient fine-tuning (PEFT) or frozen model adaptation:* PEFT techniques adapt pre-trained LLMs by training only a small subset of newly introduced parameters, preserving/freezing the original model's weights and significantly reducing computational costs [116].
- *Full fine-tuning:* In contrast, full fine-tuning modifies all parameters of a pre-trained LLM to optimise performance on a specific task, demanding more computational resources and potentially altering the model's general knowledge [117].

After selecting a foundational LLM, a customisation strategy is required to facilitate the adaptation of the LLM for specialised applications. This paper presents a spectrum of five prevalent customisation strategies, ordered by progressively increasing resource demands and computational expenditure:

- *Prompt engineering (Strategy: PEFT-adjacent)*: This strategy, which includes in-context learning and requires minimal resource investment, focuses on carefully crafting input prompts to elicit desired LLM responses, effectively steering the model's inherent capabilities without parameter modification [118].
- *Retrieval augmented generation (RAG) (Strategy: PEFT-adjacent)*: Employing moderate resource consumption, RAG augments LLM responses with external knowledge retrieval, enhancing accuracy and relevance by incorporating real-time information [119].
- *Agent frameworks (Strategy: PEFT-adjacent/Hybrid)*: These agentic frameworks, with increasingly resource-intensive demands, enable LLMs to interact with external environments and tools, facilitating the execution of complex tasks through dynamic interaction behaviour [71]. This approach can be seen as a hybrid, as some agents may modify small parameters while others do not.
- *Fine-tuning (Strategy: Full fine-tuning)*: This strategy involves significant resource allocation to adapt LLM parameters using domain-specific datasets, resulting in enhanced performance for targeted applications [117], [118].
- *Reinforcement learning from human feedback (RLHF) (Strategy: Full fine-tuning/Hybrid)*: Requiring maximum resource expenditure, RLHF refines LLM behaviour through human preference-based learning, aligning the model's outputs with desired human values and preferences [119]. This also has hybrid qualities, as some implementations only modify the "reward" model, while others modify the base LLM.

Customisation strategies designated as 'PEFT-adjacent/Hybrid' represent approaches that, while not strictly adhering to standard PEFT methodologies, either share core principles of minimising pre-trained weight modifications or integrate aspects of PEFT and full fine-tuning. Similarly, the designation 'Full fine-tuning/Hybrid' indicates strategies that predominantly rely on full fine-tuning, modifying all model parameters, but may also integrate aspects of parameter-efficient methods or possess implementations that vary, resulting in a combination of full and parameter-efficient fine-tuning characteristics.

This paper focuses on agent frameworks, a customisation strategy that enables LLMs to interact with external environments and tools. On November 25, 2024, Anthropic [120] introduced the Model Context Protocol (MCP), a new standard for connecting AI models to diverse external data systems, including business tools, content repositories, and development environments. Thus, the MCP is an open standard designed to simplify the integration of AI agents with real-world data, providing context to underlying LLMs [121]. Additionally, in collaboration with a wide range of industry partners (Deloitte, Langchain, Salesforce, Cohere, and fifty-six others), Google's launch of the Agent2Agent (A2A) protocol on April 9, 2025, an open standard that complements the MCP, is poised to accelerate the evolution of interconnected AI agents by enabling cross-vendor interoperability between agents [122]. A good understanding of AI agents necessitates thoroughly examining their core principles, inherent properties, and the range of agent types.

### 2.2.2 Defining AI agents

Building on the discussion of LLM customisation strategies, this paper explores the critical role of AI agents within agentic frameworks. To contextualise this, it is essential to clarify the concept of 'agency' in AI systems. 'Agency' denotes the extent to which an AI system's behaviour is goal-directed, directly impacts its environment, and enables it to achieve long-term objectives with minimal human intervention [71]. This implies a shift from passive AI tools to active decision-making entities. 'Agentic AI,' therefore, represents the overarching framework that empowers AI agents to operate with heightened autonomy. It encompasses the broader capabilities that enable LLMs to function as agents, facilitating complex interactions and decision-making processes [123]. In this context, an AI agent is a specific instantiation or component within an agentic AI system designed to execute particular tasks or functions. This distinction between the broader agentic AI framework and the specific AI agent implementation is crucial for understanding the nuanced application of LLMs in autonomous systems. Recognising that agentic AI is a system and an AI agent is a component within that system is very important.

While scholarly definitions of AI agents vary, they converge on the fundamental concept of systems that perceive their environment and autonomously execute actions to optimise goal achievement [25]. These systems span from basic rule-based programs to sophisticated learning and reasoning entities [30]. Despite definitional nuances, the core tenets of AI agents consistently emphasise their capacity for independent and intelligent operation within a given environment [26], [27]. Table 3 summarises AI agent definitions according to several authors.

**Table 3** AI agent definitions: Provides various scholarly definitions of AI agents, capturing themes such as autonomy, adaptability, and environment interaction

AI agent definition	Author
A programmed entity that performs operations on behalf of another user or program with a certain degree of independence.	Alrfai et al. [124]
Systems capable of executing tasks without human intervention.	Dehghantanha et al. [68]
Software or a collection of software that resides and operates on one or more computing devices, perceives its environment, and executes purposeful actions on the environment (and on itself) to achieve the agent's goals.	Kott et al. [36]
Entities that can perform certain functions without human intervention, including self-activating, self-sufficiency, and persistent computation.	Ligo et al. [38]
An entity that is a mixture of hardware and software and uses sensors to perceive its surroundings and actuators to make changes.	Sharma and Jindal [40]
Pieces of software or hardware with a processing unit capable of making wise decisions about their courses of action in uncertain and adverse environments.	Théron and Kott [35]
Computer-assisted systems that can generate text, images, audio, or videos.	Baabdullah [95]
An entity with self-adaption and intelligence in an environment can understand its environment and control its decision behaviour; it conducts self-regulation and self-learning after it perceives changes in the environment, and it can proactively complete preset tasks.	Cao et al. [107]
Systems capable of pursuing complex goals with limited supervision.	Chan et al. [71]
Systems or software that can perform tasks or functions independently, without human intervention, in various environments and situations are designed to make decisions and take actions based on their programming and the data they receive from their surroundings.	Cronin [72]
Autonomous systems that interact with their environment to achieve specific objectives, often using LLMs to enhance natural language processing capabilities.	Han et al. [92]
An intelligent entity that can autonomously execute appropriate and contextually relevant actions based on sensory input, whether in a physical, virtual, or mixed-reality environment.	Huang et al. [114]
Systems that perceive their environment and take actions to maximise their chances of achieving their goals.	Rossi [25]
Software programs that respond to states and events in their environment, independent from direct instruction by the user or owner, acting on behalf and in the owner's interest.	Bösner [30]

As evidenced by the diverse definitions presented in Table 3, the absence of a unified conceptualisation of AI agents has resulted in a proliferation of varied architectural and methodological approaches. This lack of consistency directly impacts the design and implementation of agentic systems, necessitating a thorough examination of the field's dominant trends and ongoing projects to understand the resulting complexities in modelling and application [26]. Using the various definitions in Table 3, I asked ChatGPT 4.5 to analyse themes and patterns and construct a single definition. The exact prompt is as follows:

*I have compiled a collection of AI agent definitions from various scholarly research papers. Please analyse the following definitions, identify the recurring themes and patterns, and construct a unified definition that encapsulates the core concepts shared across these definitions. Aim for a definition that is both comprehensive and concise.*

*Definitions: [I inserted my list of AI agent definitions from Table 3 here, each clearly separated.]*

*Specifically, please focus on identifying:*

- *Common elements: What core capabilities or characteristics are consistently mentioned?*
- *Variations: Where do the definitions differ, and what are the key points of divergence?*
- *Essential components: What elements are indispensable for a system to be considered an AI agent?*

*Based on your analysis, provide a unified definition of an AI agent that reflects the essence of these definitions.*

The resulting definition is as follows: *An **AI agent** is an autonomous computational entity—comprised of software or hardware and software integration—that independently perceives its environment, intelligently processes information, makes contextually appropriate decisions, and proactively executes purposeful actions to achieve defined goals, adapting dynamically through self-regulation and learning to changing circumstances.*

Although AI agent properties are explored in detail in the next section, the ChatGPT prompt above also yielded core AI agent characteristics emanating from the studies in Table 3. These are:

- *Autonomy: Capacity to operate independently.*

- *Environmental perception*: The capability to gather and interpret environmental data.
- *Decision-making*: Ability to independently select appropriate actions or behaviours.
- *Goal-oriented action*: Pursuit and achievement of specified objectives through purposeful interactions.
- *Adaptability and learning*: The capability for self-adaptation and learning.

In addition to the above, the following section provides a thorough exploration of key AI agent properties.

### 2.2.3 AI agent properties

AI agents possess several key properties that enable them to interact with their environment and make decisions to achieve specific goals. A literature review reveals several properties, as shown in Table 4.

**Table 4** AI agent properties: Lists the key properties of AI agents (e.g., autonomy, learning) identified through a literature review, with brief descriptions and source references

Core property	Description	Author
Autonomy	Operate independently without human intervention.	Phillips-Wren, Leite et al., Bösser, Gu and Li, and Liu et al. [22], [23], [30], [125], [126]
Adaptive learning	Learn from past experiences and adapt to new situations.	Rabuzin et al., Leite et al., Abiodun and Khuen, Sethy et al., and Mazumder and Liu [22], [24], [26], [127], [128]
Proactive goal pursuit	Take initiative to achieve goals.	Phillips-Wren, Leite et al., Jameel et al., Liu et al., and Pantoja et al. [22], [23], [125], [129], [130]
Reactive responsiveness	Respond to environmental changes in real-time.	Phillips-Wren, Leite et al., Jameel et al., Liu et al., and Pantoja et al. [22], [23], [125], [129], [130]
Inter-agent communication	Share information and coordinate actions with other agents.	Phillips-Wren, Leite et al., Jameel et al., Liu et al., Pantoja et al., and Sethy et al. [23], [125], [128], [129], [130]
Collaborative interaction	Collaborate in MAS to achieve shared goals.	Leng et al., Phillips-Wren, Jameel et al., Liu et al., and Sethy et al. [23], [125], [128], [129], [131]
Knowledge representation and reasoning	Store and utilise knowledge for decision-making.	Sennott et al. [132]
Ethical reasoning and decision-making	Embed ethical decision-making capabilities into AI systems.	Bösser, Rossi and Mattei, and Gu and Li [30], [126], [133]
Moral agency (where applicable)	Ensure actions by agents with significant autonomy and decision-making align with moral codes and societal norms.	Bösser, and Gu and Li [30], [126]
Social interaction	Interact with humans and other agents through various interfaces to influence their decisions, confidence, and trust through transparency and reliability.	Nickles et al., Phillips-Wren, Leite et al., Rossi and Mattei, Pitardi and Marriott, Chakraborty, He and Jazizadeh, and Peng et al. [19], [22], [125], [133], [134], [135], [136], [137]
Trustworthiness	Build user trust through reliability and transparency.	Pitardi and Marriott, and He and Jazizadeh [135], [137]
Domain-specific competence	Applications in various fields, such as smart homes, healthcare, and military operations.	Bösser, Rossi, Preethiya et al., Rashid and Kausik, and Liu et al. [25], [30], [32], [138], [139]
Explainability	Provide clear and understandable explanations for actions and decisions.	Chakraborty, Majumdar, and Simran et al. [136], [140], [141]
Resilience	Maintain performance and functionality in the face of unexpected inputs, noise, or adversarial attacks.	Cam. Falowo et al., and Rafferty and Macdermott [142], [143], [144]
Embodiment (if applicable)	In applicable instances, build a physical presence or virtual representation in an environment, enabling it to interact with the world through sensors and actuators.	Abiodun and Khuen, Weng and Ho, Bovo et al., Preethiya et al., and Liu et al. [31], [32], [94], [127], [139]
Long-term memory	Retain and utilise information over extended periods.	Chan et al., Kim and Saad, Chen et al., and Deng et al. [71], [93], [98], [123]
Agency transfer/Delegation	Transfer or delegate parts of the agency or decision-making to other agents or humans.	Neff and Nagy, Baird and Maruping, and Candrian and Scherer [65], [145], [146]

As shown in Table 4, AI agents' functionality is enhanced by various core properties, including autonomy, adaptability, and reasoning [23], [30]. While functionalities such as planning, goal-oriented behaviour, and explicit decision-making are undeniably crucial for AI agent operations [28], [30], [127], my review in Table 4 prioritises the examination of broader, foundational properties that underpin these capabilities. These specific functionalities are often observed as emergent behaviours or direct applications of core attributes like autonomy, adaptive learning, and (knowledge representation and) reasoning. For example, sound reasoning inherently enables informed decision-making [30], [132], and planning is intrinsically linked to goal-oriented actions [29], both of which are facilitated

by a high degree of autonomy [41], [147]. Furthermore, domain-specific competence integrates and applies these foundational properties within context. In addition, properties such as explainability, resilience, and long-term memory are becoming increasingly relevant as AI agents are deployed in more complex and critical applications [123], [140], [142]. Therefore, this paper focuses on the core properties—autonomy, adaptive learning, proactive goal pursuit, reasoning, and others detailed in Table 4—to provide a framework for understanding the fundamental principles that govern AI agent behaviour while acknowledging the inherent interconnectedness and hierarchical relationships among these attributes.

However, semantic gaps in communication languages challenge the development of AI agents with these properties, and there is a need for firm ethical boundaries [133], [134]. In other words, ethical considerations like fairness, data privacy and security, and transparency are essential for responsible AI development [128], [129], [136]. As Radanliev et al. [148] articulate, autonomous AI systems must be evaluated not only for their technical performance and alignment with established frameworks, such as the NIST CSF, but also, crucially, for their potential to heighten cyber risks, compromise privacy, or operate beyond adequate human control. These profound concerns fundamentally reshape the discourse, underscoring the imperative for agentic frameworks that inherently embed ethical safeguards and responsible AI principles throughout the entire deployment lifecycle.

The limitations of AI agents extend beyond the absence of a unified definition, current technological constraints, and ethical considerations. Notably, legal and regulatory frameworks present a substantial challenge, particularly concerning the classification of autonomous AI systems. Whether such systems should be granted legal personhood rather than being treated as mere objects is a subject of ongoing debate [149]. However, a thorough examination of these legal implications is beyond the scope of this paper. Building on the discussion of AI agent properties and their inherent limitations, the following section provides an overview of their classification, highlighting the diverse types and their respective applications.

#### 2.2.4 AI agent types

A clear differentiation based on their functional and task-specific architectures is required to understand the diverse applications of AI agents. In the same SLR exercise described in Section 2.1.3, one of the themes extracted from the qualitative data was under the category of ‘AI agent types.’ I asked ChatGPT 4.5 to analyse this ‘AI agent type’ data, and the exact prompt is as follows:

*You are an expert in artificial intelligence and agent-based systems, tasked with analysing the findings of a systematic literature review (SLR) on AI agent types. I will provide you with a series of notes extracted from the SLR, describing various entities that authors have identified as AI agents. Your task is to:*

1. *Categorise the data into broader, more accepted 'AI agent type' categories. These categories should reflect the core, fundamental distinctions in AI agent design and functionality established by reputable academic and industry sources.*
2. *Identify and describe any subcategories or variations mentioned in the notes. Explain how these subcategories relate to the main categories, highlighting their specific characteristics and differences.*
3. *Acknowledge and clarify any instances where authors may refer to specific implementations, applications, or variations rather than distinct agent types. Explain the nuances of these distinctions.*
4. *Present your analysis in a clear, structured table format. The table should include:*
  - *'Main AI Agent Category'*
  - *'Subcategories/Variations'*
  - *'Description and Nuances'*
5. *Provide a concise summary of the key findings and trends identified in the SLR data.*

*Please ensure your analysis is grounded in established AI agent literature and reflects a comprehensive understanding of the field.*

*[Insert your notes from the SLR here. In my case, I attached a document with my SLR notes.]*

Table 5 presents a structured analysis of the SLR data, organised into AI agent categories, including subcategories and nuances that are clarified.

**Table 5** AI agent types: Classifies AI agent architectures (e.g., reactive, cognitive) and subcategories or variations, showing how each addresses distinct design and operational nuances

AI agent category	Subcategories/Variations	Description and nuances
Autonomy-based agents	<ul style="list-style-type: none"> <li>○ Autonomous agents</li> <li>○ Adaptive autonomous agents</li> <li>○ Continual learning-enabled AI agents (learning agents)</li> </ul>	Agents perform tasks independently, with autonomous agents operating without humans [41], [150]. Adaptive autonomous agents dynamically adjust autonomy levels according to contextual requirements [100]. Learning agents continuously self-learn, adapting to novel situations autonomously [23], [24].
Interaction complexity (Agent cognition)	<ul style="list-style-type: none"> <li>○ Simple reflex agents</li> <li>○ Model-based reflex agents</li> <li>○ Goal-based agents</li> <li>○ Utility-based agents</li> <li>○ Cognitive agents</li> </ul>	Categorised by cognitive complexity, simple reflex agents react directly to current percepts without history. Model-based reflex agents maintain internal states by tracking environmental aspects [151]. Goal-based agents explicitly pursue defined objectives, guiding their actions. Utility-based agents select actions that maximise outcomes' desirability or utility [151]. Cognitive or deliberative agents are intelligent agents that aim to emulate human-like cognitive processes [152]. They are responsible for higher-level tasks such as learning, planning, conflict resolution, and task management [28].
Embodiment	<ul style="list-style-type: none"> <li>○ Embodied agents</li> <li>○ Virtual agents</li> </ul>	Embodied agents possess physical forms (such as robots and drones) capable of physical actions [31], [32]. Virtual agents exist entirely as software, operating digitally, such as chatbots and virtual assistants [31].
Behavioural approach	<ul style="list-style-type: none"> <li>○ Reactive agents</li> <li>○ Proactive (Goal-oriented) agents</li> <li>○ Hybrid agents</li> <li>○ Learning agents</li> </ul>	Reactive agents respond directly to stimuli without using internal states or past experiences [29]. Proactive agents utilise internal states, planning, and past experiences to achieve goals [29]. Hybrid agents incorporate both reactive and proactive functionalities, dynamically adjusting to their surroundings [29].
Collaboration and coordination	<ul style="list-style-type: none"> <li>○ Cooperative agents</li> <li>○ Centralised agents</li> <li>○ Decentralised agents</li> <li>○ Multi-agent systems (Swarm, Cohort, Society of agents)</li> <li>○ Autonomous collaborative agents</li> </ul>	Cooperative agents collaborate toward common goals in MASs [150], [153]. Centralised agents are centrally coordinated, whereas Decentralised agents coordinate independently through communication [153]. MASs demonstrate collective behaviours through horizontal (swarm) or vertical (cohort) coordination, enabling emergent properties and self-organisation [35], [36], [110]. Autonomous collaborative agents independently perform complete tasks, interacting with others when necessary [35], [36].
Ethics and morality	<ul style="list-style-type: none"> <li>○ Artificial moral agents</li> </ul>	Designed explicitly to facilitate ethical decision-making, adhering to moral standards is crucial in ethically sensitive applications such as healthcare or autonomous vehicles [126].
Architectural paradigms	<ul style="list-style-type: none"> <li>○ Thinking-type architecture (Symbolic/Decision-based)</li> <li>○ Response-type architecture (Event-driven)</li> <li>○ Mixed-type architecture (Hybrid)</li> </ul>	<i>Thinking-type agents execute actions through symbolic reasoning and decision-making processes. Response-type agents primarily react to external perceptions or events. Mixed-type agents combine symbolic reasoning and reactive perception-driven behaviour [107].</i>
Agent implementation contexts	<ul style="list-style-type: none"> <li>○ Specialised agents (Functional specialisation)</li> <li>○ Fully functional agents (Holistic capability)</li> </ul>	<i>Specialised agents focus on performing discrete functional roles as part of broader collaborative agent societies. Fully functional agents independently execute all necessary functions for their designed purpose and are capable of autonomous operation [35], [36].</i>
Specific implementations and applications (Contextual examples)	<ul style="list-style-type: none"> <li>○ Self-driving cars</li> <li>○ Drones</li> <li>○ Robotic vacuum cleaners</li> <li>○ Chatbots and virtual assistants</li> <li>○ Text-based generative agents</li> <li>○ Automated trading systems</li> <li>○ Robotic space exploration probes</li> <li>○ Internet information agents</li> </ul>	<i>Contextual examples demonstrating practical implementations utilising core agent characteristics (autonomy, cognition, embodiment). Examples: autonomous vehicles, drones, home robots, automated financial trading [72], robotic space exploration [40], conversational AI and generative text-based agents [72]. Such examples reflect real-world applications rather than distinct agent types.</i>

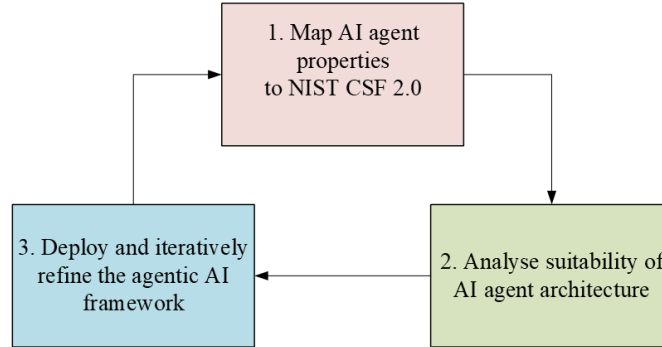
As evidenced through Table 5, the SLR identifies distinct AI agent categories based on autonomy, cognitive complexity, embodiment, behavioural approach, coordination structure, ethical considerations, and architectural design. Trends indicate a growing emphasis on adaptive autonomy, continual learning capabilities, ethical and moral agency, and multi-agent collaboration for complex problem-solving. Contextual examples provided in the literature, listed in the last row of Table 5, emphasise practical applications rather than new agent categories, highlighting how theoretical constructs of AI agents manifest in varied real-world scenarios. It is clear from Table 5 that not all agent



types are equally suited to every security task. The framework proposed in this paper aims to provide a systematic and contextually relevant approach to selecting AI agents suitable for cybersecurity tasks.

### 3. Method for developing an AI agent selection framework

Driven by the need for a structured approach to AI agent selection, design, and deployment, informed by a systematic categorisation of agent types and their characteristics, I present the AI Agent Taxonomy and Decision Framework (AIATDF). This framework, grounded in literature-derived agent properties and categories, provides a decision-making tool for mapping suitable AI agents to the NIST CSF 2.0. The framework development method, guided by the AI agent properties in Table 4, the architecture/types in Table 5, and the NIST CSF 2.0 core structure in Fig. 2, is visually represented in Fig. 5.



**Fig. 5** AIATDF development method: Outlines the overarching method employed in the study for the development of the AI agent taxonomy and decision framework

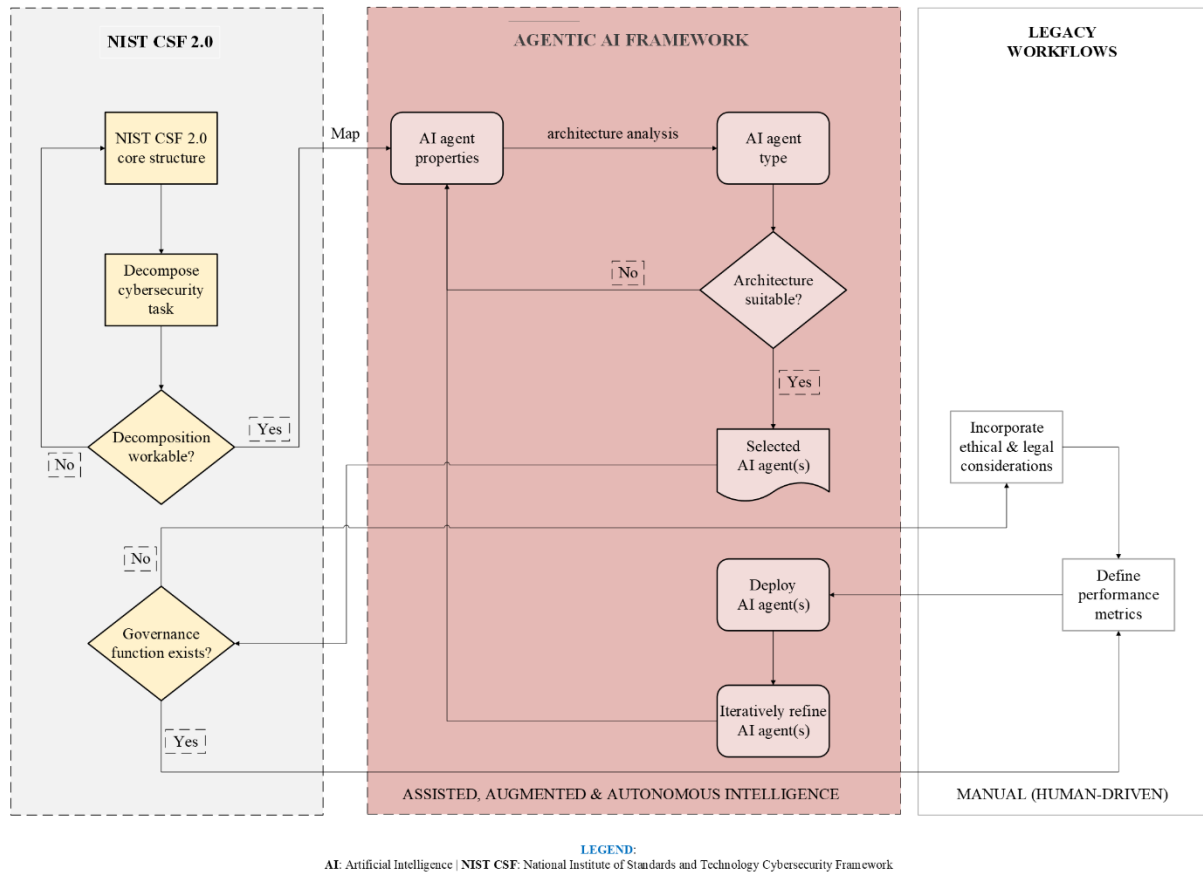
Essentially, Fig. 5 says the following:

- 1) *Establish agent property-NIST CSF 2.0 task alignment:*
  - Map AI agent properties to specific NIST CSF 2.0 task requirements.
  - Method: Develop a comprehensive mapping matrix.
- 2) *Conduct architectural suitability assessment:*
  - Evaluate the architectural suitability of various AI agent types for NIST CSF 2.0 task requirements.
  - Method: Perform a comparative analysis of agent architectures.
- 3) *Implement and iteratively refine agent deployment:*
  - Deploy selected agents and establish a continuous refinement process.
  - Method: Implement a feedback-driven iterative optimisation cycle.

Executing the steps above yields the AIATDF presented and explained in the next section.

### 4. AI agent taxonomy and decision framework

In the AIATDF description, ‘taxonomy’ emphasises the systematic categorisation of agent types derived from the literature, and ‘decision framework’ indicates a structured approach for guiding the selection, design, and deployment decisions of AI agents based on identified agent categories, variations, characteristics, and application contexts. Following the three framework development steps outlined in the methods section, Fig. 6 represents the AIATDF of this paper.



**Fig. 6** AI agent taxonomy and decision support framework: Provides an overview of the six main stages of the proposed AIATDF, showing how each step contributes to choosing and deploying AI agents aligned with the NIST CSF 2.0

The AIATDF comprises six main stages as follows:

- 1) *Contextual cybersecurity task decomposition:*
  - Break down complex cybersecurity tasks into granular subtasks within the NIST CSF 2.0.
  - *Method:* Hierarchical task analysis of the NIST CSF 2.0's subcategories through the 'informative references' functionality.
- 2) *Mapping agent properties to NIST CSF 2.0:*
  - Link AI agent properties to the NIST CSF 2.0 cybersecurity task-specific requirements.
  - *Method:* Create a mapping matrix.
- 3) *AI agent-type architectural suitability analysis:*
  - Assess AI agent architecture suitability for NIST CSF 2.0 cybersecurity task-specific requirements.
  - *Method:* Comparative analysis of agent architectures.
- 4) *Performance evaluation metrics for agent effectiveness:*
  - Define metrics to assess AI agent(s) performance in NIST CSF 2.0 objectives.
  - *Method:* Use industry best practices to define key performance indicators.
- 5) *Design, develop, and deploy AI agent(s)*
  - Create and implement AI agent solutions that meet the architectural and performance requirements identified in the previous stages.
  - *Method:* Use established software development practices (e.g., Agile or DevSecOps) and pilot testing to integrate the chosen agent architectures into the organisation's SOC or equivalent cybersecurity workflow.
- 6) *Iterative framework refinement and validation:*
  - Establish a feedback loop for continuous improvement of the agentic AI framework.
  - *Method:* Collect data and feedback from cybersecurity practitioners and end-users.

In addition to the six main stages of the AIATDF, there are two extra conditional stages in Fig. 6. These are the ‘governance decision’ and ‘ethical and legal considerations’ stages. As discussed in the literature review section, the NIST CSF 2.0 has a ‘govern’ function that ensures the integration of ethical and legal considerations, the ‘yes’ option of the ‘governance decision’ box in Fig. 6. The ‘no’ governance option in the AIATDF demonstrates that any acceptable cybersecurity framework can be utilised in place of the NIST CSF 2.0, and the agentic AI framework will still be applicable. In other words, any widely recognised cybersecurity standards, guidelines, and frameworks can be utilised with the AIATDF. In the next section, I use the NIST CSF 2.0 to validate the AIATDF.

## 5. Validation of the AIATDF

To establish the validity of the AIATDF—specifically, its ability to represent its intended construct accurately [154]—this test focuses on mapping AI agent properties to six core functions of the NIST CSF 2.0 (Fig. 2). Notably, the initial hierarchical task analysis stage of the AIATDF, which decomposes the NIST CSF 2.0 (sub)categories via the framework’s ‘informative references,’ falls beyond the scope of this agentic AI framework (Fig. 6) and is therefore not validated in this paper. Instead, the validation process commences with the crucial step of aligning AI agent properties (Table 4) to the NIST CSF 2.0 functions. As stage 2 of the AIATDF dictates, this alignment must be documented in a mapping matrix, which directly links AI agent characteristics to specific cybersecurity task-specific requirements (NIST CSF 2.0 functions in this validation exercise). Furthermore, the AIATDF validation test recognises that AI agent types (Table 5)—which directly influence the agentic AI framework architecture—are derived from these properties. Therefore, the validation extends to stage 3, which assesses the suitability of AI agent architectures for the NIST CSF 2.0 task-specific cybersecurity requirements, a process also reflected in the mapping matrix.

For demonstrative purposes, the mapping matrix (Table 6) showcases the application of reactive, cognitive, hybrid, and learning AI agent types; however, it is emphasised that this mapping strategy is universally applicable across all AI agent types. In other words, Table 6 validates a simplified version of the AIATDF conceptual framework. Thus, real-world deployments may use MAS [92] for a single cybersecurity function or combine multiple functions within the same agentic system.

**Table 6.** Mapping matrix for selection, design, and deployment of AI agents: Aligns four AI agent types—reactive, cognitive, hybrid, and learning—with the six NIST CSF 2.0 functions, offering a practical guide to optimal agent deployment

	Reactive agents	Cognitive agents	Hybrid agents	Learning agents
<b>Govern</b>	<ul style="list-style-type: none"> <li>Enforces static policy checks in real-time</li> <li>Limited scope for policy evolution</li> </ul>	<ul style="list-style-type: none"> <li>Policy-driven decision-making &amp; compliance management</li> <li>Tracks strategic goals, compliance, risk appetite</li> </ul>	<ul style="list-style-type: none"> <li>Ensures immediate policy adherence and higher-level oversight</li> <li>Facilitates cross-functional governance tasks</li> </ul>	<ul style="list-style-type: none"> <li>Learning-based governance optimisation (e.g., balancing privacy vs. security trade-offs)</li> <li>Automates policy refinement for dynamic changes</li> </ul>
<b>Identify</b>	<ul style="list-style-type: none"> <li>Rapid scanning of known assets and threats</li> <li>Trigger-based checks for new devices or vulnerabilities</li> </ul>	<ul style="list-style-type: none"> <li>Strategic planning for comprehensive asset inventories</li> <li>Incorporates organisational risk models and compliance</li> </ul>	<ul style="list-style-type: none"> <li>Balances immediate scanning with higher-level threat modelling</li> <li>Layered architecture for real-time triggers and deeper analysis</li> </ul>	<ul style="list-style-type: none"> <li>Adapts risk profiles over time via continual learning</li> <li>Discovers unknown vulnerabilities from historical patterns</li> </ul>
<b>Protect</b>	<ul style="list-style-type: none"> <li>Immediate enforcement of access rules</li> <li>Real-time reactive lockdowns under suspicious activity</li> </ul>	<ul style="list-style-type: none"> <li>Policy-driven configuration management</li> <li>Plans system-wide changes and resource allocations</li> </ul>	<ul style="list-style-type: none"> <li>Integrates rules-based triggers with strategic access control</li> <li>Quick local responses plus deliberative oversight</li> </ul>	<ul style="list-style-type: none"> <li>Uses anomaly detection to refine protective measures</li> <li>Learns evolving user access patterns for dynamic policy updates</li> </ul>
<b>Detect</b>	<ul style="list-style-type: none"> <li>Real-time intrusion triggers via signature-based checks</li> <li>High-speed event-driven anomaly alerts</li> </ul>	<ul style="list-style-type: none"> <li>Considers multi-step reasoning for stealthy threats</li> <li>Goal-based analysis of suspicious patterns</li> </ul>	<ul style="list-style-type: none"> <li>Fast detection loop (reactive layer)</li> <li>Complex threat correlation (deliberative layer)</li> </ul>	<ul style="list-style-type: none"> <li>Adaptive models refine detection thresholds over time</li> <li>Identifies zero-day exploits using ML or RL</li> </ul>
<b>Respond</b>	<ul style="list-style-type: none"> <li>Automated containment (block IP, isolate assets), minimal strategic foresight but instant action</li> </ul>	<ul style="list-style-type: none"> <li>Deliberate coordination of multi-step mitigation plans</li> <li>Accounts for long-term system impacts</li> </ul>	<ul style="list-style-type: none"> <li>Instant quarantines combined with policy-based orchestration</li> <li>Layered approach to reduce escalation time</li> </ul>	<ul style="list-style-type: none"> <li>RL to optimise response effectiveness</li> <li>Learns from each incident to improve future mitigations</li> </ul>

<b>Recover</b>	<ul style="list-style-type: none"> <li>○ Scripted fallback to restore systems quickly</li> <li>○ Lacks deeper analysis of root causes</li> </ul>	<ul style="list-style-type: none"> <li>○ Strategic resource allocation for restoration</li> <li>○ Coordinated plan to rebuild or reconFig. critical infra</li> </ul>	<ul style="list-style-type: none"> <li>○ Rapid initial recovery plus holistic follow-up</li> <li>○ Reacts quickly yet integrates deliberative analysis for improvement</li> </ul>	<ul style="list-style-type: none"> <li>○ Adaptive post-incident learning (root-cause analysis)</li> <li>○ Uses data from each breach to refine future recovery blueprints</li> </ul>
----------------	--	--	---	--

As shown in Table 6, each row corresponds to one of the six NIST CSF 2.0 functions, and each column highlights which and how selected AI agent types can address the cybersecurity task-specific requirements linked to each function. The bullet points summarise the key capabilities or actions of AI agents. The AI agent capabilities are derived from Table 5. For instance, reactive agents rely on stimulus-response mechanisms and thus react to environmental changes in near real-time [29]. Therefore, they would suit any trigger-based or event-driven cybersecurity activities in the NIST CSF 2.0. The cognitive agents are responsible for higher-level functions such as planning, reasoning, learning, and task management [28], [152]. They would be appropriate for NIST CSF 2.0 cybersecurity tasks that require coordination, strategic planning, multi-step reasoning, and decision-making. Hybrid agents combine all types of AI agents, as shown in Table 6. For instance, they combine reactive and cognitive strategies, typically arranged in layered architectures, so quick reactive behaviours can coexist with higher-level reasoning processes [22], [29]. They are, therefore, suitable for cybersecurity activities that require immediate reaction to events and strategic oversight that requires continuous planning, reasoning and decision-making. Lastly, learning agents can enhance performance by revising internal models based on new data or feedback loops [41], [150], making them well-suited for ever-evolving threat landscapes. I discuss the implications of the AIATDF in the next section.

## 6. Discussions

The proposed framework of the study, the AIATDF, was presented in Fig. 6 and conceptually validated in Table 6. The AIATDF demonstrates how AI agents can automate (assisted intelligence), autonomize (autonomous intelligence), or combine automation and autonomy (augmented intelligence) to meet the NIST CSF 2.0 cybersecurity task-specific requirements.

### 6.1 Implications of the proposed framework

#### 6.1.1 Theoretical insights

This paper introduces a framework that aligns core AI agent capabilities, such as autonomy, adaptive learning, proactivity, and reactivity, with the six fundamental functions of the NIST CSF 2.0. This integration establishes a conceptual bridge, translating abstract AI agent theory into actionable cybersecurity strategies defined by the NIST CSF. In contrast to fragmented, single-capability AI deployments, this holistic framework provides a unified perspective, defining how diverse agent types can address the complex demands inherent in each NIST CSF 2.0 function. By mapping AI agent capabilities to the task-specific outcomes of the NIST CSF 2.0, the AIATDF mitigates the risks associated with haphazard AI adoption. Rather than deploying advanced AI tools without a clear integration strategy, this approach fosters an alignment between agent theory and NIST CSF constructs. Additionally, this alignment enables the design of AI systems that excel in threat detection and response, adapt to evolving vulnerabilities, engage in strategic planning, and enhance governance. Combining theoretical concepts, such as autonomy and AI agent cooperation, with practical objectives, like reduced incident response times and enhanced compliance, empowers organisations to invest in AI-driven cybersecurity solutions with confidence. Consequently, the AIATDF enhances the theoretical rigour and real-world applicability of agent-based defence strategies, ensuring alignment with recognised cybersecurity standards.

One of the key contributions of this study is that the AIATDF also distinguishes between assisted, autonomous, and augmented modes of AI systems intelligence. This contrasts with traditional agent taxonomies that obscure critical distinctions in agent autonomy through broad, overlapping categories. The AIATDF illuminates the nuanced spectrum of AI agent capabilities by explicitly mapping each intelligence mode to specific cybersecurity functions—ranging from rapid, automated tasks (assisted intelligence) to complete human-out-of-the-loop decision-making (autonomous intelligence). In other words, the AIATDF highlights the graduated spectrum of AI agent capabilities. This clarity is crucial; for instance, it describes why a reactive AI agent may be sufficient for immediate, rule-based detection tasks within the Identify or Protect functions, while a learning or cognitive AI agent is indispensable for strategic governance or comprehensive vulnerability assessments. Consequently, the AIATDF demonstrates that "autonomy" is not a monolithic concept but a dynamic attribute that must be tailored to the specific demands of each security context. This granular approach underscores the importance of aligning AI agent architectures with functional requirements. The AIATDF advances the theoretical understanding of agentic AI operations in cybersecurity by emphasising these roles. It reveals how diverse AI agent architectures can be

synergistically deployed to achieve operational efficiency and adaptive resilience, moving beyond simplistic classifications to a more nuanced, context-aware deployment strategy.

Furthermore, the AIATDF distinguishes itself by treating AI agent architecture and operational modes as intrinsically interdependent dimensions, thereby providing a unified analytical lens for evaluating and deploying AI agents. Traditional approaches often treat these aspects in isolation, focusing either on architectural sophistication or the desired level of automation, neglecting their critical interplay. In contrast, the AIATDF demonstrates that an AI agent's architectural complexity directly dictates its potential for autonomy, and conversely, the intended autonomy influences the required architectural design. For example, while learning agents can unlock advanced autonomous capabilities, they may still operate in assisted or augmented modes when strategic human oversight is essential. This multidimensional perspective compels researchers to transcend simplistic, single-axis classifications, such as "low" versus "high" autonomy, and examine how an AI agent's structural design and functional objectives converge to address specific cybersecurity demands. This integrated approach facilitates the development of more sophisticated theoretical models, situating layered AI agent architectures within established organisational risk management standards, such as the NIST CSF. Consequently, the AIATDF fosters a holistic understanding that connects context-aware AI agent behaviours with the hierarchical needs of enterprise security, spanning real-time defence to strategic governance. This unified perspective bridges the gap between AI agent design and operational deployment, enhancing theoretical and practical applications.

While AI agents promise significant operational efficiency and scalability in cybersecurity, their deployment introduces novel cyber risks, especially within critical infrastructure domains. As Radanliev et al. [148] and Radanliev [76] emphasise, the rapid integration of AI into critical systems necessitates a thorough examination of both ethical risks and their associated governance structures. This study recognises that autonomous and semi-autonomous AI agents, particularly those interacting with sensitive data streams, can amplify adversarial threats such as model inversion attacks, data poisoning, and adversarial learning. Such vulnerabilities jeopardise not only system integrity but also intellectual property and classified information. Consequently, the AIATDF must explicitly integrate principles of responsible AI deployment to pre-empt unintended harms. Radanliev [76] and Radanliev et al. [148] advocate for ethical frameworks that prioritise transparency, human oversight, contextual accountability, and resilience against manipulation—considerations that become acutely critical in high-stakes environments, such as biomedical research, national security operations, and AI-driven policymaking, where the consequences of compromise are profound. Therefore, my proposed maturity-based framework necessitates supplementation with institutional risk assessments, comprehensive model audit trails, and thorough ethical readiness evaluations prior to full-scale deployment. Future enhancements to the AIATDF should embed ethical assurance layers, including XAI protocols, bias detection modules, and accountability matrices, directly into agentic system design, fostering not just technical adequacy but also societal trustworthiness.

### 6.1.2 Practical significance

The AIATDF offers a structured, step-by-step methodology that enables the mapping of AI agent capabilities to the subcategories of the NIST CSF 2.0, thereby empowering SOC managers and cybersecurity architects with a valuable tool. It eliminates the reliance on inefficient trial-and-error approaches or generic AI solutions, allowing practitioners to systematically identify the optimal AI agent type for each subcategory's operational objectives. For example, reactive agents can be strategically deployed for high-volume, rapid-response tasks, such as real-time anomaly detection. In contrast, cognitive agents may be reserved for strategic planning and governance functions. This targeted approach minimises resource waste and efficiently allocates computational infrastructure and staff training. By providing clear guidance, the AIATDF accelerates the deployment of effective AI-driven solutions, optimising their impact by ensuring an alignment between AI agent strengths and specific cybersecurity requirements. Open standards, such as the MCP, can further streamline these deployments by simplifying the integration of LLM-driven agents with diverse enterprise tools and data repositories, thereby expanding the AIATDF's practical reach.

Furthermore, the AIATDF offers scalability, empowering organisations across varying cybersecurity maturity levels to adopt AI-driven capabilities incrementally. The framework facilitates a strategic entry point for resource-constrained teams through assisted intelligence, focusing on automated alerts, anomaly detection, and recommendation systems that enhance existing human oversight. As organisational confidence and resource allocation expand, the AIATDF supports transitioning to more advanced autonomous solutions, including AI agent-driven incident response and strategic threat analysis. This phased approach minimises operational disruption and fosters stakeholder buy-in, ensuring that each increment of autonomy is implemented optimally for maximum impact. The framework also provides a roadmap for continuous advancement, guiding organisations from foundational automation to sophisticated, fully autonomous AI agent deployments. This strategy enables organisations to adapt and evolve their cybersecurity posture in response to the ever-changing threat landscape, ensuring sustained resilience and operational efficiency.

## 6.2 Comparisons with prior work

It was discussed in Section 2.1.3, and shown in Table 2, that only three studies from prior works mapped AI solutions to the NIST CSF. These studies were conducted by Pleshakova et al. [99], Hauptman et al. [100], and Kaur et al. [101]. This paper distinguishes itself from prior research by presenting a detailed, NIST CSF 2.0-aligned framework for the deployment of AI agents in cybersecurity. Unlike earlier works that focused on limited NIST CSF 1.1 functions [99] or explored AI-based methods without systematic mapping, the AIATDF explicitly integrates a broad spectrum of AI agent architectures with all six NIST CSF 2.0 functions, including the critical 'Govern' function. This approach addresses the fragmented scope observed in studies that concentrated on narrower functions or outdated frameworks, and it builds upon research recognising AI's importance in risk management [101] by formalising the mapping of AI agent taxonomies to security tasks. Furthermore, while some studies explored limited AI autonomy within specific NIST CSF functions [100], the AIATDF spans the entire cybersecurity lifecycle, clarifying how varying degrees of agent autonomy—assisted, augmented, and fully autonomous—intersect with NIST CSF 2.0 subcategories. By systematically linking AI agent properties to these subcategories, the AIATDF provides a promising, theory-driven mapping mechanism, a feature notably absent in prior partial alignments. Thus, the AIATDF bridges previously disparate strands of AI agent research, MAS approaches, and the evolving NIST CSF 2.0, offering a unified mapping framework that extends and reinforces the applicability of AI in modern cybersecurity practices.

## 6.3 Limitations of the study

Although the mapping framework is rigorously grounded in peer-reviewed AI agent research and the established NIST CSF 2.0 guidelines, it remains a theoretical construct. While this conceptual foundation provides a strong analytical framework, the AIATDF *lacks empirical validation* within realistic cybersecurity environments. Consequently, critical questions regarding practical performance—such as false positive rates in anomaly detection or mean time to respond during active attacks—remain unanswered. To bridge this gap, pilot studies or proof-of-concept deployments are essential. These empirical investigations would provide invaluable insights, verifying whether each AI agent type delivers the predicted benefits under real-world operational constraints, including limited computational resources and dynamic threat vectors. This real-world validation is crucial for refining the framework, ensuring its applicability and accuracy in reflecting the complexities inherent in cybersecurity operations.

Although the AIATDF provides distinct AI agent categories for analytical clarity, real-world implementations rarely adhere to such *rigid classifications*. AI agent architectures often exhibit fluidity, particularly in hybrid or evolving systems. For instance, an initially reactive agent may progressively integrate learning components, or a cognitive agent may acquire adaptive capabilities over time. Consequently, an organisation's deployment may deviate from these idealised categorisations, reflecting a more dynamic and integrated expression of AI agent behaviours. This inherent overlap does not invalidate the AIATDF's utility. Instead, it underscores the necessity for context-sensitive applications, emphasising that effective cybersecurity solutions frequently necessitate tailored combinations of AI agent functionalities rather than strict adherence to singular agent types. This context-sensitive approach acknowledges the dynamic and complex nature of real-world cybersecurity challenges, demanding a flexible and adaptable integration of AI capabilities.

The rapid convergence of AI innovation and the evolution of sophisticated threat actors necessitate a dynamic approach to this mapping. Specifically, current *AI agent properties may become obsolete* or superseded by emerging paradigms, such as quantum-safe AI algorithms or next-generation RL. Therefore, the AIATDF must be periodically reviewed and updated to maintain its operational relevance. This proactive adaptation should incorporate novel attack vectors, advanced computational models, and evolving industry best practices. Failure to do so risks deploying AI agents that, while theoretically sound in the current context, may prove inadequate in addressing the dynamic and unpredictable nature of future cyber threats. This continuous evolution is crucial to ensuring that AI-driven cybersecurity solutions remain effective and resilient in the face of emerging challenges.

While the AIATDF provides a promising technical blueprint for aligning AI agent capabilities with specific cybersecurity tasks, it does not explicitly address *critical organisational factors* that influence the sustainable success of AI-driven initiatives. These could be placed under 'legacy workflows' in Fig. 6, and their impact on the AI agent deployment studied. For example, these factors may include organisational culture, budgetary constraints, and staff readiness. In this regard, even a perfectly aligned learning AI agent may fail to deliver its intended value without organisational buy-in, adequate training budgets, or a workforce ready to collaborate effectively with autonomous systems. Consequently, organisations must conduct readiness assessments and develop change management plans to foster internal acceptance and ensure staff possess the requisite skills and tools for seamless integration with AI agents. Without these preparatory measures, even the most technically sophisticated AI deployments risk underutilisation, resistance, or premature abandonment. This underscores the imperative that

technological excellence must be coupled with strong organisational alignment to achieve optimal and sustainable outcomes.

## **7. Conclusion**

This paper aimed to develop a framework for selecting, designing, and deploying AI agents to address the NIST CSF 2.0 cybersecurity requirements, based on identified agent categories, properties, and application contexts. The study introduced an AI agent taxonomy and decision framework to address this research aim. The AIATDF systematically aligns AI agent types, including reactive, cognitive, hybrid, and learning, with the NIST CSF 2.0 functions (Identify, Protect, Detect, Respond, Recover, Govern), categories, and subcategories.

### **7.1 Summary of key findings**

This paper's key findings are encapsulated within five distinct perspectives: (i) theoretical synthesis, (ii) a practical mapping instrument, (iii) a phased adoption strategy, (iv) versatile agent architectures, and (v) framework limitations. Firstly, the AIATDF achieves a novel theoretical synthesis by integrating agent-based AI theory with the structured cybersecurity outcomes defined by the NIST CSF 2.0. This integration, primarily through the differentiation of assisted, autonomous, and augmented intelligence modes, enhances the theoretical understanding of AI agent autonomy's application in cybersecurity. Secondly, the AIATDF functions as a practical mapping instrument, utilising a matrix to demonstrate the suitability of specific AI agent types for the NIST CSF 2.0. This provides SOC managers and cybersecurity architects with a structured, risk-mitigating guide for selecting AI agents. Thirdly, the AIATDF supports a phased adoption approach, enabling organisations to incrementally integrate AI solutions, from basic automated alerts to advanced autonomous and hybrid systems. Fourthly, as detailed in Section 5, conceptual validation demonstrates the AIATDF's comprehensive scope, encompassing all six NIST CSF 2.0 functions and highlighting the potential of a MAS architecture in addressing diverse cyber threats. Finally, while the AIATDF presents a promising theoretical foundation, it requires empirical validation and further development to incorporate socio-technical factors and the evolving capabilities of AI. These five perspectives highlight the potential of a theoretically grounded and practically applicable approach to deploying cybersecurity AI agents. Additionally, they offer academic insight and actionable guidance for enhancing AI-driven defences within the NIST CSF 2.0 framework.

### **7.2 Implications for future research**

This study opens several promising research avenues. Firstly, empirical validation and field studies are necessary to establish the applicability of the AIATDF. Pilot deployments in operational environments, measuring metrics such as response times, false positive/negative rates, and resource overhead across diverse organisational contexts (e.g., small vs. large enterprises), would provide essential data for model refinement. Secondly, exploring adaptive and evolving AI agent architectures is imperative, given the rapid evolution of cyber threats and AI technologies. For example, emerging open standards such as the MCP could significantly improve the integration of external data sources into AI agent architectures, rendering the AIATDF approach more scalable and easier to deploy in diverse organisational environments. Therefore, future research should investigate how MAS architectures with adaptive learning layers, leveraging techniques such as reinforcement, transfer, or federated learning, can maintain their efficacy in dynamic threat landscapes and distributed data scenarios. Thirdly, future research should integrate socio-technical factors into AI agent deployment methodologies. Explicitly factoring in organisational culture, budget constraints, and workforce skill sets would facilitate holistic readiness assessments and deepen the understanding of seamless, autonomous, or augmented intelligence implementation. Fourthly, developing ethical and regulatory frameworks is necessary to address the escalating ethical and compliance concerns associated with advanced AI agents. Studies could define guidelines for operationalising responsible AI practices within the NIST CSF 2.0, particularly in high-stakes scenarios. Fifthly, despite the emergence of AI agents CMMs from grey literature, there are no scholarly, well-defined CMMs for adopting AI agents. Future research could provide a structured path (CMM) outlining key stages of AI agent adoption progression and actionable steps for advancement. Ultimately, longitudinal studies examining the long-term deployment of AI agents would provide valuable insights into the evolution of agent-based solutions. These studies would inform continuous improvement mechanisms, ensuring the AIATDF's sustained relevance in the face of evolving AI capabilities and cybersecurity threats.

## **Statements and Declarations:**

### **Ethical Approval and Consent to Participate**

Not applicable.

## Consent for Publication

The author declares consent for publication.

## Competing Interests

The author has no relevant financial or non-financial interests to declare.

## Funding

No funding was received to conduct this study.

## Availability of Supporting Data

No datasets were generated or analysed during the current study. Only textual data from a systematic literature review underpin the study.

## Author's Contributions

Conceptualisation, methodology, writing—original draft preparation, and writing—final draft review and editing.

## Acknowledgements

The author would like to acknowledge the current 'read and publish' agreements negotiated by the South African National Library and Information Consortium (SANLiC) as having contributed to the open-access publication of this paper without the author facing any charges.

## Authors' Information

Masike Malatji

Graduate School of Business Leadership (SBL), University of South Africa (UNISA), Midrand, Johannesburg, South Africa

## Declaration of Generative AI and AI-assisted Technologies in the Writing Process

While preparing this work, the author used Scopus AI (Beta) to review some articles, ChatGPT 4.5 to analyse textual data, Google Gemini and ChatGPT 4.5 to structure the initial ideas and logical flow of the paper and improve its readability, and Grammarly for English language editing. After using these AI-powered tools, the author reviewed and edited the content as needed and took full responsibility for the publication's content.

## References

1. Gupta M, Akiri C, Aryal K, Parker E, Praharaj L (2023) From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy. *IEEE Access* 11: 80218–45. <https://doi.org/10.1109/ACCESS.2023.3300381>
2. Mahboubi A, Luong K, Aboutorab H, Bui HT, Jarrad G, Bahutair M, Camtepe S, Pogrebna G, Ahmed E, Barry B, Gately H (2024) Evolving techniques in cyber threat hunting: A systematic review. *Journal of Network and Computer Applications* 232: 104004. <https://doi.org/10.1016/j.jnca.2024.104004>
3. Villalón-Huerta A, Ripoll-Ripoll I, Marco-Gisbert H (2022) Key requirements for the detection and sharing of behavioral indicators of compromise. *Electronics* 11(3): 416. <https://doi.org/10.3390/electronics11030416>
4. Pinto A, Herrera L, Donoso Y, Gutierrez JA (2023) Survey on intrusion detection systems based on machine learning techniques for the protection of critical infrastructure. *Sensors* 23(5): 2415. <https://doi.org/10.3390/s23052415>
5. Priyalakshmi V, Devi R (2023) Analysis and implementation of normalisation techniques on KDD'99 data set for IDS and IPS. In: Saraswat M, Chowdhury C, Kumar Mandal C, Gandomi AH (eds) *Proceedings of International Conference on Data Science and Applications*. Springer Nature, Singapore, pp. 51–70. [https://doi.org/10.1007/978-981-19-6634-7\\_5](https://doi.org/10.1007/978-981-19-6634-7_5)
6. Namakshenas D, Yazdinejad A, Dehghantanha A, Srivastava G (2024) Federated quantum-based privacy-preserving threat detection model for consumer internet of things. *IEEE Transactions on Consumer Electronics* 70(3): 5829–38. <https://doi.org/10.1109/TCE.2024.3377550>
7. Kshetri N (2021) Economics of artificial intelligence in cybersecurity. *IT Professional* 23(5): 73–77. <https://doi.org/10.1109/MITP.2021.3100177>



8. Jalalvand F, Chhetri MB, Nepal S, Paris C (2024) Alert prioritisation in security operations centres: A systematic survey on criteria and methods. *ACM Computing Surveys* 57(2): 1–36. <https://doi.org/10.1145/3695462>
9. Sindiramutty SJ (2023) Autonomous threat hunting: a future paradigm for AI-driven threat intelligence. arXiv. <https://doi.org/10.48550/arXiv.2401.00286>
10. Chen W, Zhang J (2024) Elevating security operations: the role of ai-driven automation in enhancing soc efficiency and efficacy. *Journal of Artificial Intelligence and Machine Learning in Management* 8(2): 1–13.
11. Sarker IH, Janicke H, Mohammad N, Watters P, Nepal S (2024) AI potentiality and awareness: A position paper from the perspective of human-AI teaming in cybersecurity. In: Vasant P, Panchenko V, Munapo E, Weber G-W, Thomas JJ, Intan R, Arefin MS (eds) *Intelligent Computing and Optimization*. Springer Nature Switzerland, Cham, pp. 140–49. [https://doi.org/10.1007/978-3-031-50887-5\\_14](https://doi.org/10.1007/978-3-031-50887-5_14)
12. Shoaee H, Bagherinejad J, Rezaee Nour J (2022) Towards the analysis of information technology governance and productivity based on COBIT framework: An empirical study in e-banking. *Tehnicki Vjesnik - Technical Gazette* 29 (6). <https://doi.org/10.17559/TV-20220115074214>
13. Vats V, Nizam MB, Liu M, Wang Z, Ho R, Prasad MS, Titterton V, Malreddy SV, Aggarwal R, Xu Y, Ding L, Mehta J, Grinnell N, Liu L, Zhong S, Gandamani D, Tang X, Ghosalkar R, Shen C, Shen R, Hussain N, Ravichandran K, Davis J (2024) A survey on human-AI teaming with large pre-trained models. arXiv. <https://doi.org/10.48550/arXiv.2403.04931>
14. NIST (2024) The NIST Cybersecurity Framework (CSF) 2.0. <https://www.nist.gov/cyberframework>. Accessed 07 April 2025.
15. Lungu N, Rababah AA, Dash BB, Syed AH, Barik L, Rout S, Tembo S, Lubobya C, Patra SS (2024) NIST CSF-2.0 compliant GPU shader execution. *Engineering, Technology & Applied Science Research* 14(4): 15187–93. <https://doi.org/10.48084/etasr.7351>
16. Horan C, Saiedian H (2021) Cyber crime investigation: Landscape, challenges, and future research directions. *Journal of Cybersecurity and Privacy* 1 (4): 580–96. <https://doi.org/10.3390/jcp1040029>
17. Aslan Ö, Aktuğ SS, Ozkan-Okay M, Yilmaz AA, Akin E. A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions. *Electronics* 12(6): 1333. <https://doi.org/10.3390/electronics12061333>
18. Gonzalez C, Aggarwal P, Cranford EA, Lebiere C (2023) Adaptive cyberdefense with deception: A human–AI cognitive approach. In: Bao T, Tambe M, Wang C (eds) *Cyber deception: Techniques, strategies, and human aspects*, pp. 41–57. Springer International Publishing, Cham. [https://doi.org/10.1007/978-3-031-16613-6\\_3](https://doi.org/10.1007/978-3-031-16613-6_3)
19. Peng L, Li D, Zhang Z, Zhang T, Huang A, Yang S, Hu Y (2024) Human-AI collaboration: Unraveling the effects of user proficiency and AI agent capability in intelligent decision support systems. *International Journal of Industrial Ergonomics* 103: 103629. <https://doi.org/10.1016/j.ergon.2024.103629>
20. Putta P, Mills E, Garg N, Motwani S, Finn C, Garg D, Rafailov R (2024) Agent Q: Advanced reasoning and learning for autonomous AI agents. arXiv. <https://doi.org/10.48550/arXiv.2408.07199>
21. Horn J, Hallin N, Taheri H, O’Rourke M, Edwards D (2013) Intentional state-ascription in multi-agent systems. In: Müller VC (eds) *Philosophy and theory of artificial intelligence*, pp. 225–35. Springer, Berlin. [https://doi.org/10.1007/978-3-642-31674-6\\_17](https://doi.org/10.1007/978-3-642-31674-6_17)
22. Leite A, Girardi R, Novais P (2013) Using ontologies in hybrid software agent architectures. In: 2013 IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), 3: 155–58. <https://doi.org/10.1109/WI-IAT.2013.172>
23. Liu B, Mazumder S, Robertson E, Grigsby S (2023) AI autonomy: Self-initiated open-world continual learning and adaptation. *AI Magazine* 44(2): 185–99. <https://doi.org/10.1002/aaai.12087>
24. Mazumder S, Liu B (2024) Open-world continual learning: A framework. In: Mazumder S, Liu B (eds) *Lifelong and continual learning dialogue systems*. Springer International Publishing, Cham, pp. 21–47. [https://doi.org/10.1007/978-3-031-48189-5\\_2](https://doi.org/10.1007/978-3-031-48189-5_2)
25. Rossi N (2023) Applications of artificial intelligence in healthcare. *TKS Publisher* 41(2): 49–51. [https://www.teknoscienze.com/tns\\_article/applications-of-artificial-intelligence-in-healthcare/](https://www.teknoscienze.com/tns_article/applications-of-artificial-intelligence-in-healthcare/). Accessed 07 April 2025.
26. Rabuzin K, Maleković M, Bača M (2006) A survey of the properties of agents. *Journal of Information and Organizational Sciences* 30(1): 155–70. <https://hrcak.srce.hr/20873>
27. Garro A, Mühlhäuser M, Tundis A, Mariani S, Omicini A, Vizzari G (2018) Intelligent agents and environment. In: *Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics*, 1: 309–14. Elsevier. <https://doi.org/10.1016/B978-0-12-809633-8.20327-0>
28. Fernández JM, Pavón J (2010) Talking agents: A distributed architecture for interactive artistic installations. *Integrated Computer-Aided Engineering* 17(3): 243–59. <https://doi.org/10.3233/ICA-2010-0341>

29. Kefalas P, Stamatopoulou I (2011) Towards modelling of reactive, goal-oriented and hybrid intelligent agents using p systems. In: Gheorghe M, Hinze T, Păun G, Rozenberg G, Salomaa A (eds) *Membrane Computing*. Springer, Berlin, pp. 265–72. [https://doi.org/10.1007/978-3-642-18123-8\\_21](https://doi.org/10.1007/978-3-642-18123-8_21)
30. Bösser T, (2015) Autonomous agents. In: Wright JD (ed) *International encyclopedia of the social & behavioral sciences*, 2<sup>nd</sup> edition. Elsevier, Oxford, pp. 309–13. <https://doi.org/10.1016/B978-0-08-097086-8.43011-4>
31. Weng Y-H, Ho C-h (2020) Embodiment and algorithms for human–robot interaction. In: Barfield W (ed) *The Cambridge Handbook of the Law of Algorithms*. Cambridge University Press, Cambridge, pp. 736–56. <https://doi.org/10.1017/9781108680844.035>
32. Preethiya T, Subbiah P, Pandiarajan T, Ojo S, Vijayalakshmi S (2024) Artificial intelligence in robotics. In: *Modeling, simulation, and control of ai robotics and autonomous systems*. IGI Global Scientific Publishing, pp. 152–65. <https://doi.org/10.4018/979-8-3693-1962-8.ch009>
33. Dash S (2025) Green AI: Enhancing sustainability and energy efficiency in ai-integrated enterprise systems. *IEEE Access* 13: 21216–28. <https://doi.org/10.1109/ACCESS.2025.3532838>
34. Kott A (2018) Intelligent autonomous agents are key to cyber defense of the future army networks. *The Cyber Defense Review* 3(3): 57–70. <https://cyberdefensereview.army.mil/CDR-Content/Articles/Article-View/Article/1716477/intelligent-autonomous-agents-are-key-to-cyber-defense-of-the-future-army-netwo/>. Accessed 09 April 2025.
35. Théron P, Kott A (2019) When autonomous intelligent goodwill will fight autonomous intelligent malware: a possible future of cyber defense. In: 2019 IEEE Military Communications Conference (MILCOM), pp. 1–7. <https://doi.org/10.1109/MILCOM47813.2019.9021038>
36. Kott A, Théron P, Drašar M, Dushku E, LeBlanc B, Losiewicz P, Guarino A, Mancini L, Panico A, Pihelgas M, Rządca K, Gaspari FD (2023) Autonomous intelligent cyber-defense agent (AICA) reference architecture, release 2.0. arXiv. <https://doi.org/10.48550/arXiv.1803.10664>
37. Truong TC, Diep QB, Zelinka I (2020) Artificial intelligence in the cyber domain: offense and defense. *Symmetry* 12(3): 410. <https://doi.org/10.3390/sym12030410>
38. Ligo AK, Kott A, Linkov I (2021) How to measure cyber-resilience of a system with autonomous agents: approaches and challenges. *IEEE Engineering Management Review* 49(2): 89–97. <https://doi.org/10.1109/EMR.2021.3074288>
39. Naik B, Mehta A, Yagnik H, Shah M (2022) The impacts of artificial intelligence techniques in augmentation of cybersecurity: A comprehensive review. *Complex & Intelligent Systems* 8(2): 1763–80. <https://doi.org/10.1007/s40747-021-00494-8>
40. Sharma N, Jindal N (2024) Emerging artificial intelligence applications: Metaverse, IoT, cybersecurity, healthcare - An overview. *Multimedia Tools and Applications* 83(19): 57317–45. <https://doi.org/10.1007/s11042-023-17890-6>
41. Dodig-Crnkovic G, Burgin M (2024) A systematic approach to autonomous agents. *Philosophies* 9(2): 44. <https://doi.org/10.3390/philosophies9020044>
42. Srivastava A, Stager S (2024) Cognitive computing with deep learning based cybersecurity solution for human computer interface applications In: 2024 International Conference on Data Science and Network Security (ICDSNS), pp. 1–6. <https://doi.org/10.1109/ICDSNS62112.2024.10691061>
43. Xu W, Gao Z (2024) Applying HCAI in developing effective human-AI teaming: A perspective from human-AI joint cognitive systems. *Interactions* 31(1): 32–37. <https://doi.org/10.1145/3635116>
44. Abramoff MD (2021) Autonomous artificial intelligence safety and trust. In: Grzybowski A (ed) *Artificial Intelligence in Ophthalmology*. Springer International Publishing, Cham, pp. 55–67. [https://doi.org/10.1007/978-3-030-78601-4\\_4](https://doi.org/10.1007/978-3-030-78601-4_4)
45. Freitas MP, Piai VA, Farias RH, Fernandes AMR, de Moraes Rossetto AG, Leithardt VRQ (2022) Artificial intelligence of things applied to assistive technology: A systematic literature review. *Sensors* 22(21): 8531. <https://doi.org/10.3390/s22218531>
46. Yau K-LA, Lee HJ, Chong Y-W, Ling MH, Syed AR, Wu C, Goh HG (2021) Augmented intelligence: Surveys of literature and expert opinion to understand relations between human intelligence and artificial intelligence. *IEEE Access* 9: 136744–61. <https://doi.org/10.1109/ACCESS.2021.3115494>
47. Yau K-L, Saleem Y, Chong Y-W, Fan X, Eyu JM, Chieng D (2024) The augmented intelligence perspective on human-in-the-loop reinforcement learning: Review, concept designs, and future directions. *IEEE Transactions on Human-Machine Systems* 54(6): 762–77. <https://doi.org/10.1109/THMS.2024.3467370>
48. Simmler M, Frischknecht R (2021) A taxonomy of human–machine collaboration: Capturing automation and technical autonomy. *AI & Society* 36(1): 239–50. <https://doi.org/10.1007/s00146-020-01004-z>
49. Hinsén S, Hofmann P, Jöhnk J, Urbach N (2022) How can organizations design purposeful human-ai interactions: a practical perspective from existing use cases and interviews. In: *Hawaii International Conference on System Sciences*. [https://aisel.aisnet.org/hicss-55/cl/human-ai\\_collaboration/2](https://aisel.aisnet.org/hicss-55/cl/human-ai_collaboration/2)

50. Roch N, Sievers H, Schöni L, Zimmermann V (2024) Navigating autonomy: Unveiling security experts' perspectives on augmented intelligence in cybersecurity. Usenix Association, pp. 41–60. <https://www.usenix.org/conference/soups2024/presentation/roch>. Accessed 07 April 2025.
51. Dedeke A (2017) Cybersecurity framework adoption: using capability levels for implementation tiers and profiles. *IEEE Security & Privacy* 15(5): 47–54. <https://doi.org/10.1109/MSP.2017.3681063>
52. Dimitrov V, Kaloyanova K, Petrov M (2021) Adapted SANS cybersecurity policies for NIST cybersecurity framework. In: *CEUR Workshop Proceedings*, 2933: 293–301. <https://ceur-ws.org/Vol-2933/paper29.pdf>. Accessed 26 March 2025.
53. Edwards J, (2024) A comprehensive guide to the NIST cybersecurity framework 2.0: Strategies, implementation, and best practice. John Wiley & Sons, Ltd. <https://doi.org/10.1002/9781394280391.fmatter>.
54. Velasco JM (2017) The situation and evolution of the managed services of cybersecurity, towards 3.0 and beyond. In: Ramírez JM, García-Segura LA (eds) *Cyberspace: Risks and Benefits for Society, Security and Development*. Springer International Publishing, Cham, pp. 153–64. [https://doi.org/10.1007/978-3-319-54975-0\\_9](https://doi.org/10.1007/978-3-319-54975-0_9)
55. Axon L, Fletcher K, Scott AS, Stolz M, Hannigan R, Kaafarani AE, Goldsmith M, Creese S (2022) Emerging cybersecurity capability gaps in the industrial internet of things: overview and research agenda. *Digital Threats* 3(4): 34:1-34:27. <https://doi.org/10.1145/3503920>
56. NIST (2025) Cyber AI profile. <https://www.nccoe.nist.gov/projects/cyber-ai-profile>. Accessed 05 April 2025.
57. Malatji M, Tolah A (2024) Artificial intelligence (AI) cybersecurity dimensions: A comprehensive framework for understanding adversarial and offensive AI. *AI and Ethics*. <https://doi.org/10.1007/s43681-024-00427-4>.
58. Dash B, Ansari MF, Sharma P, Ali A (2022) Threats and opportunities with ai-based cyber security intrusion detection: A review. *International Journal of Software Engineering & Applications* 13(5). <https://ssrn.com/abstract=4323258>
59. Belaïd A, (2024) Human-machine collaboration for incident response in cybersecurity operations for autonomous vehicles. *African Journal of Artificial Intelligence and Sustainable Development* 4(1): 297–321. <https://africansciencegroup.com/index.php/AJAISD/article/view/98>
60. Johnson M, Bradshaw JM, Feltovich PJ (2018) Tomorrow's human-machine design tools: from levels of automation to interdependencies. *Journal of Cognitive Engineering and Decision Making* 12(1): 77–82. <https://doi.org/10.1177/1555343417736462>
61. Ivanov SH (2023) Automated decision-making. *Foresight* 25(1): 4–19. <https://doi.org/10.1108/FS-09-2021-0183>
62. Altamimi S, Altamimi B, Côté D, Shirmohammadi S (2023) Toward a superintelligent action recommender for network operation centers using reinforcement learning. *IEEE Access* 11: 20216–29. <https://doi.org/10.1109/ACCESS.2023.3248652>
63. Tilbury J, Flowerday S (2024) Humans and automation: augmenting security operation centers. *Journal of Cybersecurity and Privacy* 4(3): 388–409. <https://doi.org/10.3390/jcp4030020>
64. Mageshkumar N, Vijayaraj A, Arunpriya N, Sangeetha A (2022) Efficient spam filtering through intelligent text modification detection using machine learning. *Materials Today: Proceedings, International Conference on Advanced Materials for Innovation and Sustainability*, 64: 848–58. <https://doi.org/10.1016/j.matpr.2022.05.364>
65. Neff G, Nagy P (2016) Automation, algorithms, and politics| talking to bots: Symbiotic agency and the case of Tay. *International Journal of Communication* 10:17. <https://ijoc.org/index.php/ijoc/article/view/6277>.
66. Girdhar M, Hong J, Moore J (2023) Cybersecurity of autonomous vehicles: a systematic literature review of adversarial attacks and defense models. *IEEE Open Journal of Vehicular Technology* 4: 417–37. <https://doi.org/10.1109/OJVT.2023.3265363>
67. Mudhivarthi BR, Thakur P, Singh G (2023) Aspects of cyber security in autonomous and connected vehicles. *Applied Sciences* 13(5): 3014. <https://doi.org/10.3390/app13053014>
68. Dehghantanha A, Yazdinejad A, Parizi RM (2024) Autonomous cybersecurity: evolving challenges, emerging opportunities, and future research trajectories. In: *Proceedings of the Workshop on Autonomous Cybersecurity*. Association for Computing Machinery, pp. 1–10, New York, NY, USA. <https://doi.org/10.1145/3689933.3690832>
69. Morovat K, Panda B (2020) A survey of artificial intelligence in cybersecurity. In: *2020 International Conference on Computational Science and Computational Intelligence (CSCI)*, pp. 109–15. <https://doi.org/10.1109/CSCI51800.2020.00026>
70. Rai HM, Galymzada A, Almas K, Nurzhan D, Alibek M (2024) Fortifying cyber defenses: a deep dive into the development of an AI-powered network intrusion detection system. In: Tanwar S, Singh PK, Ganzha M, Epiphaniou G (eds) *Proceedings of 5<sup>th</sup> International Conference on Computing, Communications, and Cyber-Security*. Springer Nature, Singapore, pp. 809–21. [https://doi.org/10.1007/978-981-97-2550-2\\_58](https://doi.org/10.1007/978-981-97-2550-2_58)
71. Chan A, Ezell C, Kaufmann M, Wie K, Hammond L, Bradley H, Bluemke E, Rajkumar N, Krueger D, Kolt N, Heim L, Anderljung M (2024) Visibility into AI agents'. In: *Proceedings of the 2024 ACM Conference on*

- Fairness, Accountability, and Transparency. Association for Computing Machinery, New York, NY, USA, pp. 958–73. <https://doi.org/10.1145/3630106.3658948>
72. Cronin I (2024) Autonomous AI agents: Decision-making, data, and algorithms. In: Cronin I (ed) *Understanding generative ai business applications: A guide to technical principles and real-world applications*, pp. 165–80. Apress, Berkeley, CA, USA. [https://doi.org/10.1007/979-8-8688-0282-9\\_11](https://doi.org/10.1007/979-8-8688-0282-9_11)
  73. Jiang Y-H, Li R., Zhou Y, Qi C, Hu H, Wei Y, Jiang B, Wu Y (2024) AI agent for education: Von Neumann multi-agent system framework. *arXiv*. <https://doi.org/10.48550/arXiv.2501.00083>
  74. Capuano N, Fenza G, Loia V, Stanzione C (2022) Explainable artificial intelligence in cybersecurity: A survey. *IEEE Access* 10: 93575–600. <https://doi.org/10.1109/ACCESS.2022.3204171>
  75. Desai B, Patil K, Mehta I, Patil A (2024) 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. 135–50. <https://doi.org/10.1109/iSemantic63362.2024.10762690>
  76. Radanliev P (2025) AI ethics: Integrating transparency, fairness, and privacy in AI development. *Applied Artificial Intelligence* 39(1): 2463722. <https://doi.org/10.1080/08839514.2025.2463722>
  77. Prasad RR, Robinson RRR, Thomas C, Balakrishnan N (2021) Evaluation of strategic decision taken by autonomous agent using explainable AI. In: *4<sup>th</sup> International Conference on Security and Privacy (ISEA-ISAP)*, pp. 1–8. <https://doi.org/10.1109/ISEA-ISAP54304.2021.9689715>
  78. Hamet P, Tremblay J (2017) Artificial intelligence in medicine. *Metabolism – Clinical and Experimental* 69:S36–40. <https://doi.org/10.1016/j.metabol.2017.01.011>
  79. Berberian B, Somon B, Sahaï A, Gouraud J (2017) The out-of-the-loop brain: A neuroeconomic approach of the human automation interaction. *Annual Reviews in Control* 44: 303–15. <https://doi.org/10.1016/j.arcontrol.2017.09.010>
  80. Hellebrandt T, Huebser L, Adam T, Heine I, Schmitt RH (2021) Augmented intelligence – mensch trifft künstliche intelligenz: intelligentes zusammenwirken von mensch und ki für bessere entscheidungen und handlungen in der produktion. *Zeitschrift Für Wirtschaftlichen Fabrikbetrieb* 116(6): 433–37. <https://doi.org/10.1515/zwf-2021-0104>
  81. Jain H, Padmanabhan B, Pavlou PA, Raghu TS (2021) Editorial for the special section on humans, algorithms, and augmented intelligence: The future of work, organizations, and society. *Information Systems Research* 32(3): 675–87. <https://doi.org/10.1287/isre.2021.1046>
  82. Karunamurthy A, Kiruthivasan R, Gauthamkrishna S (2023) Human-in-the-Loop intelligence: Advancing AI-centric cybersecurity for the future. *Quing: International Journal of Multidisciplinary Scientific Research and Development* 2(3): 20–43. <https://doi.org/10.54368/qijmsrd.2.3.0011>
  83. Ahdadou M, Ajaly A, Tahrouch M (2024) Unlocking the potential of augmented intelligence: A discussion on its role in boardroom decision-making. *International Journal of Disclosure and Governance* 21(3): 433–46. <https://doi.org/10.1057/s41310-023-00207-2>
  84. Gore S, Hamsa S, Roychowdhury S, Patil G, Gore S, Karmode S (2023) Augmented intelligence in machine learning for cybersecurity: Enhancing threat detection and human-machine collaboration. In: *2023 Second International Conference on Augmented Intelligence and Sustainable Systems (ICAISS)*, pp. 638–44. <https://doi.org/10.1109/ICAISS58487.2023.10250514>
  85. Caballero-Martin D, Lopez-Guede JM, Estevez J, Graña M (2024) Artificial intelligence applied to drone control: a state of the art. *Drones* 8(7): 296. <https://doi.org/10.3390/drones8070296>
  86. Chen L, Zhang W, Song Y, Chen J (2024) Machine learning for human-machine systems with advanced persistent threats. *IEEE Transactions on Human-Machine Systems* 54(6): 753–61. <https://doi.org/10.1109/THMS.2024.3439625>
  87. Salesforce (2025) The agentic maturity model: A 4-step roadmap for CIOs to succeed in the agentic era. <https://www.salesforce.com/news/stories/agentic-maturity-model/>. Accessed 15 April 2025
  88. Paul S, Choudhury NR, Pandit B, Dawn A (2025) Integration of AI and quantum computing in cybersecurity: A comprehensive review. In: *Integration of AI, Quantum Computing, and Semiconductor Technology*. IGI Global Scientific Publishing, pp. 287–308. <https://doi.org/10.4018/979-8-3693-7076-6.ch014>
  89. Kuru K, Kuru K (2025) UMetaBE-DPPML: Urban metaverse & blockchain-enabled decentralised privacy-preserving machine learning verification and authentication with metaverse immersive devices. *Internet of Things and Cyber-Physical Systems*. <https://doi.org/10.1016/j.iotcps.2025.02.001>
  90. Qi P, Chiaro D, Guzzo A, Ianni M, Fortino G, Piccialli F (2024) Model aggregation techniques in federated learning: A comprehensive survey. *Future Generation Computer Systems* 150: 272–93. <https://doi.org/10.1016/j.future.2023.09.008>
  91. Yu E, Yue W, Jianzheng S, Xun W (2024) Blockchain-based AI agent and autonomous world infrastructure. In: *2024 IEEE Conference on Artificial Intelligence (CAI)*, pp. 278–83. <https://doi.org/10.1109/CAI59869.2024.00061>



92. Han X, Wang N, Che S, Yang H, Zhang K, Xu SX (2024) Enhancing investment analysis: Optimizing AI-agent collaboration in financial research. In: Proceedings of the 5th ACM International Conference on AI in Finance. Association for Computing Machinery, New York, NY, USA, pp. 538–46. <https://doi.org/10.1145/3677052.3698645>
93. Chen Z, Sun Q, Li N, Li X, Wang Y, I C-L (2024) Enabling mobile AI agent in 6G era: Architecture and key technologies. *IEEE Network* 38(5): 66–75. <https://doi.org/10.1109/MNET.2024.3422309>
94. Bovo R, Abreu S, Ahuja K, Gonzalez EJ, Cheng L-T, Gonzalez-Franco M (2024) EmBARDiment: an embodied AI agent for productivity in XR. *arXiv*. <https://doi.org/10.48550/arXiv.2408.08158>
95. Baabdullah AM (2024) Generative conversational AI agent for managerial practices: the role of qi dimensions, novelty seeking and ethical concerns. *Technological Forecasting and Social Change* 198: 122951. <https://doi.org/10.1016/j.techfore.2023.122951>
96. Huang Q, Wake N, Sarkar B, Durante Z, Gong R, Taori R, Noda Y, Terzopoulos D, Kuno N, Famoti A, Llorens A, Langford J, Vo H, Fei-Fei L, Ikeuchi K, Gao J (2024) Position paper: Agent AI towards a holistic intelligence. *arXiv*. <https://doi.org/10.48550/arXiv.2403.00833>
97. Agashe S, Han J, Gan S, Yang J, Li A, Wang XE (2024) Agent S: An open agentic framework that uses computers like a human. *arXiv.Org*. <https://arxiv.org/abs/2410.08164v1>
98. Kim M, Saad W (2024) Analysis of the memorization and generalization capabilities of AI agents: Are continual learners robust? In: 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6840–44. <https://doi.org/10.1109/ICASSP48485.2024.10447575>
99. Pleshakova E, Osipov A, Gataullin S, Gataullin T, Vasilakos A (2024) Next gen cybersecurity paradigm towards artificial general intelligence: Russian market challenges and future global technological trends. *Journal of Computer Virology and Hacking Techniques* 20(3): 429–40. <https://doi.org/10.1007/s11416-024-00529-x>
100. Hauptman AI, Schelble BG, McNeese NJ, Madathil KC (2023) Adapt and overcome: perceptions of adaptive autonomous agents for human-AI teaming. *Computers in Human Behavior* 138: 107451. <https://doi.org/10.1016/j.chb.2022.107451>
101. Kaur R, Gabrijelčić D, Klobučar T (2023) Artificial intelligence for cybersecurity: Literature review and future research directions. *Information Fusion* 97: 101804. <https://doi.org/10.1016/j.inffus.2023.101804>
102. Cañas JJ (2022) AI and ethics when human beings collaborate with AI agents. *Frontiers in Psychology* 13. <https://doi.org/10.3389/fpsyg.2022.836650>
103. Roy SD, Debbarma S, Guerrero JM (2022) Machine learning based multi-agent system for detecting and neutralizing unseen cyber-attacks in AGC and HVDC systems. *IEEE Journal on Emerging and Selected Topics in Circuits and Systems* 12(1): 182–93. <https://doi.org/10.1109/JETCAS.2022.3142055>
104. Li L, El Rami J-PS, Taylor A, Rao JH, Kunz T (2022) Enabling a network AI gym for autonomous cyber agents. In: 2022 International Conference on Computational Science and Computational Intelligence (CSCI), pp. 172–77. <https://doi.org/10.1109/CSCI58124.2022.00034>
105. Ashktorab Z, Dugan C, Johnson J, Pan Q, Zhang W, Kumaravel S, Campbell M (2021) Effects of communication directionality and ai agent differences in human-ai interaction. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems. Association for Computing Machinery, New York, pp. 1–15. <https://doi.org/10.1145/3411764.3445256>
106. Zolotukhin M, Kumar S, Hämäläinen T (2020) Reinforcement learning for attack mitigation in SDN-enabled networks. In: 2020 6<sup>th</sup> IEEE Conference on Network Softwarization (NetSoft), pp. 282–86. <https://doi.org/10.1109/NetSoft48620.2020.9165383>
107. Cao Y, Wang R, Chen M, Barnawi A (2020) AI agent in software-defined network: agent-based network service prediction and wireless resource scheduling optimization. *IEEE Internet of Things Journal* 7(7): 5816–26. <https://doi.org/10.1109/JIOT.2019.2950730>
108. Franco MF, Rodrigues B, Scheid EJ, Jacobs A, Killer C, Granville LZ, Stiller B (2020) SecBot: A business-driven conversational agent for cybersecurity planning and management. In: 2020 16th International Conference on Network and Service Management (CNSM), pp. 1–7. <https://doi.org/10.23919/CNSM50824.2020.9269037>
109. Kott A, Théron P (2020) Doers, not watchers: Intelligent autonomous agents are a path to cyber resilience. *IEEE Security & Privacy* 18(3): 62–66. <https://doi.org/10.1109/MSEC.2020.2983714>
110. Théron P, Kott A, Drašar M, Rządca K, LeBlanc B, Pihelgas M, Mancini L, Panico A (2018) Towards an active, autonomous and intelligent cyber defense of military systems: The NATO AICA reference architecture. In: 2018 International Conference on Military Communications and Information Systems (ICMCIS), pp. 1–9. <https://doi.org/10.1109/ICMCIS.2018.8398730>
111. Grzonka D, Jakóbić A, Kołodziej J, Pllana S (2018) Using a multi-agent system and artificial intelligence for monitoring and improving the cloud performance and security. *Future Generation Computer Systems* 86: 1106–17. <https://doi.org/10.1016/j.future.2017.05.046>

112. Yampolskiy RV (2018) Predicting future AI failures from historic examples. *Foresight* 21(1): 138–52. <https://doi.org/10.1108/FS-04-2018-0034>
113. Petrović VM (2018) Artificial intelligence and virtual worlds – toward human-level AI agents. *IEEE Access* 6: 39976–88. <https://doi.org/10.1109/ACCESS.2018.2855970>
114. Huang X, Lian J, Lei Y, Yao J, Lian D, Xie X (2024) Recommender AI agent: Integrating large language models for interactive recommendations. *arXiv*. <https://doi.org/10.48550/arXiv.2308.16505>
115. Hochmair HH, Juhász L, Kemp T (2024) Correctness comparison of ChatGPT-4, Gemini, Claude-3, and Copilot for spatial tasks. *Transactions in GIS* 28(7): 2219–31. <https://doi.org/10.1111/tgis.13233>
116. Mao Y, Ge Y, Fan Y, Xu W, Mi Y, Hu Z, Gao Y (2025) A survey on LoRA of large language models. *Frontiers of Computer Science* 19 (7): 197605. <https://doi.org/10.1007/s11704-024-40663-9>
117. Ding N, Qin Y, Yang G, Wei F, Yang Z, Su Y, Hu S, Chen Y, Chan C-M, Chen W, Yi J, Zhao W, Wang X, Liu Z, Zheng H-T, Chen J, Liu Y, Tang J, Li J, Sun M (2023) Parameter-efficient fine-tuning of large-scale pre-trained language models. *Nature Machine Intelligence* 5(3): 220–35. <https://doi.org/10.1038/s42256-023-00626-4>
118. Trad F, Chehab A (2024) Prompt engineering or fine-tuning? a case study on phishing detection with large language models. *Machine Learning and Knowledge Extraction* 6(1): 367–84. <https://doi.org/10.3390/make6010018>
119. Anisuzzaman DM, Malins JG, Friedman PA, Attia ZI (2025) Fine-tuning large language models for specialized use cases. *Mayo Clinic Proceedings: Digital Health* 3(1): 100184. <https://doi.org/10.1016/j.mcpdig.2024.11.005>
120. Anthropic (2024) Introducing the model context protocol. <https://www.anthropic.com/news/model-context-protocol>. Accessed 07 April 2025.
121. Hou X, Zhao Y, Wang S, Wang H (2025) Model context protocol (MCP): Landscape, security threats, and future research directions. *arXiv*. <https://doi.org/10.48550/arXiv.2503.23278>
122. Surapaneni R, Jha M, Vakoc M, Segal T (2025) Announcing the Agent2Agent protocol (A2A). Google developers blog. <https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/>. Accessed 11 April 2025.
123. Deng Z, Guo Y, Han C, Ma W, Xiong J, Wen S, Xiang Y (2025) AI agents under threat: A survey of key security challenges and future pathways. *ACM Comput. Surv.* 57(7): 182:1-182:36. <https://doi.org/10.1145/3716628>
124. Alrfai MM, Alqudah H, Lutfi A, Al-Kofahi M, Alrawad M, Almaiah MA (2023) The influence of artificial intelligence on the aiss efficiency: moderating effect of the cyber security. *Cogent Social Sciences* 9(2): 2243719. <https://doi.org/10.1080/23311886.2023.2243719>
125. Phillips-Wren G (2008) Intelligent decision support to assist real-time collaboration. In: 2008 International Symposium on Collaborative Technologies and Systems, pp. 375–375. <https://doi.org/10.1109/CTS.2008.4543953>
126. Gu TL, Li L (2021) Artificial moral agents and their design methodology: retrospect and prospect. *Chinese Journal of Computers* 44(3): 632–51. <https://www.doi.org/http://dx.doi.org/10.11897/SP.J.1016.2021.00632>
127. Abiodun IA, Khuen CW (2014) A multi agent framework (MAFSNUD) for Belief-Desire-Intention (BDI) model's decision-making problem in dynamic situations: An overview. *Advanced Science Letters* 20(1): 91–96. <https://doi.org/10.1166/asl.2014.5305>
128. Sethy A, Shaik N, Yadavalli PK, Anandaraj SP (2023) 9 AI: Issues, concerns, and ethical considerations. In: de Albuquerque VH, Raj P, Yadav SP (eds) *Toward Artificial General Intelligence: Deep Learning, Neural Networks, Generative AI*. De Gruyter Brill, Berlin, Germany, pp. 189–212. <https://doi.org/10.1515/978311323749-009>
129. Jameel T, Ali R, Toheed I (2020) Ethics of artificial intelligence: Research challenges and potential solutions. In: 2020 3<sup>rd</sup> International Conference on Computing, Mathematics and Engineering Technologies (iCoMET), pp. 1–6. <https://doi.org/10.1109/iCoMET48670.2020.9073911>
130. Pantoja CE, de Jesus VS, Lazarin NM, Viterbo J (2023) A spin-off version of ason for IoT and embedded multi-agent systems. In: Naldi MC, Bianchi RAC (eds) *Intelligent Systems*. Springer Nature, Cham, pp. 382–96. [https://doi.org/10.1007/978-3-031-45368-7\\_25](https://doi.org/10.1007/978-3-031-45368-7_25)
131. Leng J, Fyfe C, Jain L (2008) Simulation and reinforcement learning with soccer agents. *Multiagent and Grid Systems* 4(4): 415–36. <https://doi.org/10.3233/MGS-2008-4407>
132. Sennott SC, Akagi L, Lee M, Rhodes A (2019) AAC and artificial intelligence (AI). *Topics in Language Disorders* 39(4): 389. <https://doi.org/10.1097/TLD.000000000000197>
133. Rossi F, Mattei N (2019) Building ethically bounded AI. In: *Proceedings of the AAAI Conference on Artificial Intelligence* 33(01): 9785–89. <https://doi.org/10.1609/aaai.v33i01.33019785>
134. Nickles M, Rovatsos M, Weiss G (2004) Empirical-rational semantics of agent communication. In: *Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems*,

2004. AAMAS 2004. 2: 94–101. <http://www.doi.org/https://doi.ieeecomputersociety.org/10.1109/AAMAS.2004.10055>
135. Pitardi V, Marriott HR (2022) Challenging vulnerability perceptions towards voice activated assistants: An abstract. In: Allen J, Jochims B, Wu S (eds) Celebrating the past and future of marketing and discovery with social impact. Springer International Publishing, Cham, pp. 255–56. [https://doi.org/10.1007/978-3-030-95346-1\\_88](https://doi.org/10.1007/978-3-030-95346-1_88)
136. Chakraborty S (2023) AI and ethics: navigating the moral landscape. In: Investigating the impact of AI on ethics and spirituality. IGI Global Scientific Publishing, pp. 25–33. <https://doi.org/10.4018/978-1-6684-9196-6.ch002>
137. He T, Jazizadeh F (2024) Trust in human-AI interaction: review of empirical research on trust in AI-powered smart home ecosystems. In: Computing in Civil Engineering 2023. ASCE, Reston, Virginia: USA, pp. 530–38. <https://doi.org/10.1061/9780784485224.064>
138. Rashid AB, and Kausik AK (2024) AI Revolutionizing industries worldwide: A comprehensive overview of its diverse applications. Hybrid Advances 7: 100277. <https://doi.org/10.1016/j.hybadv.2024.100277>
139. Liu Y, Cao X, Chen T, Jiang Y, You J, Wu M, Wang X, Feng M, Jin Y, Chen J (2025) From screens to scenes: a survey of embodied ai in healthcare. Information Fusion 119: 103033. <https://doi.org/10.1016/j.inffus.2025.103033>
140. Majumdar S, (2024) Standards for LLM security. In: Kucharavy A, Plancherel O, Mulder V, Mermoud A, Lenders V (eds) Large language models in cybersecurity: Threats, exposure and mitigation. Springer Nature, Cham, pp. 225–31. [https://doi.org/10.1007/978-3-031-54827-7\\_25](https://doi.org/10.1007/978-3-031-54827-7_25)
141. Simran, Kumar S, Hans A (2024) The AI shield and red AI framework: Machine learning solutions for cyber threat intelligence(CTI). In: 2024 International Conference on Intelligent Systems for Cybersecurity (ISCS), pp. 1–6. <https://doi.org/10.1109/ISCS61804.2024.10581195>
142. Cam H (2020) Cyber resilience using autonomous agents and reinforcement learning. In: Artificial intelligence and machine learning for multi-domain operations applications II, 11413: 219–34. SPIE. <https://doi.org/10.1117/12.2559319>
143. Falowo OI, Botsyoe LE, Koshedo K, Ozer M (2024) Enhancing cybersecurity with artificial immune systems and general intelligence: A new frontier in threat detection and response. IEEE Access 12: 123811–22. <https://doi.org/10.1109/ACCESS.2024.3454543>
144. Rafferty L, Macdermott A (2024) Adaptive defence of the Internet of Things (IoT) using the Belief-Desire-Intention (BDI) model for social robots. In: Proceedings of the 57<sup>th</sup> Hawaii International Conference on System Sciences, pp. 1722–32. Waikiki, Hawaii. <https://scholarspace.manoa.hawaii.edu/items/86168302-9592-45e4-9794-2dbbb78e614c>
145. Baird A, Maruping LM (2021) The next generation of research on IS use: A theoretical framework of delegation to and from agentic is artefacts. Management Information Systems Quarterly 45(1b): 315–41. <https://www.doi.org/10.25300/MISQ/2021/15882>
146. Candrian C, Scherer A (2022) Rise of the machines: delegating decisions to autonomous AI. Computers in Human Behavior 134107308. <https://doi.org/10.1016/j.chb.2022.107308>
147. Dorri A, Kanhere SS, Jurdak R (2018) Multi-agent systems: A survey. IEEE Access 6: 28573–93. <https://doi.org/10.1109/ACCESS.2018.2831228>
148. Radanliev P, Santos O, Brandon-Jones A, Joinson A (2024) Ethics and responsible AI deployment. Frontiers in Artificial Intelligence 7: 1–17. <https://doi.org/10.3389/frai.2024.1377011>
149. Alhalalmeh A, Al-Tarawneh A (2025) Artificial intelligence and the law: The complexities of technology and legalities. In: Hannon A, Mahmood A (eds) Intelligence-driven circular economy: Regeneration towards sustainability and social responsibility, volume 2. Springer Nature Switzerland, Cham, pp. 641–49. [https://doi.org/10.1007/978-3-031-74220-0\\_50](https://doi.org/10.1007/978-3-031-74220-0_50)
150. Bhateja N, Sethi N, Kumar D (2018) Study of ant colony optimization technique for coalition formation in multi agent systems. In: 2018 International Conference on Circuits and Systems in Digital Enterprise Technology, pp. 1–4. <https://doi.org/10.1109/ICCSDET.2018.8821175>
151. Russell SJ, Norvig P (2020) Artificial intelligence: A modern approach. 4<sup>th</sup> ed. Pearson, Hoboken, NJ, USA.
152. Ye P, Wang T, Wang F-Y (2018) A survey of cognitive architectures in the past 20 years. IEEE Transactions on Cybernetics 48(12): 3280–90. <https://doi.org/10.1109/TCYB.2018.2857704>
153. Andreadis G, Klazoglou P, Niotaki K, Bouzakis K-D (2014) Classification and review of multi-agents systems in the manufacturing section. Procedia Engineering, 69: 282–90. <https://doi.org/10.1016/j.proeng.2014.02.233>
154. Lim M (2024) A typology of validity: content, face, convergent, discriminant, nomological and predictive validity. Journal of Trade Science 12(3): 155–79. <https://doi.org/10.1108/JTS-03-2024-0016>