Human-Centric Intelligent Systems (2025) 5:151–180
https://doi.org/10.1007/s44230-025-00103-8

RESEARCH ARTICLE

Augmented Intelligence Framework for Human–Artificial Intelligence Teaming in Cybersecurity

Masike Malatji[1]

Received: 29 October 2024 / Revised: 16 May 2025 / Accepted: 26 May 2025 / Published online: 13 June 2025

Abstract

As cyberattacks grow more complex and frequent, organisations and nations face critical challenges in safeguarding their information systems and sensitive data. Recognising the limitations of traditional, solely human-centric defences, there is increasing agreement among practitioners and researchers on the need for a collaborative approach that integrates human intelligence with artificial intelligence (AI). This paper introduces the cybersecurity Augmented Intelligence Framework (*c*AIF), a conceptual framework designed to optimise human-AI teaming (HAIT) in cybersecurity. Augmented intelligence is about the role of AI enhancing rather than replacing human intelligence through a more harmonious working relationship. The methods followed consist of three phases. First, reviewing existing literature to identify foundational human–machine interaction (HMI) paradigms. A systematic review of papers from three databases led to a final selection of 20 analysis units. Second, the strengths and weaknesses of the identified paradigms for HAIT were evaluated. Lastly, outlining the core architectural components of the *c*AIF from the strengths of each paradigm. Six key HMI paradigms were identified: Human-in-the-loop (HITL), Human-out-of-the-loop (HOOTL), Human-on-the-loop (HOTL), Human-alongside-the-loop, Human-in-command, and Coactive Systems. Each paradigm offers unique strengths: for instance, HITL emphasises active and direct human intervention, while HOOTL supports full autonomy of AI operations. On the other hand, HOTL balances AI autonomy with human oversight. The analysed data suggests that strategically leveraging the strength of each paradigm allows for a hybrid intelligent framework comprising five core components: the Decision-Making Matrix, Paradigm Allocation Engine, Task-Specific Modules, Feedback and Learning System, and Interoperability Framework. The *c*AIF shows promise in enhancing human-AI collaboration, integrating human insights with AI capabilities to improve resilience and adaptability against evolving cyber threats. Future research should focus on empirically validating the *c*AIF in various cybersecurity domains, including healthcare and finance.

Keywords Artificial intelligence · Augmented intelligence · Cybersecurity · Human–Machine interaction · Human–Machine systems · Human-AI Teaming

1 Introduction

The rapidly increasing threat landscape defines today's digital world, pushing the need for robust and flexible cybersecurity strategies to the forefront [1]. As cyberattacks become more sophisticated and frequent, both organisations and countries are finding themselves in a constant battle to protect their critical information systems and data [1, 2]. It is becoming clear that relying solely on human efforts to defend against these threats is not sufficient. This realisation has led to a growing call for a more collaborative approach that combines humans' unique strengths and artificial intelligence (AI) [3, 4]. Such a strategy aims to harness the best of both worlds, achieving more significant results through teamwork and shared learning [5]. This concept, known as human-AI teaming (HAIT) [6], is at the heart of what this paper explores, especially in modern cybersecurity operations. This paper's first of two objectives is, therefore, to identify and explore selected human–machine interaction (HMI) [7] paradigms, then determine their strengths, limitations, and suitability concerning HAIT for modern

✉ Masike Malatji
malatm1@unisa.ac.za

[1] Graduate School of Business Leadership (SBL), University of South Africa (UNISA), Midrand, PO Box 392, Johannesburg 0003, South Africa

cybersecurity operations. These HMI paradigms could be thought of as types of human–machine relationships [8].

When protecting information systems, networks, and data from sophisticated cyber threats, a comprehensive approach to cybersecurity is essential for maintaining the resilience of organisations and nations [9]. This paper delves into the current landscape of cybersecurity operations, highlighting the practices in use and the challenges faced. Among the strategies discussed are *multilayered defence tactics*, which include both the concept of defence-in-depth—layering multiple security measures like firewalls, intrusion detection systems, encryption, and access controls to fend off different threats—and the zero-trust architecture [10, 11]. The latter emphasises the importance of not automatically trusting any user or system, even those within the network perimeter, requiring continuous verification of identities and strict access controls to ensure security [11]. Another critical aspect of cybersecurity operations is *threat intelligence and monitoring*. This involves constant vigilance, where a security operations centre (SOC) monitors network traffic, systems, and endpoints around the clock to spot and respond to anomalies and potential threats as they arise [12, 13]. Integrating threat intelligence platforms allows for aggregating and analysing threat data from various sources, providing actionable insights that enhance situational awareness.

Incident response and management are also crucial, requiring the development and regular updating of incident response plans. These plans are vital for guiding organisations in responding effectively to and recovering from cybersecurity incidents [14]. This includes conducting forensic analyses after an incident to understand the attack vector, scope, and impact, which helps strengthen future defences. Having metrics such as response time to understand and improve SOC operations is thus a good practice [15]. *Security automation and orchestration* are powerful methods in the realm of cybersecurity. This approach uses sophisticated tools such as security orchestration, automation, and response (SOAR) to automate the repetitive tasks that bog down security teams and streamline the processes they follow, making detecting and responding to threats faster and more efficient [16]. Think of it like setting up a highly intelligent system that can learn from past security incidents. By drawing on historical data, AI and machine learning (ML) are harnessed to spot patterns, identify oddities, and even foresee potential threats before they happen [4]. Then, there is the critical area of *compliance and risk management*. This involves regularly evaluating the cybersecurity landscape to spot any weak spots that might be exploited and figuring out which areas need the most immediate attention, all while keeping in line with the legal standards set by various regulations [17]. Whether it is abiding by the Protection of Personal Information (POPI) Act in South Africa,[1] staying compliant with the General Data Protection Regulation (GDPR) in the European Union (EU),[2] adhering to the Health Insurance Portability and Accountability Act (HIPAA) in the United States of America (USA) or following the Payment Card Industry Data Security Standard (PCI-DSS) [18], the goal is the same: to protect sensitive information and avoid the headaches of legal trouble. Lastly, there is the human element—*user awareness and training*. This is not just about having policies; it is about actively educating your team on the best practices in cybersecurity and keeping them up to date on the latest threats [19]. This could be as simple as running phishing simulations to test and improve everyone's ability to spot and stop attacks in their tracks. After all, an informed team is a secure team.

In our world, where digital connections crisscross the globe, we face familiar yet ever-evolving cybersecurity challenges. These include the *evolving threat landscape* with advanced cyber threats that seek to steal sensitive data or disrupt operations, often driven by highly sophisticated groups, including advanced persistent threat actors and state-sponsored groups [20]. There is also the ongoing *resource constraint* challenge, such as the need for more skilled professionals to keep up with these threats, making security teams feel like they are always playing catch-up [21]. On top of this, the *complexity of our digital environments* keeps growing. With more devices connecting to the internet, more people working remotely, and more data moving to the cloud, it is getting harder to keep everything secure, especially when old systems not designed for the digital environment are still in use [22, 23]. Then, there is the maze of *data privacy laws and regulations*. Different places have different rules, making protecting personal information challenging while keeping the business running smoothly and respecting user privacy1,2 [24]. *Detecting and responding to cyber incidents* quickly is another hurdle, with many organisations finding their tools and processes need to be up to the task [1, 14]. This can lead to too many false alarms, making it easy to miss real threats. *Supply chain security* is another critical issue, as attackers find ways to exploit vulnerabilities in the network of third-party vendors and partners an organisation relies on [25, 26]. Table 1 summarises the cybersecurity practices and challenges in this paper. Note that there is no direct link between the cybersecurity challenges in the first column and practices in the second column of Table 1. It is just a summarised list of cybersecurity practices and challenges introduced above.

[1] South Africa. Protection of Personal Information Act 4 of 2013. 2013. https://www.gov.za/documents/protection-personal-information-act. Accessed 25 Oct 2024.

[2] EU. Complete guide to GDPR compliance. https://gdpr.eu/. Accessed 25 Oct 2024.

Table 1 Complex and dynamic cybersecurity operations

Cybersecurity practices	Cybersecurity challenges
Multilayered defence strategies	Evolving threat landscape
Threat intelligence and monitoring	Resource constraints
Incident response and management	Complexity of digital environments
Security automation and orchestration	Data privacy laws and regulations
Compliance and risk management	Incident detection and response
User awareness and training	Supply chain security

Given these challenges, sticking to the old ways of doing things is not enough anymore [27]. We need to think differently about security, blending the unique strengths of humans and AI [28]. Humans bring strategic thinking, contextual understanding, and creativity to the table [29, 30]. At the same time, AI excels in speed, processing vast amounts of data, and spotting patterns [29, 30]. Together, they can form a powerful team better equipped to defend against cyber threats. This paper proposes creating a cybersecurity Augmented Intelligence Framework (*c*AIF) for cybersecurity to make this collaboration as effective as possible. This is the second of the two research objectives in the paper. Augmented intelligence fosters a synergistic human-AI collaboration, enhancing human capabilities rather than replacing them [31]. Unlike previous models that focus on isolated HMI paradigms, the *c*AIF adopts a hybrid approach, strategically blending paradigms to enhance decision-making processes across multiple layers of cybersecurity. This comparative advantage lies in its ability to transition seamlessly between paradigms depending on the complexity and urgency of a given task, making it better suited for complex, real-time cybersecurity operations. This is the theoretical contribution of this paper.

The remainder of this document is organised as follows: Section II delves into the literature surrounding HAIT in cybersecurity, examining the practices in use and the challenges encountered. Section III details the methodologies used in this study to select and analyse specific HMI paradigms and develop a *c*AIF tailored for cybersecurity applications. The results of this research are presented in Section IV, with a separate, in-depth discussion of these findings in Section V. The paper concludes in Section VI, where final thoughts, recommendations for future practices, the study's limitations, and potential avenues for further research are outlined.

2 Foundational HMI Paradigms

A multifaceted approach is indispensable for optimising HAIT in cybersecurity. HAIT refers to systems in which humans and multiple AI agents collaborate to improve the performance of missions that humans or AI systems can achieve alone [32]. Rather than relying solely on a single HMI paradigm, given their strengths and limitations, a strategic combination tailored to specific cybersecurity tasks is imperative. This foundational understanding propels the current research to develop a *c*AIF system integrating multiple HMI paradigms to enhance cybersecurity operations across various domains.

2.1 Human and AI Agents

Over the past few years, we have witnessed a transformation in AI capabilities, a change spurred by leaps in computing power, the widespread availability of data, and significant improvements in algorithms. This evolution has not only reshaped the capabilities of these systems but has fundamentally altered how we, as humans, interact with technology [33]. According to research by Wang et al. [33], we are now looking at a new era of cooperation where humans and AI work together, aiming to achieve common goals and outcomes. If successfully fostered, this partnership hinges on humans understanding how AI systems operate, including recognising when and why they might fail, as Bansal et al. [34] noted. This concept of collaboration casts humans and AI agents as teammates, working side by side. However, this idea sparks debate within the academic community, with scholars questioning whether AI agents can be considered teammates in the same sense as human colleagues [35]. In this context, 'human agents' refer to the people or employees who are part of any system that integrates social and technical elements, as defined by Malatji et al. [36]. This evolving partnership between humans and AI is not just about technology but also about how we reimagine the future of work, collaboration, and innovation.

In contrast to human agents, AI agents represent a different breed of collaborators in our technological landscape. As explored by Xi et al. [37], these are not typical team members. AI agents are essentially artificial entities equipped to sense their surroundings, analyse information, make decisions, and take actions based on that analysis [38]. Imagine them as highly specialised, intelligent partners explicitly designed for specific tasks.[3] Thanks to their design, these agents can continuously operate, which integrates sensors to perceive their environment and actuators to respond

[3] VentureBeat. Beyond assistants: AI agents are transforming the paradigm. 2024. https://venturebeat.com/ai/beyond-assistants-ai-agents-are-transforming-the-paradigm/. Accessed 25 Oct 2024.

accordingly [38]. Their capabilities are further detailed by Russell and Norvig [38] and their role as proactive and autonomous decision-makers capable of operating without human guidance. This paper summarises an AI agent: *a system designed to perceive, analyse, decide, and act autonomously to achieve its specified mission while learning and adapting to its environment*. They are the digital eyes, brains, and hands of our technological endeavours, gathering data, processing it, and acting to meet set objectives. The hallmark traits of these digital colleagues, as identified by [32] and [37], include:

- *Perception*: They are equipped with the means to collect environmental data or specific inputs, such as using sensors, cameras, or parsing through data streams, to identify changes or patterns crucial to their tasks.
- *Decision-making*: Utilising advanced techniques like ML, neural networks, or logic-based rules, they process the gathered information to determine the best course of action.
- *Autonomy*: These agents operate on a spectrum of independence. Some can fully autonomously navigate through their tasks, while others might be semi-autonomous, requiring periodic human oversight.
- *Actions*: Upon deciding, AI agents can take concrete steps to influence their environment or work towards achieving their set goals. This could mean interacting with physical systems like robots, engaging with software, or even delivering insights to human users.
- *Learning*: An essential aspect of AI agents is their ability to evolve by learning from data or experiences over time, enhancing their decision-making processes through sophisticated algorithms.

This nuanced understanding of AI agents illuminates their potential as indispensable allies in our quest to harness technology for better outcomes across various domains. In this regard, various AI agents are beginning to make notable strides across different sectors, showcasing the versatility and potential of these technologies. Chatbots, for instance, engage in simulated conversations with users online, mimicking human interaction with a level of sophistication that continues to improve, as highlighted by Lund and Wang [39]. Then there are autonomous vehicles, which navigate our roads by perceiving their surroundings, making split-second driving decisions, and controlling the car autonomously—a feat elaborated on by Firlej and Taeihagh [40]. Recommendation systems, another form of AI agent, curate content or product suggestions tailored to individual user preferences and behaviours, a subject explored by Zhang et al. [41]. Additionally, robotic process automation has been revolutionising business processes by automating repetitive, rule-based tasks, as discussed by Chakraborti et al. [42]. A particularly compelling application of AI agents is found within cybersecurity, where defensive AI agents proactively hunt and neutralise system threats without human intervention. This capability was notably demonstrated during the Defense Advanced Research Projects Agency (DARPA) Cyber Grand Challenge (CGC) in August 2016, where autonomous offensive and defensive agents engaged in simulated combat, operating entirely without human oversight—a milestone detailed by Maymí and Thomson [43].

Despite these advancements, exploring how humans and AI can collaborate effectively, especially in cybersecurity, remains sparse. Samtani et al. [3] have pointed out that the nuanced interaction between humans and AI in addressing critical cybersecurity challenges is an area ripe for investigation. Such research demands a multidisciplinary approach, incorporating insights from psychology, cognitive science, and HMI, among others, to truly understand and optimise this partnership. Early studies in this field, including work by Sarker et al. [44], Wang [45], and Hauptman et al. [46], have started to shed light on various aspects, such as risk factors, user interface design, and the dynamics of team collaboration between humans and AI agents. Given the limited research explicitly focusing on human-AI collaboration in cybersecurity, this paper further seeks to explore and discuss existing HAIT frameworks, paving the way for a deeper understanding and more effective implementations in the future.

2.2 Human-AI Teaming Literature

As we navigate the digital transformation from Industry 4.0, which focuses on automation and data-driven technologies, we step into Industry 5.0. This next phase marks a significant shift toward centring the human experience, or human-centricity, as highlighted by Xu et al. [47]. Industry 5.0 is not just about technology; it is about integrating sustainability, resilience, ethical considerations, and, most importantly, fostering a collaborative partnership between humans and intelligent machines [48]. This partnership aims to leverage both partners' unique strengths to achieve the best possible outcomes. This paper explores the nuances that differentiate mere interaction from true collaboration within this context. As defined by Hornbæk and Oulasvirta [49], interaction can refer to any form of communication or engagement between entities, whether one-way or reciprocal, without necessarily sharing a common goal. However, as Mourtzis et al. [50] pointed out, collaboration is a specialised form of interaction where parties work together towards a shared objective, characterised by mutual effort, shared responsibility, and open, continuous communication. The essence of collaboration between humans and intelligent machines, or augmented intelligence, is what Industry 5.0 aspires to harness.

In exploring HAIT and augmented intelligence within cybersecurity, I conducted a systematic literature review (SLR) of human-AI collaboration studies between 2014 and 2024. The goal was to infer the type of human-AI relationship in each study, thereby establishing the foundational HMI paradigms. The exact methodology of how I went about it is detailed in the methods section (Sect. 3). Table 2 summarises the SLR results.

As shown in Table 2, I examined various studies to grasp how they weave together different HMI paradigms to bolster cybersecurity across multiple domains. For instance, the works of Sarker et al. [44] investigate the application of HAIT in both information technology (IT) and operational technology (OT) settings. Their research underscores the potential of HAIT to elevate attack detection, streamline incident responses, and strengthen the overall cybersecurity stance. However, they did not present a specific HAIT framework for cybersecurity. In a different vein, Chhetri et al. [55] introduce a framework that leans on AI to automate routine alerts, enhance expert decision-making through AI-driven insights, and foster a collaborative space for addressing intricate, unprecedented threats. Their study mainly addresses the issue of alert fatigue among SOC analysts, who can become bogged down by the overwhelming number of security alerts. This, they argue, can significantly reduce the efficiency in pinpointing and mitigating real threats. Despite its strengths, this proposed framework does not incorporate a variety of HMI paradigms to elevate cybersecurity operations across different domains. In their exploration of the evolving landscape of cyber defence, Gonzalez et al. [66] introduced a novel concept focusing on the synergy between human intellect and AI, which they aptly named the human-AI cognitive teaming framework. Their research focuses exclusively on enhancing cybersecurity defence strategies rather than delving into the broader spectrum of cybersecurity operations across different domains. This specificity offers a deep dive into how we can bolster our digital fortresses against the ever-growing threat of cyberattacks. Vats et al. [65] took the discussion further by probing into the capabilities of large, pre-trained AI models and their role in augmenting human intelligence. The discourse encompasses collaborative efforts for refining AI models, fostering effective teaming between humans and machines, and contemplating the ethical dimensions of such partnerships. Their insights extend beyond cybersecurity, touching upon the broader implications for multiple sectors, thus highlighting the versatile potential of human-AI collaborations.

In addition to the studies in Table 2, a significant portion of the human-AI collaboration literature I reviewed touches upon various facets of IT and OT environments, mainly focusing on single HMI paradigms. Table 2 shows that these are predominantly the Human-in-the-loop and Human-out-of-the-loop paradigms. Despite these studies, there remains an untapped potential in weaving together multiple HMI paradigms to optimise HAIT in cybersecurity, suggesting a fertile ground for future research and development in this critical area of digital defence.

2.3 HMI Paradigms

In my journey through the rich landscape of literature on how humans and intelligent machines interact, mainly through paradigms similar to the HITL approach, I encountered a wealth of insights and frameworks. Scholars such as Holmberg,[4] [67–77], and many others, have contributed significantly to our understanding of this dynamic field. Their works span years and cover a diverse range of perspectives from [69]'s early examination of the out-of-the-loop phenomenon to more recent analyses by Sarker et al. [44] highlighting the importance of incorporating human experts in the loop and [72]'s interdependent and co-active nature of human–machine systems, among others. This review and the SLR results in Table 2 allowed me to distil the information into six overarching foundational HMI paradigms. These paradigms, summarised in Table 3, were mainly derived from the type of human-AI relationship inferred in the last column of Table 2. The HMI paradigms offer a structured way to navigate the complex terrain of human-AI collaboration. In the context of cybersecurity, they describe the multifaceted interactions between humans and AI agents.

Each HMI paradigm provides a unique lens through which we can examine the evolving roles of humans and intelligent machines in various contexts, highlighting the nuanced ways technology and humanity converge. By weaving together the threads of these HMI paradigms, I aim to present a cohesive picture that acknowledges the contributions of each and sets the stage for further exploration into how we can enhance and optimise the partnership between humans and intelligent systems. Each one of these paradigms is explored in the following six subsections.

2.3.1 Human in the Loop

The concept of HITL is not new, and a rich body of research surrounds it. [67] point out that the initial discussions around HITL, whether directly stated or implied, often revolved around control systems. These systems, which aim to manage the operation of other systems, find their use in various settings such as robotics, industrial operations, and transportation networks [78]. As shown in Fig. 1, one of the early

[4] Holmberg L. Human in command machine learning. 2021. https://urn.kb.se/resolve?urn=urn:nbn:se:mau:diva-42576. Accessed 13 Aug 2024.

Table 2 Human-AI relationship type

Author	Study purpose	Findings	Human–machine relationship type
Al-Mansoori and Salem [51]	Analyses trends and applications of AI/ML in cybersecurity domains like threat detection, risk assessment, and automated responses	AI/ML are crucial in developing dynamic defensive systems for cybersecurity as cyber threats evolve	Human in the loop; Human out of the loop
Aliberti et al. [52]	Focuses on enhancing situation awareness in human–machine systems	Presents an innovative approach to enhance situation awareness by projecting current situations into the near future, which helps anticipate developments, including cyberattacks	Human in the loop; Human out of the loop
Belaïd [53]	Discusses the dynamics of shared control in autonomous vehicles, focusing on the need for timely driver reactions during mode and period changes	Highlights the importance of human–machine collaboration in incident response for autonomous vehicles, emphasising the need for effective communication between drivers and systems	Human in the loop; Human out of the loop; Human on the loop; Human alongside the loop; Human in command
Chen et al. [54]	Explores the implications of ML in military contexts, focusing on human–machine systems and also the impact on cyber-HMI	Identifies key areas where ML is being adopted, including intelligence gathering, logistics, and healthcare, and its impact on platforms across terrestrial, naval, aerial, and cyber domains	Human in the loop; Human out of the loop; Human in command
Chhetri et al. [55]	Focuses on enhancing SOCs through human-AI collaboration to address the issue of alert fatigue, which affects SOC analysts due to overwhelming alert volumes	The $\mathscr{A}^2\mathscr{C}$ Framework is proposed for facilitating flexible decision-making by allowing transitions between automated, augmented, and collaborative operations	Human out of the loop; Human on the loop; Human alongside the loop;
Chowdhury et al. [56]	To develop a model in which AI will furnish sufficient information about its decision-making process, enabling human agents to provide feedback for improving the model	Experiments with real data and threat detection tasks show that the model significantly improves the accuracy of existing AI algorithms for these tasks	Human out of the loop; Human on the loop
Desai et al. [57]	To tackle the core issue of limited clarity and explainability in conventional AI models for cybersecurity	Highlights the need to incorporate Explainable AI (XAI) into cybersecurity operations to empower security analysts to understand, validate, and effectively respond to cyber threats	Human in the loop; Human out of the loop; Human on the loop
Gomez, Mancuso and Staheli [7]	Explore ways to apply or enable human–machine teaming where security analysts work alongside machines responsible for some duties or sub-tasks traditionally held by humans for cyber defence	Presents a roadmap of research goals for advanced human–machine teaming in cybersecurity operations	Human out of the loop; Human alongside the loop; Human in the loop
Gore et al. [58]	Investigates the application of augmented intelligence with ML and image processing techniques in cybersecurity	Augmented intelligence improves precision through human judgment and promotes productive collaboration through interactive interfaces	Human in the loop; Human out of the loop
Hauptman et al. [46]	To understand how to identify the amount of autonomous control AI agents have over their decisions and how changes to this control cognitively affect the rest of the team, with a cyber incident scenario as the context	Work cycles can be used to assign autonomy levels to adaptive AI agents based upon the degree of formal processes and predictability of the team's tasks during the cycle, and dynamic, human-like adaptation methods are vital to effective human-AI teams	Human in the loop; Human out of the loop

Table 2 (continued)

Author	Study purpose	Findings	Human–machine relationship type
Hayes and Moniz [59]	To understand explainable AI techniques to enable us to examine AI models' underlying logic	Describes and provides guidelines for XAI in collaborative robotics and cybersecurity domains	Collaboratively working rather than being alongside each other; Human in the loop; Human out of the loop
Karunamurthy, Kiruthivasan and Gauthamkrishna [60]	To explore the deep integration of AI into crucial areas of cybersecurity, such as authenticating user access, enhancing awareness of network situations, monitoring for potentially harmful behaviour, and identifying irregular traffic patterns	Introduces the human-in-the-loop intelligence cybersecurity model to synergise human intelligence with AI	Human in the loop; Human out of the loop
Maymí and Thomson [43]	To understand how new paradigms in human–machine teaming should be developed, including interfaces, types of cybersecurity operators, and XAI	Provides an overview of cyberspace threats and opportunities in the next ten years and how these will impact human–machine teaming	Human in the loop; Human out of the loop
Mikhalevich and Ryjov [61]	Discusses the possibilities of using AI to improve the information security of critical and other information infrastructures based on human–computer technology	Emphasises that the functioning of human-AI integrated systems must remain under strict human control	Human in command
Olla et al. [62]	Explores how human biases and attitudes affect the performance of human-AI teams in cybersecurity	Early study prototyping results indicate that participants prefer an AI teammate over a human in the simulation game setting	Human in the loop; Human out of the loop
Parlapalli et al. [63]	Explores the integration of augmented intelligence with machine learning and image processing to elevate threat detection capabilities and foster effective human–machine collaboration in cybersecurity	Empirically show that ML-based solutions empower cybersecurity systems to detect anomalies effectively, interpret complex network structures, and rapidly identify potential threats	Human on the loop; Human out of the loop
Roch et al. [64]	Understand the specialised security-critical context and the human requirements for successfully designing AI tools and human-AI interfaces	A decision framework for enhancing understanding of the interplay between trust in AI, mainly influenced by its transparency and different levels of autonomy	Human in the loop; Human out of the loop; Human on the loop
Sarker et al. [44]	Explores the broad landscape of AI potentiality in cybersecurity, emphasising its possible risk factors with awareness	Argue that human-AI teaming is worthwhile in cybersecurity, in which human expertise, such as intuition, critical thinking, or contextual understanding, is combined with AI's computational power to improve overall cyber defences	Human in the loop; Human out of the loop
Vats et al. [65]	Surveys the integration of Large Pre-trained Models with HAIT	Emphasises how integrating Large Pre-trained Models with HAIT enhances collaborative intelligence beyond traditional approaches	Human in the loop; Human out of the loop
Gonzalez et al. [66]	To spearhead the generation of dynamic, adaptive, and personalised cyber defence capabilities using deception	Introduced a novel concept focusing on the synergy between human intellect and AI, which they aptly named the human-AI cognitive teaming framework	Human in the loop; Human out of the loop

Table 3 HMI paradigms for human–machine systems

HMI paradigm	Humanrole	AI/Intelligent machine role	Use case
Human-in-the-loop (HITL)	Active participant: intervenes as needed	Performs tasks but relies on human input or validation	Medical diagnosis systems requiring a clinician's approval
Human-on-the-loop (HOTL)	Supervisor: monitors and intervenes when necessary	Operates autonomously but under human supervision	Autonomous vehicles with human oversight
Human-out-of-the-loop (HOOTL)	No real-time intervention: involved only in setup or analysis	Operates fully autonomously	Fully autonomous drones or trading algorithms
Human-alongside-the-loop (HATL)	Collaborator: works in parallel with AI	Complements human efforts by handling specific tasks	Creative industries where AI assists in tasks
Human-in-command (HIC)	Final decision-maker: overrides AI decisions	Provides recommendations or executes tasks autonomously	Military operations with AI recommendations
Coactive Systems	Continuous collaborator: dynamic interaction	Continuously adjusts actions based on human input	Collaborative robots (Cobots) in manufacturing

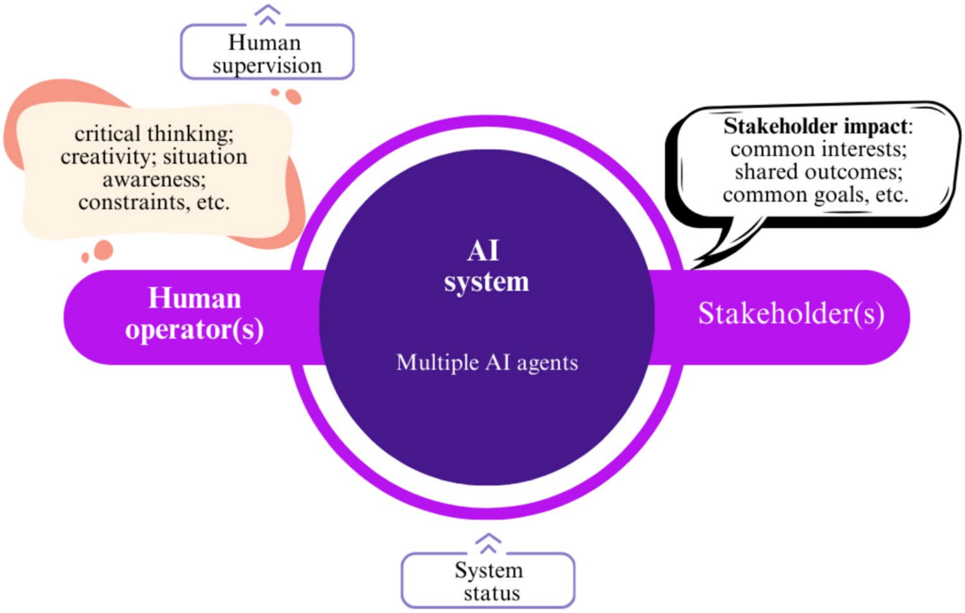

Human supervision
critical thinking; creativity; situation awareness; constraints, etc.
Stakeholder impact: common interests; shared outcomes; common goals, etc.
AI system
Multiple AI agents
Human operator(s)
Stakeholder(s)
System status

Fig. 1 Human-in-the-loop paradigm

concerns identified with HITL in military engineering was the unpredictability of human behaviour due to factors such as situation awareness.

Human operators actively engage in the control loop, adjusting their input and modifying their responses dynamically based on the specific task requirements [67]. This contrasts with passive monitoring, as human operators play a critical role in supervising automated systems, managing unforeseen situations, optimising performance, and executing necessary maintenance procedures [75]. Therefore, the human involvement level is very high within the HITL paradigm. In discussing Fig. 1, the concept of human supervisory control warrants further explanation. This idea essentially looks at how humans occasionally interact with machines while continuously getting updates and information from them [79]. According to Sheridan [79], this interaction vividly shows how we work alongside AI in cybersecurity efforts. Imagine this teamwork as a dynamic system where humans and AI play crucial roles. AI might be doing a lot of the heavy lifting, performing various tasks autonomously. However, it depends on humans to steer it in the right direction, validating its actions and ensuring everything is on track. This mutual exchange of information involves continuously sharing task-relevant data, system feedback, and contextual insights. For example, AI systems provide humans with real-time analytics, anomaly detection alerts, and pattern recognition outputs. In contrast, humans offer contextual understanding, override decisions, and provide corrective feedback to improve AI performance and adaptability.

Take, for example, AI systems used in medical diagnoses. These systems can sift through data and suggest diagnoses, but a healthcare practitioner's expertise is crucial to approving these suggestions [80]. Similarly, think about how we deal with email spam filters. They might flag certain emails as spam, but occasionally, we need to jump in and correct them when they mistakenly tag legitimate emails [81]. This interaction showcases how these HITL systems fundamentally rely on us to oversee and guide essential parts of their processes. This setup ensures that humans are not just bystanders but active participants who have the final say in AI systems' decisions. As highlighted by Fanni et al. [82], and reinforced by the European Commission,[5] AI is here to assist, offering insights and suggestions rather than making unilateral decisions. This collaborative approach ensures that AI systems serve to enhance human decision-making, keeping us firmly in control.

2.3.2 Human Out of the Loop

When we turn our attention to the HOOTL paradigm, we are venturing into territory where human intervention is markedly less prevalent, essentially the flip side of the HITL approach. In this scenario, as Docherty et al. [83] highlighted, systems are given the autonomy to operate without human input or oversight during their routine operations. Berberian et al. [69] have raised a vital flag about this level of automation, especially regarding safety–critical systems. Their concern stems from a phenomenon known as the 'out of the loop' performance problem. This issue, further dissected by researchers like Parasuraman and Wickens [84] and Endsley [85], primarily revolves around a significant drawback—the erosion of situation awareness. When humans are removed from the operational loop, they might struggle to detect and comprehend issues as they arise and find it challenging to develop effective solutions due to their lack of engagement with the system's ongoing operations [69].

Examples of the HOOTL paradigm in action include fully autonomous drones [86], trading algorithms [87], and SOAR tools for autonomous threat-hunting. These instances underscore intelligent systems where human involvement is confined mainly to the drawing board, during the conceptualisation, design, and development phases. Fanni et al. [82] describe this as a form of passive human agency, where humans set the stage for autonomous operations but step back during the live action, entrusting the system to manage itself. The HOOTL paradigm underscores a delicate balance between leveraging technological advancements for efficiency and the critical need for human oversight to ensure safety and reliability, particularly in high-stakes situations. For example, in fully autonomous AI-driven threat-hunting tools used in SOCs [88], human participation tends to be minimal following the initial setup phase. Once established, these systems are responsible for autonomously monitoring, identifying, and addressing potential threats.

However, this approach can lead to challenges related to trust in AI decision-making [89] due to its susceptibility to biases stemming from the quality of input data or the design of algorithms [6, 39]. For example, instances where false positives occur [81] or anomalies that necessitate human interpretation may undermine confidence in the system's accurate decision-making process. As Hou et al. [89] discuss, building trust in autonomous systems is paramount for their successful integration into critical workflows. An emerging solution to this trust issue is XAI, which Gunning et al. [90] advocate for. As AI systems become increasingly integrated into cybersecurity defences, ensuring transparency and accountability through XAI mechanisms is critical. XAI enables human operators to understand AI decisions, enhancing trust and making it easier to justify decisions during audits or incidents [91]. Moreover, ethical concerns about autonomous AI decisions [92], especially in sensitive domains like healthcare and finance [80, 93], require careful consideration. These challenges and evolving cyber threats underscore the need for an adaptive, human-XAI partnership.

2.3.3 Human on the Loop

In exploring the nuances of human–machine collaboration, we encounter a concept known as the HOTL paradigm. While similar to the HITL approach, this paradigm positions the human operator in a slightly more detached, yet still critical, oversight role. Li et al. [73] provided an insightful examination of this model, highlighting its distinctive nature. The essence of HOTL lies in creating a balance where humans oversee automated, self-adapting systems without being immersed in every minutia of their operation. Anderson and Fort [67] described this relationship as one where humans engage with technology through rather than in these systems. This distinction is pivotal. Self-adaptive systems, by design, are engineered to autonomously adjust their functions and behaviours in real-time as they respond to evolving environmental conditions [73]. This capability ensures they remain effective and efficient without constant human intervention [67, 73]. This concept is illustrated in Fig. 2, where the human operator watches the system's interaction with its surroundings.

The operator's role is more about periodic oversight and less about the direct, continuous control characteristic of the HITL paradigm. Intervention is strategic and occurs only when necessary, such as when the system's behaviour might deviate into the realm of the unexpected or potentially

[5] European Commission. Ethics guidelines for trustworthy AI. Publications Office of the European Union. 2019. https://data.europa.eu/doi/10.2759/346720. Accessed 25 Oct 2024.

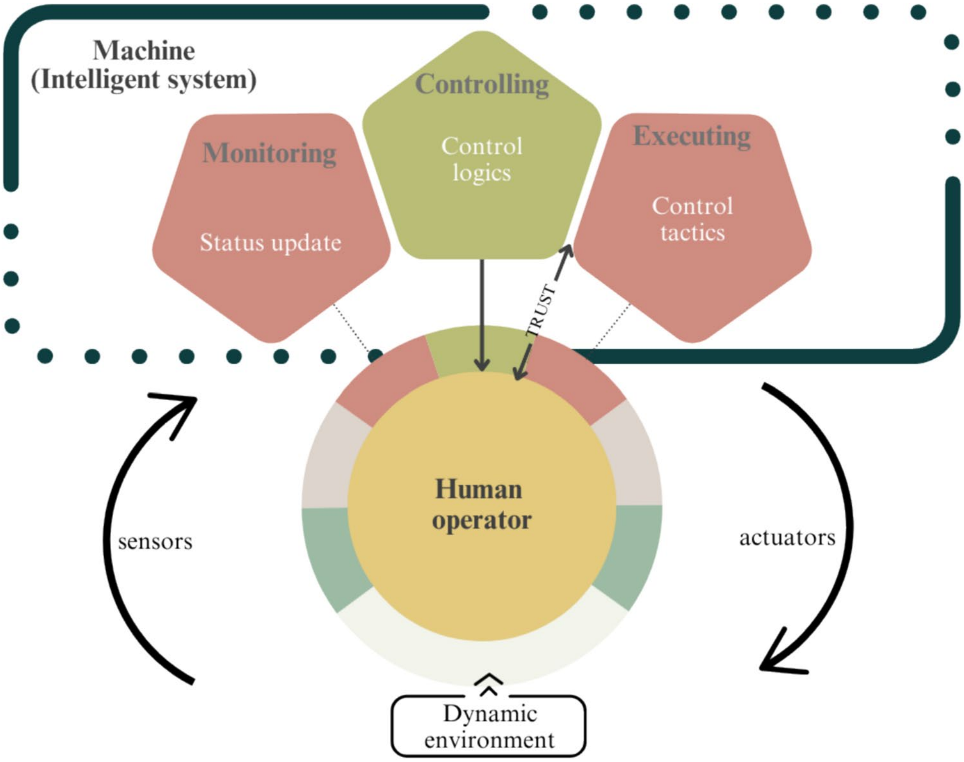


Fig. 2 A human-on-the-loop self-adaptive system

problematic [73]. The 'trust' arrow represents the intervention action by the human operator between the controlling and execution stages. It is essentially about the human operator's confidence in the autonomous machine's decision and the explainability (XAI) of such decisions [73]. In other words, if time passes without explanation, the complex analysis and planning of the autonomous machine will probably make human operators lose trust, i.e., reducing the probability of true-positive and true-negative. This HOTL paradigm represents a significant stride towards leveraging technology's potential while ensuring human judgment remains a cornerstone of critical decision-making processes [82]. It acknowledges the strengths of humans and machines, seeking to optimise their collaboration for better outcomes. A use case example of the HOTL paradigm is autonomous vehicles with human oversight.

2.3.4 Human Alongside the Loop

The HATL approach champions the idea that AI excels at tasks well-suited for machines, like analysing large volumes of data in real-time, while humans take on roles that require judgment, creativity, or empathy. Engstrom and Ho [71], in a study commissioned by the Administrative Conference of the United States (ACUS) of America, shed light on how federal agencies could harness the potential of algorithmic tools in governance. The ACUS is an independent agency of the USA government dedicated to improving federal agency administrative processes and procedures. These researchers proposed a novel way to leverage HATL: having human and AI-generated decisions reviewed together, allowing a human reviewer to make the final call. This process ensures a thorough evaluation and enriches the AI model with diverse insights, essentially teaching the AI through exposure to human decision-making. Engstrom and Ho's [71] argument is that pairing traditional human judgment with AI's analytical prowess can significantly enhance decision-making processes. As Machireddy et al. [94] pointed out, relying solely on historical data can restrict an AI model's understanding and adaptability to new trends or anomalies. Thus, pairing traditional human judgment with AI's analytical prowess provides AI systems with what is essentially exogenous training data, helping to overcome the limitations of models that rely solely on historical data and predefined rules [71].

However, introducing AI as a quasi-team member in such an integrated process comes with challenges. Berberian et al. [69] have delved into the complexities of this partnership, highlighting the new coordination demands and the potential for misunderstandings or failures in the human–machine relationship. They liken the role of humans in this setup to passive information processors who must stay vigilant and ready to identify and address any deviations or issues that arise. This vigilant oversight is crucial for ensuring that the AI systems operate within expected parameters and maintaining a smooth partnership between humans and machines. Like the HOOTL scenario, trust in AI decision-making is another factor in HATL. This evolution from direct human involvement (HITL) to a more supervisory role (HOTL) and eventually to a system where humans primarily review outcomes (HOOTL and HATL) necessitates a leap of faith in AI's reliability and responsibility. Campbell [70] echoes this sentiment, suggesting that progressing towards less direct human oversight will require time and proven reliability.

2.3.5 Human in Command

In the evolving landscape of AI, the HIC paradigm represents a significant shift towards emphasising human oversight and decision-making authority over AI systems. Anderson and Fort [67] point out that the HIC concept is relatively new, having emerged around 2019. This term gained prominence through the work of the European Commission's High-Level Expert Group (HLEG) on AI, tasked with crafting guidelines for trustworthy AI. The HLEG's definition of HIC underscores the importance of human oversight across all aspects of AI system operations. This includes the immediate outcomes of AI decisions and their broader economic, societal, legal, and ethical impacts [92, 93]. The degree of human control sets the HIC paradigm apart, especially compared to the HITL approach. In HITL settings, while AI systems perform tasks and make recommendations, they still rely on human input or validation. However, under the HIC model, the human operator is not just another checkpoint; they are the ultimate authority, with the power to override AI decisions as they see fit. This delineation emphasises a more active form of human agency, as noted by Fanni et al. [82].

A practical example of the HIC paradigm in action can be seen in military operations, where AI-powered autonomous weapon systems might suggest strategies or actions. Still, the final decisions rest firmly in the hands of human commanders as to when, where and why the weapon will be employed [95]. This approach ensures that, despite the advanced capabilities of AI, the human perspective, rooted in ethical, legal, and societal considerations, remains at the forefront of decision-making processes. The HIC paradigm, therefore, champions a vision of AI as a tool for enhancing human judgment rather than replacing it [29], ensuring that technology serves humanity in the most responsible way possible.

2.3.6 Coactive Systems

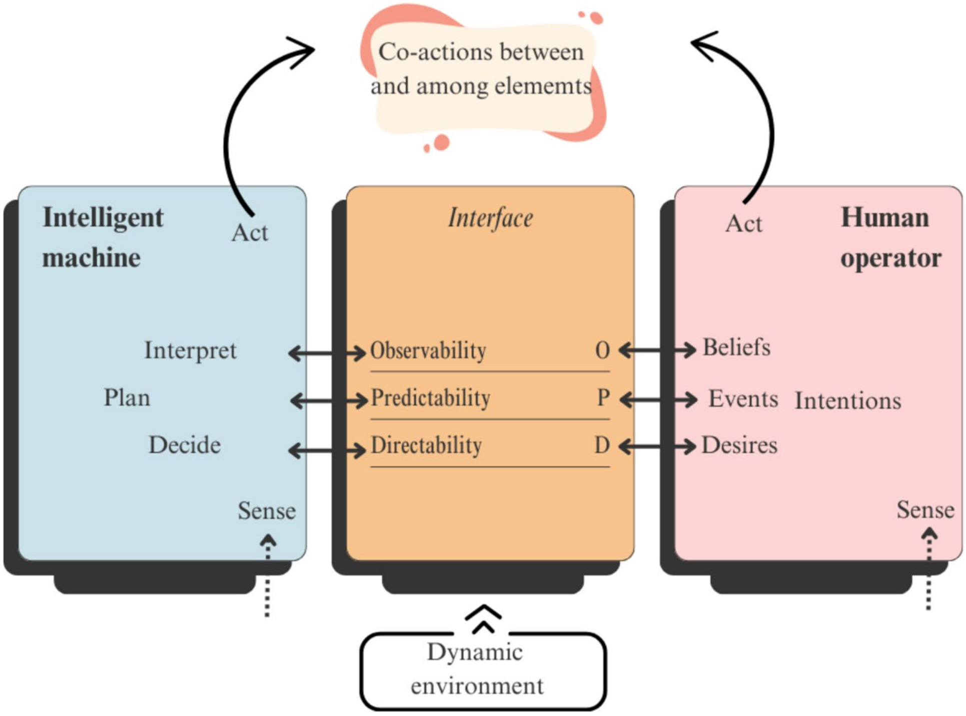
In the intricate dance of collaboration between humans and technology, Coactive Systems emerged as a testament to the power and potential of truly integrated teamwork. The essence of coactivity lies in its focus on interdependence, highlighting a relationship where humans and intelligent machines are not just working alongside each other (HATL) but are deeply intertwined in their efforts [72]. This concept pushes us to rethink the dynamics of HMI, aiming for a seamless blend of capabilities that enhance the team's overall performance. It is about continuous and mutual contribution. Here, humans and intelligent machines are equal partners in the task, each bringing unique strengths. The interaction between the two is dynamic, characterised by real-time feedback and adjustments. This level of integration ensures that the system can adapt and evolve in response to new information or changing conditions, making it both flexible and resilient. Tai [96] shed further light on how such systems can achieve their full potential. By carefully aligning each team member's capabilities and limitations and harnessing their interdependence's power, Coactive Systems are designed to optimise the collective output. This approach requires a deep understanding of the nuances of human and machine contributions, recognising that each brings something invaluable to the collaboration.

Further emphasising this point, Coactive Systems can be viewed through the lens of the International Standardisation Organisation[6] as an intelligent human–machine system, where continuous interaction between humans and machines, as system elements, and interoperability are fundamental. This perspective is supported by the work of Jarrahi [29] and Jarrahi et al. [30], who highlight the importance of leveraging the respective strengths and weaknesses of each participant to enhance the system's overall capability. Similarly, Kaur et al. [28] support this perspective and provide insights into such systems' practical applications and benefits. A simplified depiction of a Coactive System in Fig. 3, as illustrated by Johnson et al. [72], serves as a visual representation of this concept.

At the heart of the Coactive System model lies a trio of principles known as observability, predictability, and directability (OPD)—concepts introduced by Johnson et al. [72]. These principles are the linchpins of how humans and intelligent machines interact within such systems, ensuring that teamwork is not just possible but highly effective. The first principle, observability, is about transparency. It emphasises the importance of team members, whether human or machine, making their status, knowledge, and understanding of the team's objectives, tasks, and surrounding environment visible to one another. This openness ensures that everyone, or every 'thing', in the team is on the same page, fostering a sense of unity and shared purpose. Predictability, the second principle, focuses on reliability. In any team, anticipating each other's actions is crucial. This predictability does not mean actions are rigid or robotic; instead, it suggests that actions are consistent enough that team members can

[6] ISO/IEC/IEEE 15288:2023. ISO/IEC/IEEE 15288:2023(en), Systems and software engineering—System life cycle processes. 2023. https://www.iso.org/obp/ui/en/#iso:std:iso-iec-ieee:15288:ed-2:v1:en. Accessed 25 Oct 2024.

Fig. 3 Coactive system



count on them when planning their next steps. This reliability is critical in dynamic environments where timing and coordination are critical for success. Directability, the third cornerstone, revolves around influence and adaptability. It pertains to the ability of team members to guide each other's actions and, in turn, be open to direction. This mutual guidance allows for a fluid, adaptive approach to achieving goals, where feedback and instructions can be given and received to enhance the team's overall effectiveness.

Johnson et al. [72] point out that not every task requires all three OPD elements to be actively supported, suggesting that the application of this framework should be tailored based on the specific requirements of each task. Tai [96] emphasised that OPD is a guideline for facilitating interdependent interactions between humans and machines, enhancing their collaboration during task execution. Digging deeper, the Coactive Systems theory, according to Mascolo [97], suggests that our actions and growth are outcomes of interactions not just among individuals but between individuals and their environments. Stein [76] describes these systems as inherently inter-participatory and mutually inclusive. This means that the system elements involved in producing actions and experiences—be they people, machines, or environmental factors—are deeply interconnected, influencing and shaping each other in a continuous, dynamic process. This exchange concept within Coactive Systems is not a simple give-and-take or mirrored reflection but a complex, ongoing activation and reactivation of all system elements. It is a process where interactions are not just happening in parallel but are deeply intertwined, each affecting and being affected by the other. In essence, Coactive Systems embody a philosophy where the collaboration between humans and intelligent machines is seen as a living, evolving dance—a partnership where both parties learn, adapt, and grow together.

2.4 Key Components of the HMI Paradigms

The section we have just discussed offers a rich exploration of the varied and intricate ways humans and AI systems can work together. These collaborative models are tailored for various scenarios, each with unique demands for autonomy, human input, and consideration of potential risks. To make sense of this complexity, I have compiled Table 4, which captures the essence of six distinct HMI paradigms. This table draws on the insights and research findings from a host of field experts [67–77], providing an overview of the critical components that define each paradigm.

Each HMI paradigm in Table 4 offers unique contributions toward creating a cybersecurity human-AI conceptual framework, the *c*AIF. A framework designed to bolster the adaptability and effectiveness of human-AI cybersecurity measures across a diverse range of sectors. The research process I utilised to develop the *c*AIF, including the methods and considerations involved, is outlined in the following section.

Table 4 Key components of HMI paradigms

HMI paradigm	Level of human intervention	Key component
HITL	High	Humans are active participants Iterative process of interaction Decision-making shared
HOTL	Moderate	Humans are overseers Humans intervene intermittently
HOOTL	Low, to non-existent	System autonomy Predefined algorithmic rules
HATL	Medium	Concurrent operations Shared workspace Complementary roles Continuous collaboration
HIC	High	Humans are the ultimate authority Human-centred design System as a support tool
Coactive Systems	High, and ongoing	Shared control Mutual adaptation Emergent behaviour

3 Methods

At this juncture, I would like to emphasise that two main research objectives (ROs) drive this study:

- *RO 1*: My first aim is to delve into the six selected HMI paradigms. I seek to understand their unique advantages and limitations, evaluating their fitness for use with HAIT strategies to fortify cybersecurity efforts in today's digital era.
- *RO 2*: My second objective is to develop a HAIT framework based on the insights gained from RO 1. This hybrid model aims to leverage the identified strengths and overcome the limitations of the selected HMI paradigms, ultimately enhancing cybersecurity measures across different domains.

I have adopted a specific approach to meet these objectives, ensuring a thorough exploration of these areas.

3.1 Study Approach

Three phases were executed to address the two research objectives of this paper:

- Phase 1: Selecting the HMI paradigms
- Phase 2: Identifying strengths, limitations, and suitability of HMI paradigms
- Phase 3: Developing the cybersecurity augmented intelligence framework

3.1.1 Phase 1: Selecting the HMI Paradigms

Figure 4, derived from Page et al. [98], summarises the systematic literature review process followed in this study.

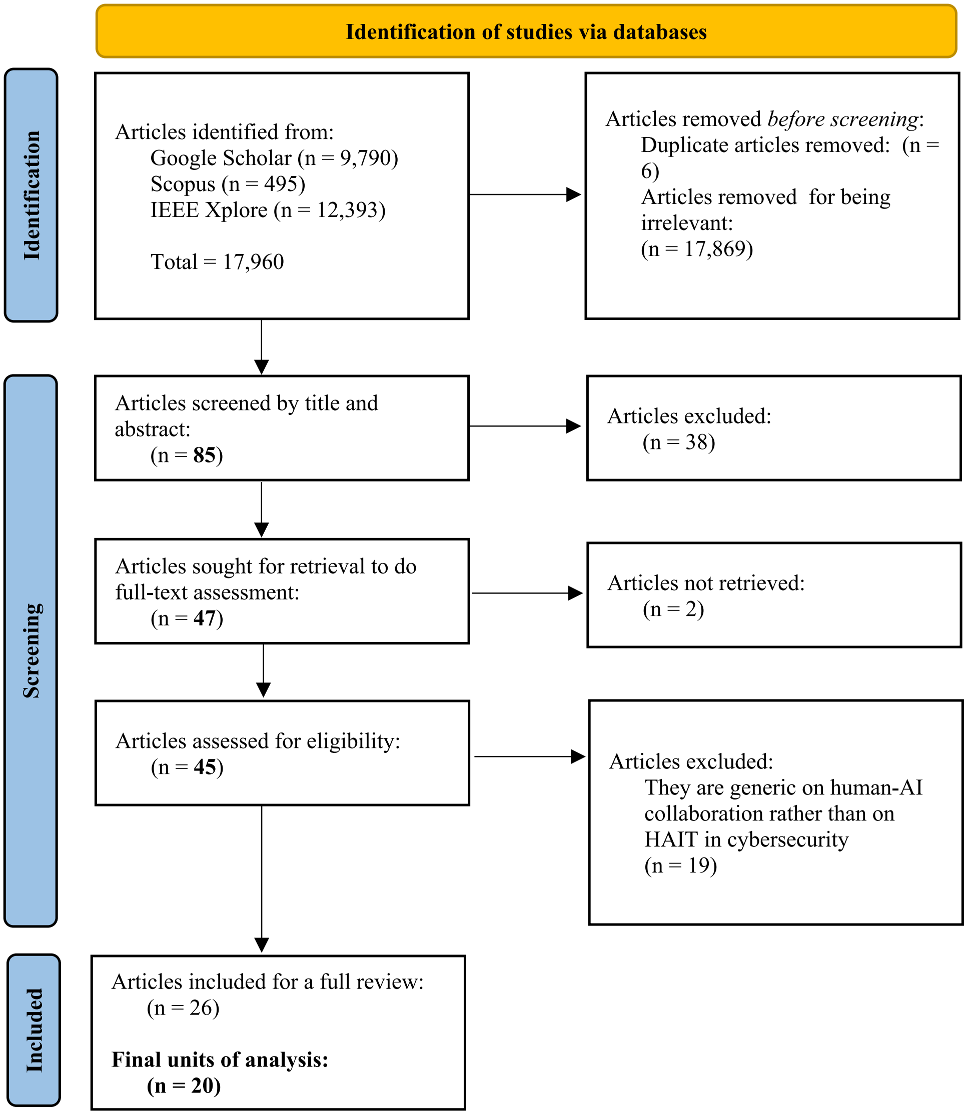
To ensure academic rigour and transparency in identifying and selecting relevant literature, this study adopted a systematic approach guided by the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework [99]. The PRISMA framework emphasises clear documentation of the search strategy, inclusion/exclusion criteria, and screening process to enhance reproducibility and credibility [98]. The literature search was conducted using Google Scholar, Scopus, and IEEE Xplore databases, chosen for their relatively good coverage of scholarly publications related to AI, HMI and cybersecurity. To ensure a broad yet focused coverage, the following primary search terms were used, combined with Boolean operators to refine the search:

- "Human-AI" OR "human-artificial intelligence teaming"
- "Human–machine interaction" OR "human–machine teaming"
- "AI collaboration" OR "cybersecurity AI"
- "Augmented intelligence" AND "cybersecurity"

Generally, the search string is as follows: *"(human-AI OR human-artificial intelligence teaming OR human–machine interaction OR human–machine teaming OR AI collaboration) AND (cybersecurity OR augmented intelligence)."*

Because each database has its specific syntax requirements, below is the search string utilised between 06 and 13 January 2025:

Fig. 4 PRISMA flow diagram for the study


Identification of studies via databases
Identification
Articles identified from:
Google Scholar (n = 9,790)
Scopus (n = 495)
IEEE Xplore (n = 12,393)
Total = 17,960
Articles removed *before screening*:
Duplicate articles removed: (n = 6)
Articles removed for being irrelevant:
(n = 17,869)
Screening
Articles screened by title and abstract:
(n = **85**)
Articles excluded:
(n = 38)
Articles sought for retrieval to do full-text assessment:
(n = **47**)
Articles not retrieved:
(n = 2)
Articles assessed for eligibility:
(n = **45**)
Articles excluded:
They are generic on human-AI collaboration rather than on HAIT in cybersecurity
(n = 19)
Included
Articles included for a full review:
(n = 26)
Final units of analysis:
(n = 20)

- Scopus (Search within: All fields)*: "(human-ai OR human-artificial AND intelligence AND teaming OR human–machine AND interaction OR human–machine AND teaming OR ai AND collaboration) AND (cybersecurity OR augmented AND intelligence)."*
- IEEE Xplore (Advanced search: Full text and metadata): *"(human-AI OR human-artificial intelligence teaming OR human–machine interaction OR human–machine teaming OR AI collaboration) AND (cybersecurity OR augmented intelligence)."*
- Google Scholar (Advanced search: with all of the words; Where my words occur: anywhere in the article): *"(human-AI OR human-artificial intelligence teaming OR human–machine interaction OR human–machine teaming OR AI collaboration) AND (cybersecurity OR augmented intelligence)."*

The search was limited to publications in English, peer-reviewed articles and conference papers, and articles published between 2014 and 2024 to capture recent advancements in the field. The search yielded 9,790 results in Google Scholar, 495 in Scopus, and 12,393 in IEEE Xplore. After eliminating duplicates and irrelevant papers, 85 unique records were identified for the screening and selection process, utilising the Zotero reference management software. The inclusion criteria required articles to:

- Be written in English
- Address human-AI collaboration, HMI, or related paradigms.
- Include applications or theoretical discussions relevant to cybersecurity.
- Propose or evaluate frameworks, models, or paradigms.

Studies were excluded if:

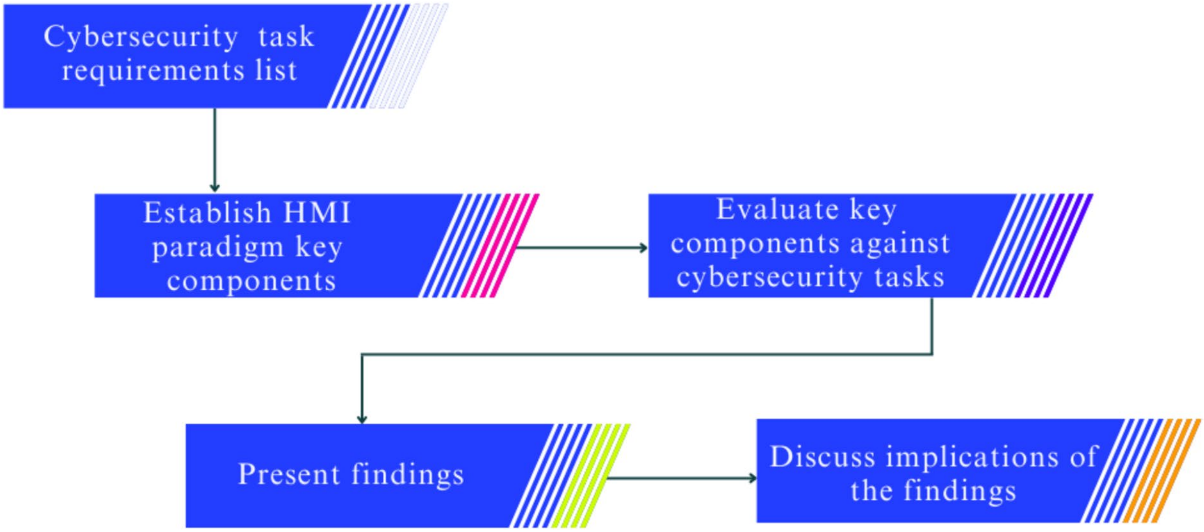

Cybersecurity task requirements list
Establish HMI paradigm key components
Evaluate key components against cybersecurity tasks
Present findings
Discuss implications of the findings

Fig. 5 Approach to exploring the suitability of HMI paradigms

- They focused solely on AI applications outside of human interaction.
- They were unrelated to cybersecurity or augmented intelligence.

Regarding the article screening process, I know PRISMA techniques traditionally require multiple reviewers. However, with improvements in the natural language processing (NLP) capabilities of AI chatbots, AI excels in speed, data processing, and pattern recognition [29, 30], and this is unmatched by humans or any number of reviewers using traditional techniques. While AI systems offer unparalleled capabilities, their susceptibility to biases necessitates a human-centred approach that integrates human insight and values. The author/reviewer provided this in the screening process, utilising the SciSpace AI-powered tool. Thus, the screening process involved one reviewer utilising the SciSpace tool to summarise the articles. The author then provided human insight by manually reviewing the summaries of all 26 papers to determine if they met the inclusion criteria. Upon this review, six papers were excluded because they were either focused on AI applications outside of human interaction or unrelated to cybersecurity or augmented intelligence. Thereafter, a complete manual review was conducted on the final 20 units of analysis. These were listed earlier in Table 2.

3.1.2 Phase 2: Identifying Strengths, Limitations, and Suitability of HMI Paradigms

To identify the strengths, limitations, and suitability of HMI paradigms, I followed the steps outlined in Fig. 5.

The introductory section of this paper describes the landscape of dynamic cybersecurity operations, encompassing both practices and challenges across diverse domains, as encapsulated in Table 1. An analytical evaluation was conducted to understand various HMI paradigms' strengths, limitations, and appropriateness for cybersecurity operations. This entailed an assessment of each HMI paradigm’s key components in Table 4 against the cybersecurity operations outlined in Table 1. The advanced capabilities of the large language model, ChatGPT 4o, were employed to extract and analyse the relevant textual data. The primary objective was to ascertain the most fitting HMI paradigm for various cybersecurity tasks. Following the identification of the optimal paradigm, an intelligent framework (cAIF) was proposed to enhance the efficacy of human-AI cybersecurity collaborations.

3.1.3 Phase 3: Developing the Augmented Intelligence Framework

The ultimate goal of proposing an intelligent framework in cybersecurity is to represent a multifaceted approach to human-AI collaboration that acknowledges the unique contributions of human and AI agents. I followed the steps outlined in Fig. 6 to develop the *c*AIF framework.

It warrants emphasis that the strategic integration of diverse HMI paradigms to develop the *c*AIF is contingent upon the *specific requirements* of individual cybersecurity tasks. Table 1 delineates merely the overarching cybersecurity tasks and operations as they pertain to the scope of this paper. This representation is admittedly simplified when contrasted with the intricate nature of cybersecurity operations as delineated in the National Institute of Standards and

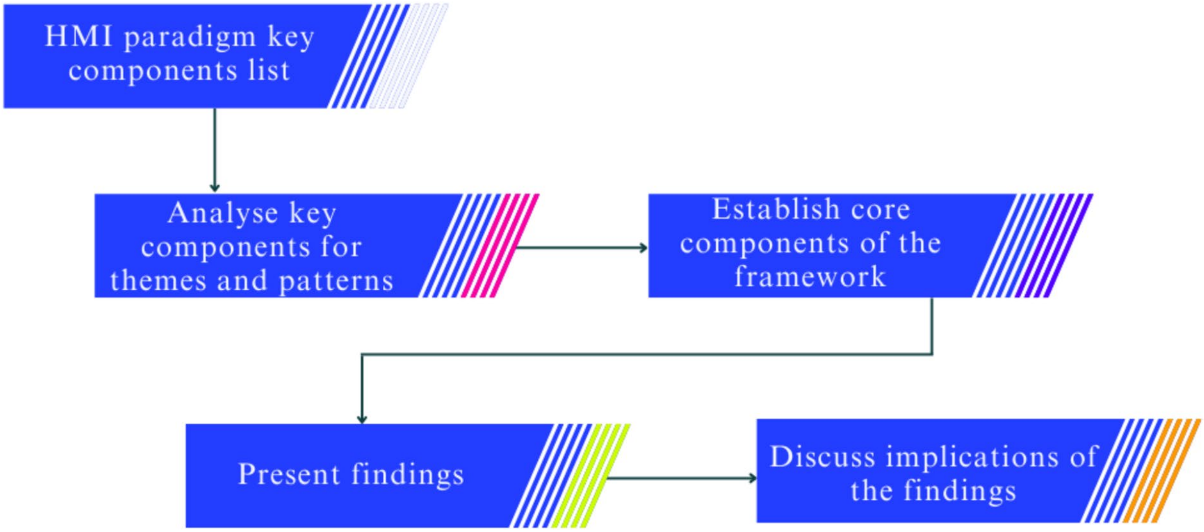

HMI paradigm key components list
Analyse key components for themes and patterns
Establish core components of the framework
Present findings
Discuss implications of the findings

Fig. 6 Approach to developing an augmented intelligence framework

Technology (NIST)'s Cybersecurity Framework 2.0 (NIST CSF).[7] Consequently, AI agents must be customised to specific tasks to facilitate optimal collaboration with human counterparts within the proposed *c*AIF.

3.2 Data Analysis Procedure

I employed ChatGPT 4o as a data analysis tool to assist in processing and identifying patterns across selected HITL, HOOTL, HOTL, HATL, HIC, and Coactive Systems literature. ChatGPT 4o's NLP capabilities were leveraged to analyse qualitative data and synthesise critical insights from the reviewed literature, enhancing the objectivity and efficiency of the data analysis process.

3.2.1 HMI Paradigm Strength, Limitation, and Suitability Data Analysis

The HMI paradigm textual data (Tables 3 & 4) was analysed against cybersecurity operations textual data (Table 1). This analysis determined each HMI paradigm's strengths, limitations, and suitability for each cybersecurity operations task.

3.2.2 Framework Development Data Analysis

An in-depth analysis of the textual data relating to these paradigms was conducted to determine the appropriateness of specific HMI paradigms for designated cybersecurity tasks. This analysis aimed to identify the *core components* of the *c*AIF. The rationale behind pinpointing these core components was predicated on the understanding that cybersecurity tasks vary considerably, from routine threat detection to intricate incident response activities. These variations necessitate differing levels of human participation and AI autonomy. By integrating the strengths inherent in various HMI paradigms, the objective was to architect a flexible and adaptive framework to meet cybersecurity operations' diverse and changing demands. To facilitate this analysis, ChatGPT 4o was employed to scrutinise the HMI paradigm textual data (Tables 3 and 4 and the outcomes of RO 1), assisting in conceptualising potential architectural designs for the framework. The ChatGPT 4o input prompt was as follows:

"The findings from the initial research objective predominantly revolved around the strengths, limitations, and appropriateness of the six HMI paradigms for fostering human-AI collaboration in cybersecurity. The conclusion was that strategically integrating these paradigms would maximise the synergy between human expertise and AI efficiency. With this insight, the ambition is to design and construct a hybrid human-AI teaming framework, the 'Augmented Intelligence Framework', that incorporates multiple paradigms tailored to various cybersecurity tasks' unique, albeit complex and dynamic, requirements. The task is to design this framework's architecture and justify each core component identified."

Following this, the findings from the analysis of the first and second research objectives are presented, offering an understanding of the designed framework.

[7] NIST. The NIST Cybersecurity Framework 2.0. 2024. https://www.nist.gov/cyberframework. Accessed 25 Oct 2024.

Table 5 HMI paradigm suitability for dynamic and complex cybersecurity operations

HMI paradigm	Strength	Limitation	Suitability in cybersecurity operations
HITL	High accuracy due to human validation; Useful for complex decisions	Slower response time; Scalability challenges	Multifaceted defence mechanisms Comprehensive risk and compliance management Proactive incident response and management
HOTL	Balances autonomy with human oversight; Enables swift interventions	Potential for human error in oversight; Requires constant attention	Continuous threat detection and intelligence gathering Automated security operations and orchestration
HOOTL	Fully autonomous, enabling rapid response; Efficient for routine tasks	Lack of human oversight can lead to unaddressed anomalies; Trust issues	Automated security operations and orchestration Continuous threat detection and intelligence gathering
HATL	Collaborative; Combines human judgment with AI efficiency	Potential for role confusion; Requires well-defined task allocation	Multifaceted defence mechanisms Robust user security education and training
HIC	Ensures human control over critical decisions; High accountability	Slower decision-making; May limit AI's potential	Comprehensive risk and compliance management Proactive incident response and management
Coactive Systems	Dynamic and adaptive; Real-time collaboration enhances effectiveness	Complex to implement; Requires advanced integration of human-AI tasks	Multifaceted defence mechanisms Continuous threat detection and intelligence gathering Proactive incident response and management

4 Analysis Findings

The results of the strengths, limitations, and appropriateness of the six HMI paradigms for facilitating collaboration between humans and AI systems in cybersecurity are detailed in Sect. 4.1, addressing the initial research objective of this study. Correspondingly, the identification and explication of the *c*AIF's core components are described in Sect. 4.2, responding to the second research objective of this paper.

4.1 HMI Paradigm Suitability Results

The cybersecurity operations were outlined in Table 1, while Tables 3 and 4 reviewed the key components of the HMI paradigms. An analysis juxtaposing the key components of the HMI paradigms with the fundamental cybersecurity operations culminated in the findings presented in Table 5.

Upon determining the appropriateness of various HMI paradigms, a subsequent analysis of the findings in Table 5 and the textual data encapsulated in Tables 3 and 4 was undertaken. This analysis aimed to identify and outline the architectural core components of the *c*AIF. The outcomes of this analytical process are outlined in the subsequent section.

4.2 Architectural Core Components of the Framework

The analysis of Tables 3, 4, and 5 textual data yielded five core components in Table 6.

The results in Table 6 indicate that the core components of the *c*AIF are:

- Task-Specific Modules
- Decision-Making Matrix
- Paradigm Allocation Engine
- Interoperability Framework
- Feedback and Learning System

It is imperative to underscore that *all six* HMI paradigms engage with each of the five core components of the *c*AIF. Listing one or two core components alongside each HMI paradigm in Table 6 does not denote these components as exclusively relevant to the respective paradigms. Instead, this tabulation reflects the extraction of core components through thematic analysis and pattern identification within the textual data. The implications of these findings are elaborated upon in the subsequent section.

Table 6 Core components of the *c*AIF

HMI paradigm	Themes and patterns	Core component
HITL	Humans and AI agents collaborate in decision-making Humans consistently engage with the AI agents by providing feedback to refine system outputs	Decision-making matrix Feedback and learning system
HOTL	The AI agents operate independently most of the time Human agents step in only when necessary	Paradigm allocation engine Interoperability framework
HOOTL	Full control by the AI agents without human agent intervention	Paradigm allocation engine Interoperability framework
HATL	Human and AI agents work simultaneously on different tasks	Task-specific modules
HIC	Humans have the final decision-making power The AI agents are tailored to human needs and preferences	Decision-making matrix
Coactive Systems	Humans and systems collaborate equally AI agent behaviour is influenced by human agent input and vice versa	Task-specific modules Feedback and learning system

5 Discussions of the Findings

The present study is concerned with developing a *c*AIF system that strategically integrates six HMI paradigms, intending to augment AI-driven cybersecurity operations across various domains. Within this segment, an interpretive analysis of the findings presented in the preceding section leads to the inductive construction of the *c*AIF.

5.1 Implications of HMI Paradigm Suitability

This study sheds light on the nuanced reality that no single HMI paradigm is a panacea for all cybersecurity tasks. For instance, the HITL paradigm stands out in scenarios demanding intricate decision-making processes, such as deploying complex defence strategies and thorough risk and compliance management. This prominence is attributed to its foundation on human validation, capitalising on human capabilities in contextual comprehension, strategic analysis, and creativity, as highlighted by Jarrahi [29] and Jarrahi et al. [30]. However, the HITL approach faces limitations in scalability. It may also exhibit slower response times compared to the rapid pace of AI, which excels in processing vast datasets and identifying patterns but is restrained by the necessity of continual human validation in the HITL scenario. Conversely, the HOOTL paradigm is ideally suited for automating routine security operations, including adopting SOAR tools. This approach facilitates swift action without human delay, enhancing continuous threat detection and intelligence efforts [16]. However, it falls short in adaptability, a hallmark of human intervention, particularly when confronting novel or unexpected threats. Trust in the HOOTL system's ability to manage critical workflows autonomously remains a pressing concern, as underscored by [70].

The HOTL paradigm was recognised for its efficacy in continuous threat monitoring and automated security measures, effectively balancing AI independence with human oversight. This model, however, is not without its vulnerabilities, particularly to human error, and demands consistent vigilance from operators. The challenge arises during periods of AI-led task execution, where a potential diminishment in human situational awareness can occur, leading to oversight failures akin to those observed in HOOTL scenarios [69, 84, 85]. Additionally, the HATL model promotes a symbiotic relationship, exemplified in initiatives like comprehensive user security training, by merging human discernment with AI’s operational efficiency. However, the division of responsibilities must be explicit to prevent role confusion. The collaborative dynamic between humans and AI, akin to integrating new team members, necessitates mutual trust and understanding of each party’s function, underscoring the belief that autonomous systems represent the future of task execution [69, 70].

In addition, the findings identified the HIC paradigm as crucial in situations where the need for accountability is non-negotiable, such as in incident response and compliance management. The HIC paradigm ensures that humans retain authority over critical decisions. However, this dependency on human judgment might hinder the timeliness of responses and possibly constrain the exploitation of AI's capabilities to their fullest extent. This issue mirrors the challenges of scalability and delayed response times previously outlined in the HITL paradigm. Despite AI's remarkable abilities in speed, data processing, and pattern recognition [29, 30], its effectiveness is somewhat curtailed by its integration with human decision processes in the HITL and HIC paradigms.

Moreover, Coactive Systems has emerged as a promising, albeit complex, paradigm that facilitates real-time, dynamic collaboration between humans and AI. This paradigm is especially well-suited for developing adaptive and sophisticated defence mechanisms. Nevertheless, its successful application requires integrating advanced human insights and AI capabilities seamlessly. Echoing the sentiments regarding the HATL model, establishing clear roles and fostering trust between humans and AI systems is critical [69, 70].

Reflecting on these insights, the study concludes that the key to effective human-AI collaboration in cybersecurity is independent of the exclusive use of any single HMI paradigm. Instead, a strategic blend of paradigms, specifically chosen to meet the distinct demands of various cybersecurity tasks, is vital for harnessing the full potential of human-AI partnerships. This insight lays the groundwork for the primary objective of this paper, which is to conceptualise and develop a *c*AIF that integrates different HMI paradigms to bolster cybersecurity efforts across a range of domains. However, it is essential to note, as supported by the literature [69, 84], that adopting such an intelligent human–machine system will likely introduce new challenges and complexities. Addressing these emerging issues will necessitate a comprehensive and interdisciplinary approach spanning computer science, engineering, social sciences, neurology, and humanities.

5.2 Implications of the Framework's Core Components

While previous HAIT frameworks, such as those proposed by Sarker et al. [44], Chhetri et al. [55], Gonzalez et al. [66], Tsamados et al. [100], and Vats et al. [65], among others, have successfully enhanced specific aspects of cybersecurity operations, these models focus on individual paradigms or applications. The *c*AIF expands upon these foundations by integrating multiple HMI paradigms into a dynamic, adaptable structure that can respond to real-

5.2.1 Task-Specific Modules

The Task-Specific Modules, a cornerstone of the framework, are designed to cater to the nuanced demands of various cybersecurity operations. The following are just examples used for demonstration in this paper, as it has already been stated that the *c*AIF is contingent upon the *specific requirements* of individual tasks, which, based on the NIST CSF 2.07, could be hundreds of these:

- *Threat intelligence and monitoring*: This module would leverage the HOOTL and HOTL paradigms for the swift detection and evaluation of potential threats. The quick identification ensures that threats are recognised and assessed in real-time, providing a pivotal advantage in cybersecurity defence.
- *Incident response and management*: When managing and responding to incidents, we could rely on the HITL paradigm or Coactive Systems. This approach places human judgment at the forefront of critical decision-making processes [67, 97], while AI offers substantial support in data analysis and automation. This blend ensures that responses are both swift and informed, balancing the speed of automation with the nuance of human insight.
- *Risk and compliance management*: For tasks related to regulatory compliance and risk assessments, the framework emphasises using HOTL or HITL paradigms. This ensures comprehensive human oversight, which is critical in navigating the complex regulatory requirements and risk evaluation landscape.

Each of these modules operates within the framework dictated by the Paradigm Allocation Engine (see Sect. 5.2.3). This sophisticated component assigns the most suitable paradigm to each cybersecurity task, ensuring optimal efficacy. In essence, the *c*AIF system is not a one-size-fits-all solution. However, it is instead highly adaptable and tailored to meet the specific needs of different cybersecurity tasks through its Task-Specific Modules. An AI agent could be configured for each specific cybersecurity task.

5.2.2 Decision-Making Matrix

The Decision-Making Matrix is the pivotal element within *c*AIF, serving as the strategic backbone for selecting suitable HMI paradigms following the nature of cybersecurity tasks. This matrix judiciously evaluates tasks through a multifaceted lens, focusing on three essential dimensions to ascertain the most fitting approach:

- *Task complexity*: This dimension delves into the intricacy of tasks, gauging factors such as the requisite level of human intervention, the probability of encountering unexpected variables, and the demand for sophisticated decision-making processes [27, 82]. For tasks characterised by high complexity, paradigms that emphasise human judgment, such as HITL or Coactive Systems, are preferred to ensure that humans' nuanced understanding and critical thinking capabilities are effectively leveraged.
- *Risk level*: This aspect appraises the potential implications of tasks on the security posture of organisations, considering both the likelihood and the severity of adverse outcomes should the task be improperly executed. Tasks deemed high-risk, notably those involving incident response or threat mitigation, typically necessitate paradigms that guarantee human oversight, such as HOTL [73] or direct human intervention (HITL) [75], to mitigate risks effectively.
- *Urgency*: The urgency dimension assesses the time sensitivity associated with tasks. For instances requiring immediate action, such as automated threat detection or real-time incident response, paradigms that allow for autonomous AI execution under the HOOTL frame-

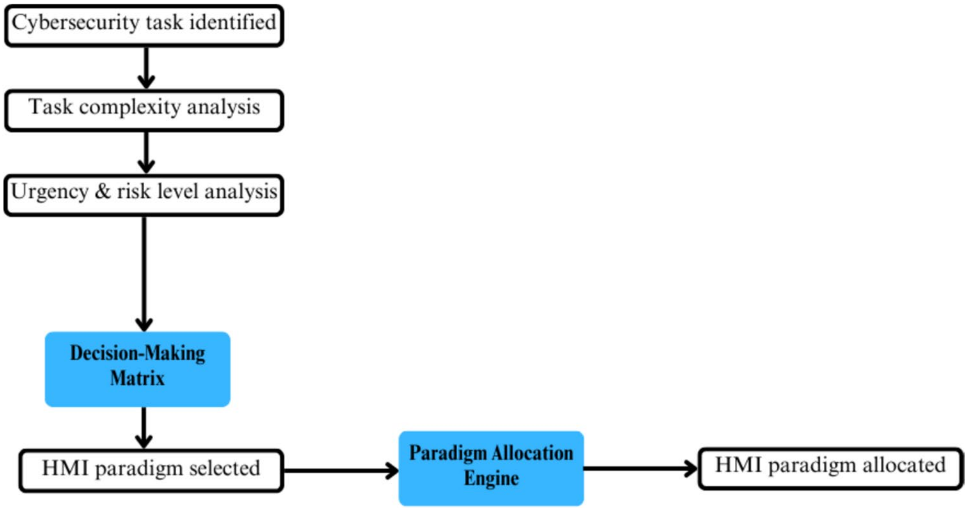


Fig. 7 Decision-making process in the *c*AIF

work might be most appropriate [82]. Conversely, tasks with lesser immediacy may benefit from a collaborative model, fostering a synergistic human-AI interaction HATL, that capitalises on the strengths of both entities [71].

Figure 7 shows the workflow diagram for decision-making in the *c*AIF. The decision-making process starts with identifying a cybersecurity task, then assessing the complexity, urgency, and risk level. Based on these factors, the Decision-Making Matrix selects the most suitable HMI paradigm, and the Paradigm Allocation Engine allocates the paradigm to the task.

By dynamically adapting to the evolving conditions of cybersecurity tasks, the Decision-Making Matrix ensures that the most suitable HMI paradigm is selected in real-time, optimising the interaction mode to address each situation's unique demands effectively. This adaptive approach underscores the *c*AIF's commitment to enhancing cybersecurity operations through a nuanced understanding of task-specific requirements, marrying the unparalleled capabilities of AI with the irreplaceable insight of human expertise.

5.2.3 Paradigm Allocation Engine

The Paradigm Allocation Engine, a sophisticated AI-driven component, plays a pivotal role in the cybersecurity framework by executing decisions outlined by the Decision-Making Matrix. This engine is pivotal for several reasons:

- *Real-time assessment*: It is dedicated to continuously surveilling cybersecurity operations, monitoring and evaluating tasks based on the established criteria of the Decision-Making Matrix. This ongoing scrutiny ensures that cybersecurity operations remain robust and adaptive to emerging threats and challenges, drawing on recent studies for its underpinning methodologies [12, 13].
- *Paradigm switching*: The engine's agility in facilitating smooth transitions between HMI paradigms is crucial. For example, a threat initially identified autonomously under the hands-off, tools-on paradigm (HOOTL), upon escalating in complexity, prompts the engine to seamlessly transition to a more collaborative hands-on, tools-light (HOTL) or hands-in, tools-light (HITL) paradigm. This mechanism ensures that human expertise is engaged when necessary, enhancing the system's responsiveness and adaptability.
- *Resource optimisation*: By judiciously allocating human and AI resources, the engine ensures that human intervention is reserved for the most impactful tasks. Conversely, routine or low-risk tasks are efficiently managed by AI systems [55]. This strategic distribution of tasks maximises resource utilisation and allows human experts to focus on areas where their expertise is indispensable.

Figure 8 shows the workflow diagram illustrating the *c*AIF's paradigm-switching process. It shows how the decision-making process flows from task assessment through the Decision-Making Matrix to the Paradigm Allocation Engine, which then allocates the appropriate HMI paradigm based on the needs of the cybersecurity task.

The Paradigm Allocation Engine is an intelligent part of the *c*AIF system, embodying the principles of efficiency, adaptability, and strategic resource management. Its role in real-time assessment, paradigm switching, and resource optimisation underscores its critical contribution to the dynamic and complex field of cybersecurity operations. This also highlights that the *c*AIF system is structured in a layered approach, each layer representing a different HMI paradigm. These layers interact, allowing the system to transition between paradigms seamlessly.

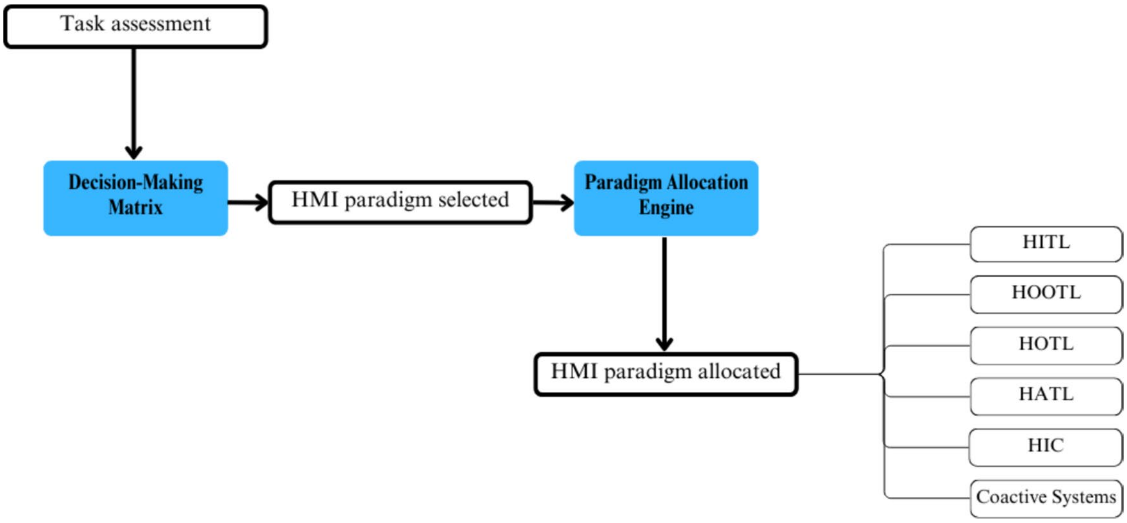


Fig. 8 HMI paradigm switching process

5.2.4 Interoperability Framework

Developing an Interoperability Framework within the *c*AIF plays a crucial role in ensuring seamless integration with existing cybersecurity infrastructures, thereby enhancing the effectiveness and adaptability of HAIT in cybersecurity contexts. This framework is designed to include several key components, including:

- *Application Programming Interface (API) integration*: At the heart of the Interoperability Framework is the provision of a standardised set of APIs. These APIs facilitate robust communication [101] between the *c*AIF and other cybersecurity tools and platforms. This level of integration is pivotal for enabling efficient data exchange and automating processes across different cybersecurity ecosystems, thereby fostering a more cohesive and responsive security infrastructure.
- *Modular design*: Recognising organisations' diverse needs and capabilities, the *c*AIF is structured to offer flexibility through its modular design. This approach allows organisations to adopt and implement specific components or paradigms of the framework that directly address their immediate cybersecurity needs. Over time, as an organisation's capabilities and requirements evolve, additional modules can be seamlessly integrated, ensuring that the cybersecurity strategy remains dynamic and responsive to changing operational landscapes.
- *Scalability*: A fundamental aspect of the Interoperability Framework is its scalability, ensuring that the *c*AIF can grow with the organisation. Whether accommodating an increasing volume of data or more complex cybersecurity challenges, the framework's scalable nature ensures that the *c*AIF remains a robust and viable solution over the long term.

Figure 9 shows the workflow diagram illustrating the *c*AIF's interoperability process.

By weaving together these core components—API integration, modular design, and scalability—the Interoperability Framework ensures that the *c*AIF can be effectively and efficiently integrated into various cybersecurity infrastructures. This adaptability is critical for optimising HAIT, creating a strategic balance between human expertise and AI autonomy. Ultimately, incorporating such a framework enhances the overall effectiveness and resilience of cybersecurity operations, positioning organisations to better navigate the complexities of the digital landscape.

5.2.5 Feedback and Learning System

The Feedback and Learning System represents another critical aspect in the ongoing development and refinement of the *c*AIF, ensuring its continuous evolution and alignment with the dynamic field of cybersecurity. This system comprises several integral components, each contributing to enhancing the *c*AIF's effectiveness and efficiency.

- *Performance metrics*: Central to this system is the rigorous tracking of various performance indicators that gauge the effectiveness of different HMI paradigms in managing specific tasks. Key metrics, such as response time, accuracy, and the efficiency of human-AI collaboration, are monitored. This approach provides a quantitative basis for evaluating the *c*AIF's performance and highlights areas for potential improvement. The significance of such metrics is underscored by research from Agye-

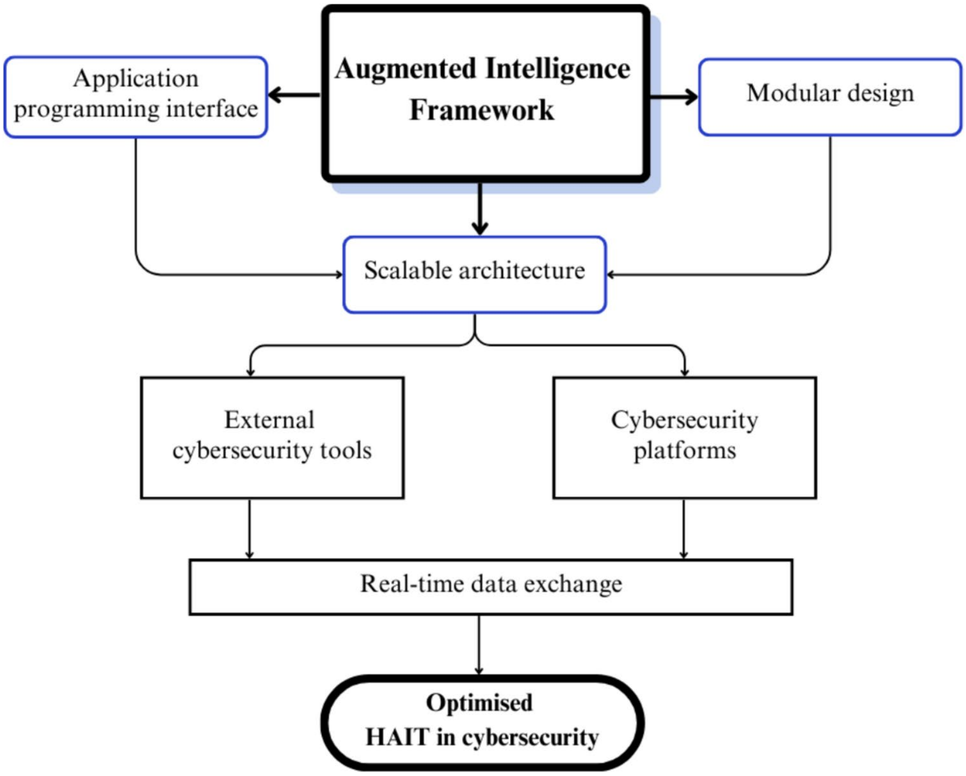

Augmented Intelligence Framework
Application programming interface
Modular design
Scalable architecture
External cybersecurity tools
Cybersecurity platforms
Real-time data exchange
Optimised HAIT in cybersecurity

Fig. 9 *c*AIF's interoperability workflow

pong et al. [15] and Jarrahi et al. [29], who emphasise the critical role of performance measurement in optimising human-AI interaction.

- *Continuous learning*: Leveraging AI-driven analytics, this component focuses on extracting valuable insights from past operations. This learning process enables the *c*AIF to refine its decision-making criteria and enhance its paradigm allocation strategies. By continuously adapting to new data and evolving threat landscapes, the *c*AIF ensures its relevance and effectiveness in the face of changing cybersecurity challenges.
- *User feedback*: Recognising the invaluable perspective of human operators, the Feedback and Learning System actively incorporates user feedback into the refinement of the *c*AIF. This process addresses critical aspects such as trust, usability, and operational efficiency, ensuring the interface between humans and AI/ML systems is as intuitive and effective as possible. The importance of integrating user feedback is further supported by findings from Engstrom and Ho [71] and policy recommendations from the European Commission5, which highlights the need for user-centric design in AI systems.

By synthesising these components—performance metrics, continuous learning, and user feedback—the Feedback and Learning System plays a pivotal role in the *c*AIF's ability to optimise HAIT in cybersecurity. This integrated approach enhances the *c*AIF's operational capabilities. It ensures its adaptability and resilience, positioning it as an intelligent human–machine system tailored to meet the complex demands of modern cybersecurity operations.

5.3 Intelligent Human–Machine System

Based on the insights derived from the core components of the framework, the operation of the *c*AIF can be summarised in the following manner:

- *Task-Specific Module*: This component customises AI agents to effectively address distinct cybersecurity tasks or operations, ensuring that the technology meets specific needs.
- *Decision-Making Matrix*: This element evaluates each task's complexity, associated risks, and urgency to select the most suitable HMI paradigm(s) for execution.
- *Paradigm Allocation Engine*: This mechanism facilitates the implementation of the selected paradigms by continually performing real-time assessments of tasks, enabling paradigm switching and optimising resource allocation as needed.
- *Interoperability Framework*: The *c*AIF is designed to integrate with existing cybersecurity infrastructures seamlessly. This is achieved using APIs, a modular design, and an emphasis on scalability, allowing for effective collaboration across systems.

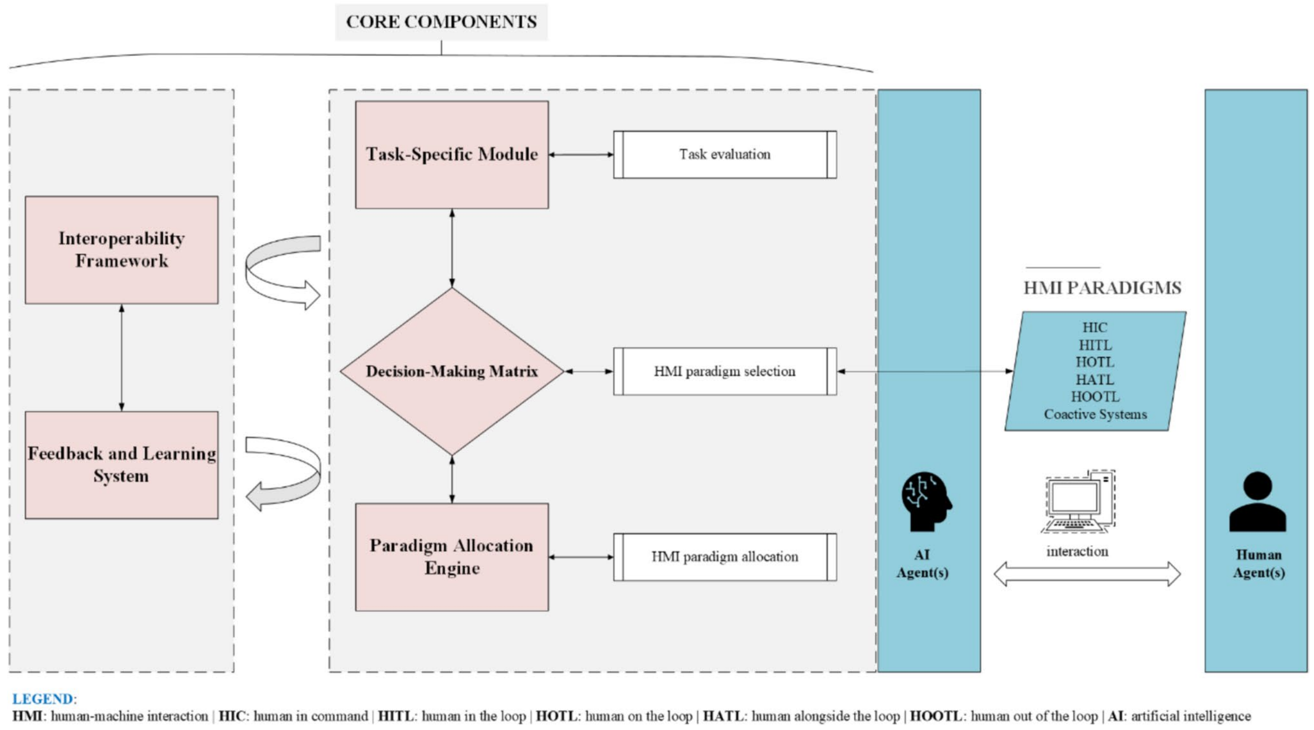

CORE COMPONENTS
Interoperability Framework
Feedback and Learning System
Task-Specific Module
Task evaluation
Decision-Making Matrix
HMI paradigm selection
Paradigm Allocation Engine
HMI paradigm allocation
HMI PARADIGMS
HIC
HITL
HOTL
HATL
HOOTL
Coactive Systems
AI Agent(s)
interaction
Human Agent(s)
LEGEND:
HMI: human-machine interaction | HIC: human in command | HITL: human in the loop | HOTL: human on the loop | HATL: human alongside the loop | HOOTL: human out of the loop | AI: artificial intelligence

Fig. 10 Augmented intelligence framework

- *Feedback and Learning System*: This component emphasises the importance of continuous improvement. By utilising performance metrics, fostering ongoing learning, and incorporating user feedback, the *c*AIF evolves iteratively to enhance its effectiveness.

A visual representation of the *c*AIF can be found in Fig. 10, illustrating how these components interact to create a cohesive and adaptive framework for cybersecurity operations.

Figure 10 illustrates the layered architecture of the *c*AIF, with each layer representing a specific core component. These layers are designed to be interconnected, allowing smooth transitions between different interaction paradigms. The Paradigm Allocation Engine facilitates these transitions and works with the other core components. For example, during an incident response scenario, the system might initially function in a fully autonomous mode (HOOTL) for rapid threat detection. Once a threat is confirmed, the *c*AIF can seamlessly transition to a mode that incorporates human oversight (HOTL) or direct human intervention (HITL), effectively merging human expertise with AI capabilities (HATL; Coactive Systems) to manage the situation more effectively.

The interconnectedness of the HMI paradigm layers is depicted by the arrows in Fig. 10, which show that all six HMI paradigms utilise the five core components to optimise HAIT in cybersecurity. The HMI paradigms can adjust their positions along the core components' vertical axis, adapting to specific needs. These arrows also represent the flow of decision-making and information processing across the core components. The upward and downward movement of the arrows signifies how output or decisions from one layer can inform or influence the processes in the layers above and below, reflecting a bi-directional flow of information. For instance, the bi-directional connection between the Task-Specific Modules and the Feedback and Learning System layers highlights the continuous feedback loop between human operators and AI systems. As humans engage with the system, their inputs and decisions are analysed by AI engines, allowing for algorithm refinement and enhanced future performance. This iterative learning process ensures that the framework remains effective even as the cybersecurity landscape evolves. The same bi-directional relationship applies between the Decision-Making Matrix, the Task-Specific Modules, and the other core components. Consequently, the core components of the *c*AIF are intricately interconnected with the HMI paradigms, with each component deriving its functionality from the paradigms it supports. Together, they create an intelligent human–machine system that dynamically adapts to the requirements of cybersecurity tasks, ensuring effective collaboration between human operators and AI agents.

As the literature states, AI systems undoubtedly offer tremendous processing speed and pattern recognition advantages, but their susceptibility to biases poses significant challenges. In complex cybersecurity environments, biases in decision-making models could lead to inappropriate responses to novel or unexpected threats. This underscores the importance of integrating explainability mechanisms within the *c*AIF to mitigate these risks and ensure human operators retain confidence in the system's outputs.

5.4 Empirical Validation of the *c*AIF

Empirical validation is a critical research process involving testing models, conceptual frameworks or theories against real-world data to ensure their accuracy and applicability [102]. It is the assessment of a model's answer by its correspondence to relevant empirical data [103]. For a framework like the *c*AIF, which integrates diverse HMI paradigms, empirical validation is essential to demonstrate its effectiveness in enhancing cybersecurity operations. Potential *c*AIF validation methods include case studies, simulations, and pilot implementations in real-world settings.

5.4.1 Case Studies

Case study methodology, as described by Yin [104], emphasises the in-depth exploration of complex phenomena in specific contexts, making it well-suited for empirically testing the *c*AIF. Although case studies could involve partnerships with organisations of any size in many sectors, such as healthcare, finance, and critical infrastructure, the subsequent study will include partnering with one or two small-to-medium enterprises (SMEs) in South Africa. This will involve applying the *c*AIF to improve the detection and remediation of cyber threats targeting SMEs. The key steps for validating the *c*AIF would include (i) Gathering qualitative and quantitative data on SMEs' system performances, such as incident detection accuracy and time-to-response metrics, before and after implementing the *c*AIF; and (ii) Measuring improvements in cybersecurity operational efficiency, decision-making quality, and overall cybersecurity resilience of SMEs.

5.4.2 Simulations

Controlled simulation environments could offer a safe and repeatable platform to test the *c*AIF's effectiveness against predefined cybersecurity threats. Simulation platforms provide a means for examining complex interactions and changes within a system over time, including the influence of social actors [105]. Cyber ranges, such as the MITRE ATT&CK[8] or similar platforms can be employed to replicate sophisticated cyberattacks. The key steps for validating the *c*AIF would include (i) Designing scenarios involving common cybersecurity challenges like ransomware attacks or insider threats and then deploying the *c*AIF's components to assess their real-time performance; and (ii) Measuring key indicators such as the speed of human-AI collaboration, the accuracy of threat identification, and system adaptability.

5.4.3 Pilot Implementations

Pilot studies test the feasibility of methods and procedures for larger-scale studies [106]. Pilot programs could enable limited deployment of the *c*AIF in real-world organisational settings. These programs can be executed in phases, allowing gradual integration and testing. For instance, in phase one, a specific cybersecurity operation, such as incident response management, could be selected, and then the *c*AIF could be deployed to evaluate its decision-making capabilities. The scope could be expanded in phase two to include multi-layered defence strategies, incorporating human operators' and supervisors' feedback. The *c*AIF's effectiveness and usability could be gauged using mixed-methods approaches combining surveys, interviews, and system performance analytics.

5.4.4 Challenges and Considerations

Several challenges may arise while empirically validating the *c*AIF. Resource constraints are one of the biggest challenges as organisations, including the author's institution, may lack the technical infrastructure or skilled personnel to implement and test the framework effectively, as advanced AI/ML tools must be built. This is one of the reasons why empirically validating the *c*AIF is not within the scope of this paper. Furthermore, testing the *c*AIF in sensitive domains like healthcare or finance necessitates robust mechanisms to ensure compliance with data protection regulations, such as those in South Africa1 and the EU2. Organisations at different stages of cybersecurity readiness may experience disparate results, requiring tailored adaptation of the *c*AIF, which would be challenging due to resource constraints. Moreover, achieving cross-sector collaborations, which could provide richer datasets for validating the framework's generalisability, would also prove challenging. Strategies such as phased implementations, extensive training programs, and partnerships with academic or government institutions can be adopted to address some of these challenges. At the time of writing, partnerships with some of South Africa's government institutions for infrastructure and

[8] MITRE. ATT&CK matrix for enterprise. 2015. https://attack.mitre.org/. Accessed 13 Jan 2025.

research funding are being explored to embark on a project to validate the *c*AIF empirically.

6 Conclusion

6.1 Concluding Remarks

The first research objective aimed to identify and examine various HMI paradigms, assessing their strengths, limitations, and suitability for enhancing human-AI teaming in modern cybersecurity operations. Six specific HMI paradigms were analysed: Human-in-the-loop, human-out-of-the-loop, human-on-the-loop, human-alongside-the-loop, human-in-command, and Coactive Systems. The findings reveal that each HMI paradigm offers unique strengths and limitations, making them better suited for particular tasks within the cybersecurity framework. For instance, the HITL paradigm is particularly effective in complex decision-making scenarios that require direct human intervention. Conversely, the HOOTL paradigm excels in environments where automated responses can be carried out without human involvement. The HOTL paradigm balances autonomy and human oversight, allowing adaptive decision-making while maintaining human awareness. In situations requiring significant accountability, such as risk management, regulatory compliance, and incident response, the HIC paradigm ensures that a human retains control over critical decisions. The HATL paradigm allows human and AI agents to operate simultaneously on different tasks without interfering with one another. At the same time, Coactive Systems fosters a collaborative environment where humans and AI influence each other's actions in real-time.

However, the effectiveness of these paradigms can vary greatly depending on situational factors. This underscores the importance of adopting a flexible approach to optimise human-AI collaboration amid the ever-changing landscape of cyber threats. Organisations can leverage human expertise and AI capabilities by strategically combining these HMI paradigms to enhance their dynamic cybersecurity operations. This leads to the second research objective, which focuses on developing a hybrid HAIT framework tailored to optimise cybersecurity operations across different domains, grounded in the selected HMI paradigms' strengths, limitations, and suitability. The result is an Augmented Intelligence Framework consisting of five core architectural components: Task-Specific Modules, a Decision-Making Matrix, a Paradigm Allocation Engine, an Interoperability Framework, and a Feedback and Learning System. These components are intricately linked to the HMI paradigms, ensuring the *c*AIF remains adaptable to various cybersecurity challenges. This dynamic architecture facilitates flexible human-AI collaboration, effectively harnessing human insight and AI capabilities to strengthen resilience, efficiency, and adaptability in an ever-evolving cyber threat environment.

6.2 Recommendations

Organisations looking to enhance their AI-driven cybersecurity operations can leverage the *c*AIF by integrating various HMI paradigms tailored to specific tasks. This multifaceted approach enhances proactive measures and strengthens reactive capabilities in the face of cyber threats. However, practical considerations are essential for effectively adopting the *c*AIF in cybersecurity operations. Key factors include the skill levels of human operators in cybersecurity and AI/ML technologies, implementation methodologies, budget constraints for capital and operational expenditures, and adherence to regulatory compliance. In summation, this paper offers the following recommendations for the successful adoption and implementation of the *c*AIF:

- *Adopt a phased implementation approach*: Given the complexity of incorporating multiple HMI paradigms, a phased strategy is advisable. By starting with the most critical components and gradually expanding from there, organisations can mitigate risks and make iterative improvements based on real-world feedback.
- *Invest in advanced AI capabilities*: To maximise the potential of the *c*AIF, organisations should invest in advanced AI/ML technologies. Tools that promote real-time decision-making, adaptive learning, and context-aware operations will significantly improve the model's responsiveness to evolving cybersecurity threats and enhance overall operational efficiency.
- *Enhance user training and collaboration*: Solid training for cybersecurity professionals is crucial for successfully deploying the *c*AIF. Collaboration between (multi)skilled human operators and AI systems will enable trust and ensure the framework is utilised effectively.
- *Ensure compliance and ethical AI use*: As the *c*AIF becomes increasingly autonomous, aligning its operations with regulatory requirements and ethical standards is vital. Regular audits and the incorporation of XAI tools can help maintain transparency, accountability, and trust in the decision-making processes within the framework.

As with any research endeavour, there are limitations to how the study was conducted, influencing the overall findings. There are also practical aspects to be considered when validating systems like the *c*AIF. The following section will address some of these limitations and propose potential avenues for future research. By exploring these areas, we can refine the capabilities of the *c*AIF, making it an even more helpful tool for cybersecurity and beyond.

6.3 Limitations and Future Research

Integrating multiple HMI paradigms within the *c*AIF presents various design challenges. Harmonising different approaches, such as HITL, HOOTL, and Coactive Systems, into a cohesive framework can introduce complexities during integration. Furthermore, while the theoretical foundation of the framework provides a strong base, it may not fully account for the practical challenges encountered in real-world cybersecurity operations. Therefore, thorough empirical testing and validation are crucial for refining the framework to ensure effective implementation across various organisational contexts. Future research should thus prioritise the empirical validation of the *c*AIF in diverse cybersecurity domains, including healthcare and finance. Furthermore, as AI systems become more autonomous in cybersecurity operations, ethical concerns about accountability, transparency, and privacy become paramount. The reliance on AI for critical decision-making raises questions about who is responsible for failure or misjudgment. Future research should also explore frameworks governing these ethical concerns, ensuring that human oversight remains central, especially in high-risk contexts such as healthcare or finance. Enhancing the *c*AIF's contextual understanding through advanced AI/ML tools will be essential for its adaptability and resilience. This focus will help maintain the *c*AIF's alignment with its intended objectives by fostering research into its XAI mechanisms. Such research will enhance AI agents' observability, predictability, and directability as they function alongside human counterparts. Several areas for further exploration are therefore recommended:

- *Development of advanced AI/ML tools:*
  - o Contextual AI for dynamic paradigm switching: Future studies could focus on developing AI/ML tools that dynamically enable the framework to switch between HMI paradigms based on real-time context and situational awareness. This will require creating AI systems capable of understanding the nuances of various cybersecurity tasks and anticipating when a shift is needed. Techniques such as reinforcement learning and context-aware algorithms could significantly improve the framework's decision-making processes.
  - o Integration of XAI: Another vital area for research is incorporating XAI tools into the framework. These tools can provide transparency in AI decision-making, helping human operators understand and trust the actions of AI systems. This is especially crucial in paradigms like HITL and HOTL, where human oversight is essential.

- *Application in other domains:*
  - o Healthcare cybersecurity: Given the sensitive nature of healthcare data and the rising number of cyberattacks on healthcare systems, the *c*AIF could be adapted and tested within this sector. Future research could investigate how the framework's dynamic paradigm integration can enhance the protection of patient data, ensure compliance with healthcare regulations, and improve incident response in medical settings.
  - o Financial services: With its complex regulatory landscape and high-stakes cybersecurity requirements, the financial sector represents another arena for applying the *c*AIF. Research could explore how the framework can be customised to manage financial data integrity, fraud detection, and regulatory compliance across various jurisdictions. This would involve tailoring the Decision-Making Matrix core component to meet the specific needs of financial institutions.

- *Cross-domain adaptability:*
  - o Generalised framework development: Future work should aim to create a generalised version of the *c*AIF that can be readily adapted to various fields beyond cybersecurity. This requires identifying core components and paradigms that are universally applicable and designing a flexible architecture that allows domain-specific customisation.
  - o Interdisciplinary research: Collaboration among cybersecurity experts, AI researchers, and specialists in fields like healthcare, finance, and critical infrastructure could foster the development of interdisciplinary frameworks that incorporate the strengths of the *c*AIF while addressing the unique challenges of each domain.

- *Human factors and user experience:*
  - o User-centric design: Additional research could focus on optimising the user experience within the *c*AIF, particularly its interfaces and interactions with human operators. Studies could explore designing more intuitive interfaces that facilitate seamless collaboration between human and AI agents, thereby reducing cognitive load and improving overall efficiency.

- o Trust and collaboration studies: Investigating factors influencing trust in AI systems within the *c*AIF could lead to better collaborative outcomes. Understanding how different paradigms affect trust levels will be crucial in developing strategies to enhance human-AI collaboration, especially in high-stakes environments.

- *Ethical and legal considerations:*

  - o Ethical AI deployment: As AI systems gain autonomy, considerations around ethical decision-making, privacy, and accountability become increasingly important. Future research should address the ethical implications of deploying the *c*AIF in cybersecurity, particularly in achieving a balance between AI autonomy and human oversight.
  - o Legal frameworks: Research could also focus on establishing legal frameworks that support the responsible deployment of the *c*AIF. This includes ensuring cybersecurity and data privacy regulations compliance and exploring how legal standards may need to evolve to accommodate advanced AI-driven systems.

Pursuing these areas of research will ensure that the *c*AIF becomes an invaluable tool for human-AI collaboration, equipped to adapt to emerging challenges and contribute to the overarching goal of securing digital environments across diverse industries.

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.

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

Funding No funding was received to conduct this study.

Data availability No datasets were generated or analysed during the current study. The only textual data underlying the study are in Table 2, as generated from the literature review.

Declarations

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

Ethical approval and consent to participate Not applicable.

Consent for publication The author declares consent for publication.

References

1. Camacho NG. The role of AI in cybersecurity: addressing threats in the digital age. J Artif Intell Gen Sci. 2024;3:143–54. https://doi.org/10.60087/jaigs.v3i1.75.
2. Yamin MM, Ullah M, Ullah H, Katt B. Weaponized AI for cyber attacks. J Inf Secur Appl. 2021;57: 102722. https://doi.org/10.1016/j.jisa.2020.102722.
3. Samtani S, Kantarcioglu M, Chen H. Trailblazing the artificial intelligence for cybersecurity discipline: a multi-disciplinary research roadmap. ACM Trans Manag Inf Syst. 2020;11:1–19. https://doi.org/10.1145/3430360.
4. Chen W, Zhang J. Elevating security operations: the role of AI-driven automation in enhancing SOC efficiency and efficacy. J Artif Intell Mach Learn Manag. 2024;8:1–13.
5. Dubey A, Abhinav K, Jain S, Arora V, Puttaveerana A. HACO: a framework for developing human-AI teaming. In: Proceedings of the 2020 Innovations in Software Engineering Conference. Association for Computing Machinery, New York; 2020. pp. 1–9.
6. Pflanzer M, Traylor Z, Lyons JB, Dubljević V, Nam CS. Ethics in human–AI teaming: principles and perspectives. AI Ethics. 2023;3:917–35. https://doi.org/10.1007/s43681-022-00214-z.
7. Gomez SR, Mancuso V, Staheli D. Considerations for human-machine teaming in cybersecurity. In: Schmorrow DD, Fidopiastis CM, editors. Augmented cognition. Cham: Springer; 2019. p. 153–68.
8. Xiang C-G, Yu Z. Human-machine hybrid augmented intelligence: human-machine relationship, collaboration and mutual enhancement. In: Proceedings of the 2023 China automation congress. IEEE Xplore, New York; 2024 pp. 7471–7478.
9. Khalil MI, Abdel-Rahman M. Advanced cybersecurity measures in IT service operations and their crucial role in safeguarding enterprise data in a connected world. Eig Rev Sci Technol. 2023;7:138–58.
10. Ferdous J, Islam R, Mahboubi A, Islam MdZ. A review of state-of-the-art malware attack trends and defense mechanisms. IEEE Access. 2023;11:121118–41. https://doi.org/10.1109/ACCESS.2023.3328351.
11. Scalise P, Boeding M, Hempel M, Sharif H, Delloiacovo J, Reed J. A systematic survey on 5G and 6G security considerations, challenges, trends, and research areas. Future Internet. 2024;16:67. https://doi.org/10.3390/fi16030067.
12. Villalón-Huerta A, Ripoll-Ripoll I, Marco-Gisbert H. Key Requirements for the detection and sharing of behavioral indicators of compromise. Electronics. 2022;11:416. https://doi.org/10.3390/electronics11030416.
13. Wallis T, Leszczyna R. EE-ISAC—practical cybersecurity solution for the energy sector. Energies. 2022;15:2170. https://doi.org/10.3390/en15062170.
14. Schlette D, Caselli M, Pernul G. A comparative study on cyber threat intelligence: the security incident response perspective. IEEE Commun Surv Tutor. 2021;23:2525–56. https://doi.org/10.1109/COMST.2021.3117338.

15. Agyepong E, Cherdantseva Y, Reinecke P, Burnap P. Challenges and performance metrics for security operations center analysts: a systematic review. J Cyber Secur Technol. 2020;4:125–52. https://doi.org/10.1080/23742917.2019.1698178.
16. Bartwal U, Mukhopadhyay S, Negi R, Shukla S. Security orchestration, automation, and response engine for deployment of behavioural honeypots. In: Proceedings of the 2022 Dependable and Secure Computing Conference. IEEE Xplore, New York; 2022. pp. 1–8.
17. George AS, Baskar T, Srikaanth PB. Cyber threats to critical infrastructure: Assessing vulnerabilities across key sectors. Partn Univers Int Innov J. 2024;2:51–75. https://doi.org/10.5281/zenodo.10639463.
18. Lincke S. Complying with the PCI DSS Standard. In: Lincke S, editor. Information security planning: a practical approach. Cham: Springer; 2024. p. 45–63.
19. Alahmari S, Renaud K, Omoronyia I. Moving beyond cyber security awareness and training to engendering security knowledge sharing. Inf Syst E-Bus Manag. 2023;21:123–58. https://doi.org/10.1007/s10257-022-00575-2.
20. Djenna A, Harous S, Saidouni DE. Internet of things meet internet of threats: new concern cyber security issues of critical cyber infrastructure. Appl Sci. 2021;11:4580. https://doi.org/10.3390/app11104580.
21. Sen R, Heim G, Zhu Q. Artificial intelligence and machine learning in cybersecurity: applications, challenges, and opportunities for MIS academics. Commun Assoc Inf Syst. 2022;51:179–209. https://doi.org/10.17705/1CAIS.05109.
22. Abdullahi M, Baashar Y, Alhussian H, Alwadain A, Aziz N, Capretz LF, Abdulkadir SJ. Detecting cybersecurity attacks in Internet of Things using artificial intelligence methods: a systematic literature review. Electronics. 2022;11:198. https://doi.org/10.3390/electronics11020198.
23. Aslan Ö, Aktuğ SS, Ozkan-Okay M, Yilmaz AA, Akin E. A comprehensive review of cyber security vulnerabilities, threats, attacks, and solutions. Electronics. 2023;12:1333. https://doi.org/10.3390/electronics12061333.
24. Srinivas J, Das AS, Kumar N. Government regulations in cyber security: framework, standards and recommendations. Future Gener Comput Syst. 2019;92:178–88. https://doi.org/10.1016/j.future.2018.09.063.
25. Wallis T, Johnson C, Khamis M. Interorganizational cooperation in supply chain cybersecurity: a cross-industry study of the effectiveness of the UK implementation of the NIS directive. Inf Secur Int J. 2021;48:36–68. https://doi.org/10.11610/isij.4812.
26. Hammi B, Zeadally S, Nebhen J. Security threats, countermeasures, and challenges of digital supply chains. ACM Comput Surv. 2023;55(316):1–316. https://doi.org/10.1145/3588999.
27. Mern J, Hatch K, Silva R, Hickert C, Sookoor T, Kochenderfer MJ. Autonomous attack mitigation for industrial control systems. In: Proceedings of the 2022 Dependable Systems and Networks Workshops. IEEE Xplore, New York; 2022. pp. 28–36. https://doi.org/10.1109/DSN-W54100.2022.00015.
28. Kaur R, Gabrijelčič D, Klobučar T. Artificial intelligence for cybersecurity: literature review and future research directions. Inf Fusion. 2023;97: 101804. https://doi.org/10.1016/j.inffus.2023.101804.
29. Jarrahi MH, Lutz C, Newlands G. Artificial intelligence, human intelligence and hybrid intelligence based on mutual augmentation. Big Data Soc. 2022;9:20539517221142824. https://doi.org/10.1177/20539517221142824.
30. Jarrahi MH. Artificial intelligence and the future of work: human-AI symbiosis in organizational decision making. Bus Horiz. 2018;61:577–86. https://doi.org/10.1016/j.bushor.2018.03.007.
31. Cu MK, Gamboa VL, J. J. Abraham JJ, Tan SM, Ong E. Humans + AI: exploring the collaboration between AI and human labor in the workplace. In: Proceedings of the 2023 HCI and UX Conference. IEEE Xplore, New York; 2024. pp. 35–40. https://doi.org/10.1109/CHIuXiD59550.2023.10452733.
32. Ezer N, Bruni S, Cai Y, Hepenstal SJ, Miller CA, Schmorrow DD. Trust engineering for human-AI teams. In: Proceedings of the human factors and ergonomics society annual meeting. Thousand Oaks: Sage Journals; 2019. pp. 63:322–326. https://doi.org/10.1177/1071181319631264.
33. Wang D, Churchill E, Maes P, Fan X, Shneiderman B, Shi Y, Wang Q. From human-human collaboration to human-AI collaboration: Designing AI systems that can work together with people. In: Proceedings of the 2020 human factors in computing systems conference. Association for Computing Machinery, New York; 2020. pp. 1–6. https://doi.org/10.1145/3334480.3381069.
34. Bansal G, Nushi B, Kamar E, Weld DS, Lasecki WS, Horvitz E. Updates in human-AI teams: understanding and addressing the performance/compatibility tradeoff. In: Proceedings of the AAAI conference on artificial intelligence. Burnaby: public knowledge project; 2019. pp. 33:2429–2437. https://doi.org/10.1609/aaai.v33i01.33012429.
35. Xu W, Gao Z. Applying HCAI in developing effective human-AI teaming: a perspective from human-AI joint cognitive systems. Interactions. 2024;31:32–7. https://doi.org/10.1145/3635116.
36. Malatji M, Marnewick A, Von Solms S. Validation of a socio-technical management process for optimising cybersecurity practices. Comput Secur. 2020;95: 101846. https://doi.org/10.1016/j.cose.2020.101846.
37. Xi Z, Chen W, Guo X, He W, Ding Y, Hong B, Zhang M, Wang W, Jin S, Zhou E, Zheng R, Fan X, Wang X, Xiong L, Zhou Y, Wang W, Jiang C, Zou Y, Liu X, Yin Z, Dou S, Weng R, Cheng W, Zhang Q, Qin W, Zheng Y, Qiu X, Huang X, Gui T. The rise and potential of large language model based agents: a survey. arXiv. 2023. https://doi.org/10.48550/arXiv.2309.07864. Accessed 15 May 2025.
38. Russell SJ, Norvig P. Artificial intelligence: a modern approach. 4th ed. Prentice Hall; 2020.
39. Lund BD, Wang T. Chatting about ChatGPT: how may AI and GPT impact academia and libraries? Libr Hi Tech News. 2023. https://doi.org/10.1108/LHTN-01-2023-0009.
40. Firlej M, Taeihagh A. Regulating human control over autonomous systems. Regul Gov. 2021;15:1071–91. https://doi.org/10.1111/rego.12344.
41. Zhang Q, Lu J, Jin Y. Artificial intelligence in recommender systems. Complex Intell Syst. 2021;7:439–57. https://doi.org/10.1007/s40747-020-00212-w.
42. Chakraborti T, Isahagian V, Khalaf R, Khazaeni Y, Muthusamy V, Rizk Y. From robotic process automation to intelligent process automation. In: Asatiani A, García JM, Helander N, Jiménez-Ramírez A, Koschmider A, Mendling J, Meroni G, Reijers HA, editors. Business process management: blockchain and robotic process automation forum. Cham: Springer; 2020. p. 215–28.
43. Maymí FJ, Thomson R. Human-machine teaming and cyberspace. In: Schmorrow DD, Fidopiastis CM, editors. Augmented cognition: intelligent technologie. Springer: Cham; 2018. p. 299–315.
44. Sarker IH, Janicke H, Mohammad N, Watters P, Nepal S. 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-H, Thomas JJ, Intan R, Arefin MS, editors. Intelligent computing and optimization: lecture notes in networks and systems, vol. 874. Cham: Springer; 2024. p. 140–9.
45. Wang Y. Human-centered design of AI-driven user interfaces for autonomous vehicle cybersecurity. J AI-Assist Sci Discov. 2022;2:1–24.
46. Hauptman AI, Schelble BG, McNeese NJ, Madathil KC. Adapt and overcome: perceptions of adaptive autonomous agents for

human-AI teaming. Comput Hum Behav. 2023;138: 107451. https://doi.org/10.1016/j.chb.2022.107451.
47. Xu X, Lu Y, Vogel-Heuser B, Wang L. Industry 4.0 and Industry 5.0—Inception, conception and perception. J Manuf Syst. 2021;61:530–5. https://doi.org/10.1016/j.jmsy.2021.10.006.
48. Leng J, Sha W, Wang B, Zheng P, Zhuang C, Liu Q, Wuest T, Mourtzis D, Wang L. Industry 5.0: prospect and retrospect. J Manuf Syst. 2022;65:279–95. https://doi.org/10.1016/j.jmsy.2022.09.017.
49. Hornbæk K, Oulasvirta A. What is interaction?. In: Proceedings of the 2017 Human Factors in Computing Systems Conference. Association for Computing Machinery, New York; 2017. pp. 5040–5052. https://doi.org/10.1145/3025453.3025765.
50. Mourtzis D, Angelopoulos J, Panopoulos N. The future of the human–machine interface (HMI) in society 5.0. Future Internet. 2023;15:162. https://doi.org/10.3390/fi15050162.
51. Al-Mansoori S, Salem MB. The role of artificial intelligence and machine learning in shaping the future of cybersecurity: trends, applications, and ethical considerations. Int. J. Soc. Anal. 2023;8. https://norislab.com/index.php/ijsa/article/view/36.
52. Aliberti L, D'Aniello G, Fortino G, Gaeta M. Situation projection for enhanced human-machine interaction based on rule mining. In: Proceedings of the 2024 international conference on human-machine systems. IEEE Xplore, New York; 2024. pp. 1–6. https://doi.org/10.1109/ICHMS59971.2024.10555614.
53. Belaïd A. Human-machine collaboration for incident response in cybersecurity operations for autonomous vehicles. Afr. J. Artif. Intell. Sustain. Dev. 2024; 4. https://africansciencegroup.com/index.php/AJAISD/article/view/98.
54. Chen L, Zhang W, Song Y, Chen J. Machine learning for human–machine systems with advanced persistent threats. IEEE Trans Hum Mach Syst. 2024;54:753–61. https://doi.org/10.1109/THMS.2024.3439625.
55. Chhetri MB, Tariq S, Singh S, Jalalvand F, Paris C, Nepal S. Towards human-AI teaming to mitigate alert fatigue in security operations centres. ACM Trans Internet Technol. 2024;12:1–12. https://doi.org/10.1145/3670009.
56. Chowdhury A, Nguyen H, Ashenden D, Pogrebna G. POSTER: a teacher-student with human feedback model for human-AI collaboration in cybersecurity. In: Proceedings of the 2023 Asia conference on computer and communications security. New York: Association for computing machinery; 2023. pp. 1040–1042. https://doi.org/10.1145/3579856.3592829.
57. Desai B, Patil K, Mehta I, Patil A. Explainable AI in cybersecurity: a comprehensive framework for enhancing transparency, trust, and human-AI collaboration. In: Proceedings of the 2024 international seminar on application for technology of information and communication. New York: IEEE Xplore; 2024. pp. 135–150. https://doi.org/10.1109/iSemantic63362.2024.10762690.
58. Gore S, Hamsa S, Roychowdhury S, Patil G, Gore S, Karmode S. Augmented intelligence in machine learning for cybersecurity: enhancing threat detection and human-machine collaboration. In: Proceedings of the 2023 international conference on augmented intelligence and sustainable systems. New York: IEEE Xplore; 2023. pp. 638–644. https://doi.org/10.1109/ICAISS58487.2023.10250514.
59. Hayes B, Moniz M. Trustworthy human-centered automation through explainable AI and high-fidelity simulation. In: Cassenti DN, Scataglini S, Rajulu SL, Wright JL, editors. Advances in simulation and digital human modelling. Cham: Springer; 2021. p. 3–9.
60. Karunamurthy A, Kiruthivasan AR, Gauthamkrishna S. Human-in-the-loop intelligence: advancing AI-centric cybersecurity for the future. Quing Int J Multidiscip Sci Res Dev. 2023. https://doi.org/10.54368/qijmsrd.2.3.0011.
61. Mikhalevich IF, Ryjov AP. Augmented intelligence framework for protecting against cyberattacks. In: Proceedings of the 2018 engineering and telecommunication conference. New York: IEEE Xplore; 2019. pp. 143–145. https://doi.org/10.1109/EnT-MIPT.2018.00039.
62. Olla R, Hand E, Louis SJ, Houmanfar R, Sengupta S. A cybersecurity game to probe human-AI teaming. In: Proceedings of the 2024 conference on games. New York: IEEE Xplore; 2024. pp. 1–5. https://doi.org/10.1109/CoG60054.2024.10645666.
63. Parlapalli V, Jayaram V, Aarella SG, Peddireddy K, Palle RR. Enhancing cybersecurity: a deep dive into augmented intelligence through machine learning and image processing. In: Proceedings of the 2023 international workshop on artificial intelligence and image processing. New York: IEEE Xplore; 2024. pp. 96–100. https://doi.org/10.1109/IWAIIP58158.2023.10462845.
64. Roch N, Sievers H, Schöni L, Zimmermann V. Navigating autonomy: Unveiling security experts' Perspectives on augmented intelligence in cybersecurity. In: Proceedings of the 2024 symposium on usable privacy and security. Berkeley: usenix association; 2024. pp. 41–60. https://www.usenix.org/conference/soups2024/presentation/roch.
65. 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 DN, Tang X, Ghosalkar R, Shen C, Shen R, Hussain N, Ravichandran K, Davis J. A survey on human-AI teaming with large pre-trained models. arXiv. 2024. https://doi.org/10.48550/arXiv.2403.04931. Accessed 15 May 2025.
66. Gonzalez C, Aggarwal P, Cranford EA, Lebiere C. Adaptive cyberdefense with deception: a human–AI cognitive approach. In: Bao T, Tambe M, Wang C, editors. Cyber deception: techniques, strategies, and human aspects. Cham: Springer; 2023. p. 41–57.
67. Anderson M, Fort K. Human where? A new scale defining human involvement in technology communities from an ethical standpoint. Int Rev Inf Ethics. 2022. https://doi.org/10.29173/irie477.
68. Aschenbrenner D, Colloseus C. Human in command in manufacturing. In: Alfnes E, Romsdal A, Strandhagen JO, von Cieminski G, Romero D, editors. Advances in production management systems Production management systems for responsible manufacturing, service, and logistics futures. Cham: Springer; 2023. p. 559–72.
69. Berberian B, Somon B, Sahaï A, Gouraud J. The out-of-the-loop brain: a neuroergonomic approach of the human automation interaction. Annu Rev Control. 2017;44:303–15. https://doi.org/10.1016/j.arcontrol.2017.09.010.
70. Campbell M. Tomorrow's applications require IT operations that are autonomous, ubiquitous, and smarter—In a word, invisible. Computer. 2023;56:129–33. https://doi.org/10.1109/MC.2022.3227656.
71. Engstrom DF, Ho DE. Algorithmic accountability in the administrative State. Yale J Regul. 2020;37:800–54.
72. Johnson M, Bradshaw JM, Feltovich PJ, Jonker CM, van Riemsdijk MB, Sierhuis M. Coactive design: designing support for interdependence in joint activity. J Hum-Robot Interact. 2014;3:43–69. https://doi.org/10.5898/JHRI.3.1.Johnson.
73. Li N, Adepu S, Kang E, Garlan D. Explanations for human-on-the-loop: a probabilistic model checking approach. In: Proceedings of the 2020 international symposium on software engineering for adaptive and self-managing systems. New York: Association for Computing Machinery; 2020. pp. 181–187. https://doi.org/10.1145/3387939.3391592.
74. Nahavandi S. Trusted autonomy between humans and robots: toward human-on-the-loop in robotics and autonomous systems.

IEEE Syst Man Cybern Mag. 2017;3:10–7. https://doi.org/10.1109/MSMC.2016.2623867.

75. Rahwan I. Society-in-the-loop: programming the algorithmic social contract. Ethics Inf Technol. 2018;20:5–14. https://doi.org/10.1007/s10676-017-9430-8.
76. Stein Z. Between the natural and the normative: the concept of development in integrative theories of human development. In: Mascolo MF, Bidell TR, editors. Handbook of integrative developmental science. Oxfordshire: Routledge; 2020. p. 38–62.
77. Wu X, Xiao L, Sun Y, Zhang J, Ma T, He L. A survey of human-in-the-loop for machine learning. Future Gener Comput Syst. 2022;135:364–81. https://doi.org/10.1016/j.future.2022.05.014.
78. Nise NS. Control systems engineering. Hoboken, NJ, USA: Wiley; 2020.
79. Sheridan TB. Human supervisory control of automation. In: Handbook of human factors and ergonomics 736–760. 5th ed. Hoboken, NJ USA: Wiley; 2021.
80. Meskó B, Topol EJ. The imperative for regulatory oversight of large language models (or generative AI) in healthcare. Npj Digit Med. 2023;6:1–6. https://doi.org/10.1038/s41746-023-00873-0.
81. Middleton SE, Letouzé E, Hossaini A, Chapman A. Trust, regulation, and human-in-the-loop AI: within the European region. Commun ACM. 2022;65:64–8. https://doi.org/10.1145/3511597.
82. Fanni R, Steinkogler VE, Zampedri G, Pierson J. Active human agency in artificial intelligence mediation. In: Proceedings of the 2020 EAI international conference on smart objects and technologies for social good. New York: Association for computing machinery; 2020. pp. 84–89. https://doi.org/10.1145/3411170.3411226.
83. Docherty B, Neunschwander E, Karir M, Flinner K. Losing humanity: the case against killer robots. Commun: ACM; 2012.
84. Parasuraman R, Wickens CD. Humans: still vital after all these years of automation. Hum Factors. 2008;50:511–20. https://doi.org/10.1518/001872008X312198.
85. Endsley MR. From here to autonomy: lessons learned from human-automation research. Hum Factors. 2017;59:5–27.
86. Caballero-Martin D, Lopez-Guede JM, Estevez J, Graña M. Artificial intelligence applied to drone control: a state of the art. Drones. 2024;8:296. https://doi.org/10.3390/drones8070296.
87. Dakalbab F, Talib MA, Nasir Q, Saroufil T. Artificial intelligence techniques in financial trading: a systematic literature review. J King Saud Univ Comput Inf Sci. 2024;36:102015. https://doi.org/10.1016/j.jksuci.2024.102015.
88. Mahboubi A, Luong K, Aboutorab H, Bui HT, Jarrad G, Bahutair M, Camtepe S, Pogrebna G, Ahmed E, Barry GH. Evolving techniques in cyber threat hunting: a systematic review. J Netw Comput Appl. 2024;232: 104004. https://doi.org/10.1016/j.jnca.2024.104004.
89. Hou K, Hou T, Cai L. Exploring trust in human–AI collaboration in the context of multiplayer online games. Systems. 2023;11:217. https://doi.org/10.3390/systems11050217.
90. Gunning D, Stefik M, Choi J, Miller T, Stumpf S, Yang G-Z. XAI—explainable artificial intelligence. Sci Robot. 2019;4:1–7. https://doi.org/10.1126/scirobotics.aay7120.
91. Yang W, Wei Y, Wei H, Chen Y, Huang G, Li X, Li R, Yao N, Wang X, Gu X, Amin MB, Kang B. Survey on explainable AI: from approaches, limitations and applications aspects. Hum Centric Intell Syst. 2023;3:161–88. https://doi.org/10.1007/s44230-023-00038-y.
92. Laptev VA, Feyzrakhmanova DR. Application of artificial intelligence in justice: current trends and future prospects. Hum-Centric Intell Syst. 2024;4:394–405. https://doi.org/10.1007/s44230-024-00074-2.
93. Treacy S. Mechanisms and constraints underpinning ethically aligned artificial intelligence systems: an exploration of key performance areas. Hum Centric Intell Syst. 2023;3:189–96. https://doi.org/10.1007/s44230-023-00036-0.
94. Machireddy JR, Rachakatla SK, Ravichandran P. Leveraging AI and machine learning for data-driven business strategy: a comprehensive framework for analytics integration. Afr J Artif Intell Sustain Dev. 2021;1:12–150.
95. Bächle TC, Bareis J. ‘Autonomous weapons’ as a geopolitical signifier in a national power play: analysing AI imaginaries in Chinese and US military policies. Eur J Futur Res. 2022;10:20. https://doi.org/10.1186/s40309-022-00202-w.
96. Tai JEM. Coactive design in systems engineering: human-machine teaming in search and rescue (SAR) operations. Monterey, CA USA: Naval Postgraduate School; 2021.
97. Mascolo MF. Chapter Eight—Developing through relationships: an embodied coactive systems framework. In: Lerner RM, Benson JB, editors. Advances in child development and behaviour, vol. 45. Oxfordshire: Routledge; 2013. p. 185–225.
98. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, Shamseer L, Tetzlaff JM, Akl EA, Brennan SE, Chou R, Glanville J, Grimshaw JM, Hróbjartsson A, Lalu MM, Li T, Loder EW, Mayo-Wilson E, McDonald S, McGuinness LA, Stewart LA, Thomas J, Tricco AC, Welch VA, Whiting P, Moher D. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. J Clin Epidemiol. 2021;134:178–89. https://doi.org/10.1016/j.jclinepi.2021.03.001.
99. Brignardello-Petersen R, Santesso N, Guyatt GH. Systematic reviews of the literature: an introduction to current methods. Am J Epidemiol. 2025. https://doi.org/10.1093/aje/kwae232.
100. Tsamados A, Floridi L, Taddeo M. Human control of AI systems: from supervision to teaming. AI Ethics. 2024;5:1535–48. https://doi.org/10.1007/s43681-024-00489-4.
101. Kamruzzaman A, Thakur K, Ali ML. Cybersecurity threats using application programming interface (API). In: 2024 International conference on computing, internet of things and microwave systems (ICCIMS). 2024; pp. 1–6. https://doi.org/10.1109/ICCIMS61672.2024.10690413.
102. Lamperti F. An information theoretic criterion for empirical validation of simulation models. Econom Stat. 2017;5:83–106. https://doi.org/10.1016/j.ecosta.2017.01.006.
103. Tieleman S. Towards a validation methodology for macroeconomic agent-based models. Comput Econ. 2022;60:1507–27. https://doi.org/10.1007/s10614-021-10191-w.
104. Yin RK. Case study research and applications: design and methods. 6th ed. Thousand Oaks; 2017.
105. Kavak H, Padilla JJ, Vernon-Bido D, Diallo SY, Gore R, Shetty S. Simulation for cybersecurity: state of the art and future directions. J Cybersecurity. 2021;7:1–13. https://doi.org/10.1093/cybsec/tyab005.
106. Teresi JA, Yu X, Stewart AL, Hays RD. Guidelines for designing and evaluating feasibility pilot studies. Med Care. 2022;60:95–103. https://doi.org/10.1097/MLR.0000000000001664.