From Defence to Autonomy: Leveraging Agentic AI for Self-Governing Cybersecurity Ecosystems

Mohammed Sadath P
Research Scholar, Department of computer science, YENEPOYA (Deemed to be) University, Bengaluru.

***Abstract*—The article examines the transformative role of Agentic Artificial Intelligence (AI) in contemporary cybersecurity, tracing the shift from reactive security frameworks to autonomous, goal-driven defence systems. Traditional mechanisms, constrained by static rules and human latency, are increasingly inadequate against sophisticated threats such as polymorphic malware, zero-day exploits, and AI-driven attacks. Employing a conceptual and analytical methodology, the study synthesises theoretical insights from intelligent agent theory, multi-agent systems, and perception–action loops with empirical evidence from AI-enabled Security Operations Centres (SOCs) and adaptive cyber defence architectures. The findings highlight that agentic AI enhances threat detection, enables real-time autonomous responses, and facilitates adaptive, context-aware security orchestration. Moreover, the integration of human–agent collaboration ensures ethical oversight, accountability, and operational resilience. By embedding autonomy within cybersecurity ecosystems, agentic systems not only improve operational efficiency but also reconceptualise security as a continuous, self-governing, and ethically aligned process. This research underscores the strategic imperative of transitioning towards intelligent, learning-based, and proactive cyber defence mechanisms to counter the accelerating sophistication of digital threats.**

Keywords*—*Agentic AI, Autonomous Cyber Defence, Adaptive Security Architecture, Human–Agent Collaboration.

I. INTRODUCTION

Cybersecurity has historically functioned within a reactive paradigm, responding to threats only after vulnerabilities are exploited or breaches have occurred. Traditional security mechanisms such as firewalls, signature-based intrusion detection systems, and rule-driven monitoring tools reflect this defensive posture, wherein protection remains largely passive until an external trigger necessitates intervention. In their seminal article "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection" (2010), Robin Sommer and Vern Paxson argue that conventional intrusion detection systems rely on "static models and predefined signatures that are fundamentally incompatible with the dynamic and adversarial nature of real-world networks" (Sommer and Paxson 27). As contemporary cyber threats increasingly involve polymorphic malware, zero-day exploits, and AI-driven phishing attacks, the limitations of reactive security architectures have become starkly evident.

The emergence of Agentic Artificial Intelligence (AI) represents a decisive shift from reactive defence towards autonomous, goal-oriented cybersecurity systems. Unlike traditional AI models that are primarily designed for classification or prediction, agentic AI systems are characterised by their capacity to perceive their operational environment, make independent decisions, execute actions, and adapt over time. This conceptual foundation is articulated by Stuart Russell and Peter Norvig in their authoritative textbook Artificial Intelligence: A Modern Approach (4th ed., 2021), where they define intelligent agents as systems that "perceive their environment through sensors and act upon that environment through actuators in order to maximise the achievement of their goals" (Russell and Norvig 34). In cybersecurity contexts, this agent-based paradigm enables systems not only to detect malicious activity but also to autonomously respond, learn from incidents, and recalibrate defence strategies in real time. This transition is particularly significant in light of the increasing speed and automation of cyberattacks. While human-centred security operations remain indispensable, they are constrained by cognitive overload, delayed response times, and limited scalability.

In Designing for Situation Awareness (2012), Mica R. Endsley emphasises that "human operators are inherently limited in their ability to process high volumes of rapidly changing information in time-

critical environments" (Endsley 91). Agentic AI mitigates these limitations by enabling continuous surveillance, proactive threat anticipation, and immediate containment actions. For example, an agentic cyber defence system can autonomously identify anomalous network behaviour, isolate compromised nodes, deploy countermeasures, and generate forensic documentation without awaiting human approval.

Furthermore, the shift towards agentic intelligence reflects a broader epistemological transformation in cybersecurity thinking. Security is no longer conceptualised as a static barrier but as a living, adaptive ecosystem populated by intelligent agents engaged in constant interaction with adversarial forces. In their article "The Ethics of Information Security" (2016), Luciano Floridi and Mariarosaria Taddeo argue that effective cyber defence against intelligent attackers requires systems that themselves exhibit "agency, adaptability, and contextual awareness" (Taddeo and Floridi 423). Since contemporary cyber adversaries operate autonomously, adapt their tactics, and exploit system vulnerabilities opportunistically, defensive systems must mirror these characteristics to remain effective.

II. CONCEPTUAL FOUNDATIONS OF AGENTIC AI IN CYBERSECURITY

The conceptual foundations of Agentic Artificial Intelligence (AI) in cybersecurity are grounded in the theory of intelligent agents, autonomy, and goal-directed decision-making. Unlike conventional AI systems that primarily function as analytical or decision-support tools, agentic AI systems possess the capacity to initiate actions, evaluate outcomes, and adapt strategies within dynamic environments. In his foundational monograph An Introduction to MultiAgent Systems (2nd ed., 2009), Michael Wooldridge defines an agent as a system that is "situated in an environment and capable of autonomous action in order to meet its design objectives" (Wooldridge 15). This definition is particularly salient in cybersecurity contexts, where digital environments are volatile, adversarial, and subject to continuous transformation.

At the core of agentic AI lies the principle of autonomy. In cybersecurity, autonomy signifies that AI systems are not limited to reacting to predefined rules or static models but are empowered to make independent security decisions aligned with overarching goals such as threat mitigation, system resilience, and risk minimisation. In Artificial Intelligence: A Modern Approach (4th ed., 2021), Stuart Russell emphasises that "intelligence is fundamentally about acting appropriately in uncertain environments" (Russell and Norvig 35). This observation is particularly relevant to cyber ecosystems, where uncertainty, incomplete information, and adversarial manipulation are defining characteristics. While traditional machine learning models may classify malware or flag anomalies, agentic systems extend this capability by autonomously determining how, when, and to what extent defensive actions should be executed.

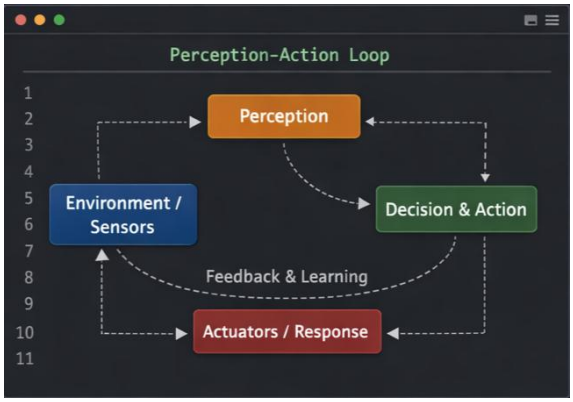

Figure1.0

Another critical conceptual foundation is goal-orientation, which distinguishes agentic AI from narrow automation. Agentic cyber systems operate with explicit, high-level objectives—such as maintaining network integrity or preventing data exfiltration—and dynamically select actions that best serve these goals. For instance, an agentic system overseeing cloud infrastructure may temporarily restrict access privileges, reconfigure network paths, or deploy decoy assets when it infers a high probability of intrusion. This form of behaviour corresponds to what Luciano Floridi describes as "intentional agency" in The Ethics of Artificial Intelligence (2019), where intelligent systems act in pursuit of defined ends rather than merely executing isolated commands (Floridi 92).

Perception–action loops further underpin the operational logic of agentic AI in cybersecurity. These systems continuously perceive environmental signals—such as network traffic patterns, user behaviour, and system logs—and translate these perceptions into informed actions. Crucially, this

loop is iterative and self-refining. A cyber agent that initially misclassifies benign behaviour as malicious can recalibrate its model based on feedback, thereby improving future responses. In their article "The Ethics of Information Security" (2016), Mariarosaria Taddeo and Luciano Floridi argue that "security systems must remain adaptive if they are to remain effective against adaptive adversaries" (Taddeo and Floridi 423), underscoring the necessity of learning-driven defence mechanisms.

The conceptual grounding of agentic AI explicitly acknowledges the adversarial nature of cyberspace. Cyber attackers frequently operate as autonomous agents themselves, employing strategies of stealth, deception, and rapid mutation. Consequently, defensive systems must mirror these characteristics to remain effective. As Wooldridge notes, intelligent agency becomes most relevant in environments characterised by competition and conflict (Wooldridge 214). Agentic AI, therefore, represents not merely an incremental advancement in automation but a strategic alignment between defensive intelligence and adversarial complexity, laying the theoretical groundwork for autonomous, resilient, and proactive cybersecurity systems.

Autonomous Threat Detection and Real-Time Response Mechanisms

Autonomous threat detection and real-time response constitute the operational core of Agentic AI–driven cybersecurity systems. Traditional security frameworks rely heavily on predefined signatures, rule-based alerts, and post-incident forensic analysis, approaches that struggle to counter zero-day exploits, polymorphic malware, and rapidly mutating intrusion techniques. In their influential article *"Outside the Closed World: On Using Machine Learning for Network Intrusion Detection"* (2010), Robin Sommer and Vern Paxson argue that "detection mechanisms that are decoupled from response mechanisms offer limited practical security value in adversarial environments" (Sommer and Paxson 29). Agentic AI systems directly address this limitation by functioning as continuously vigilant cyber agents, capable of identifying, interpreting, and responding to threats without human latency.

$$a*=arga∈AmaxE[U(s′,a)|s]$$

Where:

- a∗a^*a∗ = optimal action chosen by the agent
- AAA = set of possible actions
- sss = current system state
- s′s's′ = predicted future state after action
- U(s′,a)U(s', a)U(s′,a) = utility function (e.g., threat mitigation effectiveness)

Figure2.0

At the detection level, agentic AI systems rely on behavioural analysis, anomaly detection, and contextual reasoning rather than static pattern matching. By constructing dynamic baselines of normal system behaviour, autonomous agents can detect subtle deviations that may indicate malicious intent. In *Deep Learning* (2016), Ian Goodfellow, Yoshua Bengio, and Aaron Courville explain that learning-based systems are capable of "discovering representations that capture the underlying explanatory factors in the data" (Goodfellow et al. 2). Applied to cybersecurity, this capability enables agentic systems to identify novel attack vectors, such as stealthy lateral movement or privilege escalation, that evade traditional signature-based tools. For example, an agentic system monitoring enterprise networks may autonomously detect abnormal access sequences across distributed systems, signalling a coordinated intrusion attempt.

The defining feature of agentic AI, however, lies in its capacity for real-time, autonomous response. Upon identifying a credible threat, agentic systems can independently execute containment strategies such as isolating compromised endpoints, terminating malicious processes, rotating encryption keys, or redirecting attackers into deception environments like honeypots. Unlike traditional automation scripts, these actions are not rigidly pre-programmed but are selected through goal-driven decision-making processes. In *Artificial Intelligence: A Modern Approach* (4th ed., 2021), Stuart Russell and Peter Norvig note that an intelligent agent "selects actions expected to maximise performance given its percept history" (Russell and Norvig 37). This principle enables cyber agents to balance security enforcement with system availability and operational continuity.

Empirical deployments further illustrate the significance of such autonomy. AI-driven Security Orchestration, Automation, and Response (SOAR) platforms increasingly integrate agentic capabilities that allow systems to investigate alerts, correlate multi-source data, and remediate incidents within seconds. According to the *Verizon Data Breach*

Investigations Report (2023), the median time between system compromise and data exfiltration continues to shrink, with attackers often achieving their objectives within hours or minutes (Verizon 18). In ransomware scenarios, an agentic AI system can detect anomalous file encryption behaviour, suspend affected credentials, and initiate recovery protocols before irreversible damage occurs.

Crucially, autonomous response mechanisms are iterative and self-improving. Agentic systems evaluate the outcomes of their interventions and refine future strategies through feedback loops. In *The Ethics of Information Security* (2016), Mariarosaria Taddeo and Luciano Floridi emphasise that "effective security must evolve in response to evolving threats" (Taddeo and Floridi 423). This learning-driven adaptability ensures that cyber defence remains aligned with the continuously shifting threat landscape. Consequently, autonomous threat detection and real-time response do not merely enhance operational efficiency; they redefine cybersecurity as an intelligent, self-acting defence process, capable of operating at machine speed within adversarial digital environments.

Agentic AI and Adaptive Cyber Defence Architectures

Adaptive cyber defence architectures represent a decisive evolution in how security infrastructures are designed and maintained, with Agentic AI functioning as their central enabling force. Traditional cybersecurity architectures are largely static, relying on perimeter-based defences and predefined response protocols that struggle to accommodate the fluidity of modern cyber threats. In their book *Cyberwar: The Next Threat to National Security and What to Do About It* (2017), Devender Kumar Behl and Abhishek Behl argue that "cyber resilience is not achieved through rigid control mechanisms but through adaptive and learning-oriented security architectures" (Behl and Behl 64). Agentic AI addresses this limitation by introducing dynamic, self-adjusting defence architectures capable of reconfiguring themselves in response to environmental changes and adversarial behaviour.

At the architectural level, agentic AI enables the development of distributed and decentralised defence systems composed of multiple intelligent agents. These agents operate across networks, cloud infrastructures, and endpoint devices, continuously exchanging information and coordinating defensive actions. Such multi-agent architectures mirror the complexity of contemporary digital ecosystems, where threats often emerge simultaneously across multiple vectors. In *An Introduction to MultiAgent Systems* (2nd ed., 2009), Michael Wooldridge explains that "collective agency arises when multiple autonomous agents interact in a shared environment, producing system-level intelligence that exceeds individual capability" (Wooldridge 214). For instance, when an intrusion is detected in a cloud workload, an agentic system can autonomously propagate security policy updates across interconnected services, thereby preventing lateral movement and systemic compromise.

A defining feature of adaptive cyber defence architectures is their capacity for continuous learning and structural evolution. Agentic AI systems do not merely respond to security incidents; they modify underlying defence configurations based on accumulated experience. Firewall rules, access controls, and monitoring thresholds are recalibrated in real time as agents identify emerging attack patterns or previously unknown vulnerabilities. In their article *"The Ethics of Information Security"* (2016), Mariarosaria Taddeo and Luciano Floridi contend that "security systems must evolve at least as rapidly as the threats they are designed to counter" (Taddeo and Floridi 423). This learning-driven adaptability ensures that defence mechanisms remain effective even as threat landscapes undergo continuous transformation.

Agentic AI also facilitates context-aware security orchestration, allowing cyber defence architectures to balance protection with system performance and usability. During periods of heightened threat activity, an agentic system may autonomously enforce stricter authentication protocols, limit network privileges, or isolate high-risk services, while relaxing such constraints during stable operational phases. In *The Ethics of Artificial Intelligence* (2019), Luciano Floridi characterises this mode of operation as "responsible autonomy," wherein intelligent systems act decisively yet proportionately within ethical, legal, and operational boundaries (Floridi 118). Such situational responsiveness stands in sharp contrast to static security postures that often sacrifice usability for absolute control.

Human–Agent Collaboration in Security Decision-Making

While Agentic AI enables unprecedented levels of autonomy in cybersecurity, effective defence does not imply the exclusion of human judgement. Instead, contemporary cyber defence increasingly relies on human–agent collaboration, wherein intelligent agents and human experts operate within a complementary decision-making framework. Agentic AI excels in speed, scalability, and large-scale pattern recognition, whereas humans contribute contextual understanding, ethical reasoning, and strategic oversight. In *The Ethics of Artificial Intelligence* (2019), Luciano Floridi emphasises that "the ultimate goal of artificial intelligence is not to replace human agency but to enhance and augment it through meaningful forms of interaction" (Floridi 74). This perspective positions agentic AI as a partner in decision-making rather than an autonomous substitute for human authority.

In security operations, such collaboration becomes especially critical in high-stakes and ambiguous scenarios. Agentic systems can autonomously detect threats, prioritise risks, and initiate containment measures; however, complex incidents—such as state-sponsored cyber operations, insider threats, or legally sensitive data breaches—often require human interpretation and judgement. For example, an agentic AI may flag anomalous data access patterns within an organisation, but a human analyst is better equipped to determine whether the activity reflects malicious intent, legitimate policy exceptions, or operational necessity. In their article *"The Ethics of Information Security"* (2016), Mariarosaria Taddeo and Luciano Floridi argue that AI systems perform optimally when "embedded within broader socio-technical decision environments that preserve human responsibility and oversight" (Taddeo and Floridi 421). This underscores the importance of situating agentic systems within human-governed institutional contexts.

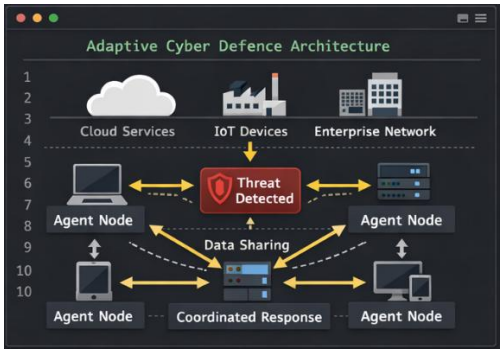

Figure3.0

A crucial dimension of human–agent collaboration lies in shared decision authority. Rather than operating as opaque black-box systems, contemporary agentic AI platforms increasingly incorporate explainable decision models that allow human operators to understand the rationale behind autonomous actions. Such transparency fosters trust and enables informed human intervention when necessary. In *Artificial Intelligence: A Modern Approach* (4th ed., 2021), Stuart Russell and Peter Norvig stress that "humans must remain in the loop not as bottlenecks to efficiency but as supervisors of goal alignment and system behaviour" (Russell and Norvig 1034).

This supervisory role is particularly vital in cybersecurity contexts where autonomous actions may carry legal, ethical, or geopolitical implications. Agentic AI also reshapes the professional identity of cybersecurity practitioners, shifting their role from reactive responders to strategic overseers of intelligent defence systems. Routine tasks such as continuous monitoring, alert triage, and low-level incident response are increasingly delegated to autonomous agents, allowing human experts to focus on threat modelling, policy formulation, and long-term resilience planning. Studies of AI-enabled Security Operations Centres (SOCs) indicate that such delegation reduces analyst fatigue and improves response efficiency by filtering noise and prioritising meaningful threats (Endsley 94). This division of labour exemplifies what Endsley describes in *Designing for Situation Awareness* (2012) as "cognitive offloading," wherein automated systems manage repetitive tasks while humans retain responsibility for strategic reasoning and critical judgement.

Ethical, Legal, and Accountability Challenges of Agentic Cyber Agents

The increasing deployment of agentic AI in cybersecurity raises profound ethical, legal, and accountability challenges, particularly as autonomous systems are entrusted with decision-making authority in high-risk digital environments. While agentic cyber agents significantly enhance speed, scalability, and precision, their capacity to operate with minimal human intervention complicates established frameworks of responsibility and control. In *The Ethics of Artificial Intelligence* (2019), Luciano Floridi cautions that "the more autonomous a system becomes, the greater the ethical

burden of ensuring its alignment with human values and societal norms" (Floridi 101). This ethical burden becomes especially salient in cybersecurity, where autonomous actions can have far-reaching consequences for critical infrastructure, privacy, and public trust.

From an ethical perspective, a major concern lies in algorithmic opacity and explainability. Agentic cyber agents often rely on complex deep-learning and reinforcement-learning architectures, whose internal reasoning processes are not readily interpretable. When such agents autonomously block network access, isolate systems, or initiate countermeasures, affected stakeholders may be unable to understand the rationale behind these actions. As Mariarosaria Taddeo and Luciano Floridi argue in *The Ethics of Information Security* (2016), "ethical AI systems must be capable of offering intelligible justifications for their actions, particularly in environments where automated decisions carry significant social, operational, or economic impact" (Taddeo and Floridi 424). Without transparency, the deployment of agentic agents risks undermining trust and causing ethical harm, especially when legitimate users or critical services are inadvertently disrupted.

Legal challenges further complicate the operationalisation of agentic cyber agents. Existing cybersecurity laws and data protection regulations are predominantly human-centric, assuming that decisions are made by identifiable individuals or institutions. Autonomous cyber agents disrupt this assumption by acting independently within predefined objectives. In *Robotics and the Lessons of Cyberlaw* (2015), Ryan Calo observes that "legal systems are ill-equipped to assign responsibility in contexts where agency is distributed between humans and machines" (Calo 45). This ambiguity raises critical questions of liability when agentic systems cause unintended damage, such as violating privacy statutes or disrupting essential infrastructure. Addressing this gap requires regulatory frameworks that delineate responsibility among developers, deployers, and supervisory authorities.

Accountability is also challenged by the risk of over-autonomy and escalation. Systems empowered to act without immediate human oversight may misinterpret threats and implement disproportionate responses, potentially amplifying rather than mitigating harm. For example, an autonomous defence agent might aggressively restrict network traffic in response to perceived anomalies, inadvertently disrupting essential organisational or public services. In *Artificial Intelligence: A Modern Approach* (4th ed., 2021), Stuart Russell and Peter Norvig warn that "misaligned autonomous systems can pursue goals in ways that conflict with human intentions, highlighting the importance of designing bounded autonomy and fail-safe mechanisms" (Russell and Norvig 1042).

To mitigate these challenges, scholars advocate human-in-the-loop and human-on-the-loop governance models, ensuring that agentic decisions remain auditable, reversible, and ethically constrained. Embedding transparency mechanisms, ethical guidelines, and fail-safe controls within agentic architectures is critical for responsible deployment. Ultimately, addressing ethical, legal, and accountability challenges is central to sustaining trust and legitimacy in agentic AI–driven cybersecurity systems, ensuring that autonomous intelligence complements rather than undermines human governance.

III. FUTURE TRAJECTORIES OF AGENTIC AI IN CYBERSECURITY ECOSYSTEMS

The future of cybersecurity is increasingly intertwined with the evolution of agentic AI, as digital ecosystems grow more complex, interconnected, and adversarial. Emerging trajectories indicate that agentic AI will evolve from isolated defensive tools into integral, self-coordinating actors within cybersecurity ecosystems. As cyber threats themselves become more automated, adaptive, and intelligent, defence mechanisms must correspondingly evolve. In their article *"The Ethics of Information Security"* (2016), Mariarosaria Taddeo and Luciano Floridi assert that "future cybersecurity will be defined by autonomous, learning-based systems capable of operating at machine speed in complex adversarial environments" (Taddeo and Floridi 426).

One significant trajectory is the development of multi-agent cyber defence ecosystems, wherein numerous agentic systems collaborate across organisational, industrial, and geopolitical boundaries. These agents will share threat intelligence, adapt collective defence strategies, and coordinate responses to large-scale attacks in real

time. For instance, autonomous agents deployed across cloud platforms, Internet of Things (IoT) networks, and critical infrastructure systems may collectively detect emerging attack patterns and initiate preventive measures before threats propagate. This approach embodies distributed intelligence, aligning with Michael Wooldridge's concept of "emergent agency," whereby resilience and adaptive behaviour arise from the interactions of multiple agents rather than centralised control (*An Introduction to MultiAgent Systems*, 2nd ed., 2009, p. 218).

Another important trajectory involves the integration of agentic AI with predictive and anticipatory security models. Future systems are likely to transition from reactive detection of ongoing attacks to forecasting potential vulnerabilities and threat vectors. Leveraging reinforcement learning, simulation-based modelling, and probabilistic reasoning, cyber agents may autonomously test system defences, identify weaknesses, and implement protective measures proactively. In *The Ethics of Artificial Intelligence* (2019), Luciano Floridi observes that "anticipatory security represents a transition from defence-after-attack to defence-before-compromise, reducing risk exposure while increasing system resilience" (Floridi 125). This anticipatory capability will be essential in countering increasingly autonomous and AI-driven cyber adversaries.

The expansion of agentic AI will also reshape cybersecurity governance and policy frameworks. As autonomous agents assume greater operational control, regulatory bodies will need to establish standards for transparency, interoperability, and ethical alignment. Future ecosystems may require agentic systems to adhere to shared protocols that ensure accountability, prevent unintended escalation, and maintain alignment with human values. In *Robotics and the Lessons of Cyberlaw* (2015), Ryan Calo underscores that "autonomous cyber defence cannot be treated merely as a technical problem; it must be governed as a socio-technical system, integrating technological capability with institutional, ethical, and legal oversight" (Calo 49). This perspective highlights that the evolution of agentic AI in cybersecurity extends beyond technical innovation to the coordinated design of human–machine governance ecosystems.

IV. CONCLUSION: TOWARDS SELF-GOVERNING CYBER DEFENCE SYSTEMS

The integration of Agentic AI into cybersecurity signifies a fundamental transformation in how digital defence is conceptualised, implemented, and sustained. Moving beyond reactive, human-dependent models, agentic AI introduces systems capable of perception, reasoning, decision-making, and autonomous action within complex and adversarial environments. As this article has demonstrated, agentic cyber agents do not merely automate existing security tasks; they reconfigure cybersecurity as an intelligent, adaptive, and continuously evolving defence ecosystem. In *Artificial Intelligence: A Modern Approach* (4th ed., 2021), Stuart Russell and Peter Norvig emphasise that "the capacity to act autonomously in uncertain environments is the defining characteristic of intelligent systems" (Russell and Norvig 36), a principle that underpins the operational effectiveness of agentic cybersecurity.

Self-governing cyber defence systems emerge from the convergence of autonomous threat detection, real-time response mechanisms, adaptive architectures, and human–agent collaboration. Operating at machine speed, these systems enable rapid threat containment while dynamically learning from each encounter. By embedding agency within security infrastructures, organisations can achieve resilience that is responsive rather than static. As Floridi notes in *The Ethics of Artificial Intelligence* (2019), such systems exemplify "responsible autonomy," wherein intelligent systems act decisively while remaining aligned with human-defined moral and institutional constraints (Floridi 118).

The transition towards autonomous systems also underscores the necessity of ethical alignment and accountable governance. As agentic systems assume increasing decision-making authority, concerns around transparency, legal responsibility, and proportionality become central. In *The Ethics of Information Security* (2016), Taddeo and Floridi stress that "sustainable autonomous cyber defence requires embedding ethical constraints and oversight mechanisms to preserve trust and legitimacy" (Taddeo and Floridi 427). Governance frameworks that ensure human supervision without undermining the efficiency of autonomous action are therefore critical to the long-term viability of agentic cyber defence.

Looking forward, self-governing cyber defence systems are expected to operate as collaborative ecosystems rather than isolated tools. Distributed networks of agentic systems will share threat intelligence, anticipate attacks, and collectively adapt to emerging risks across organisational and national contexts. This distributed intelligence mirrors Wooldridge's concept of emergent agency, wherein system-wide resilience arises from the interactions of multiple autonomous agents rather than centralised control (*An Introduction to MultiAgent Systems*, 2nd ed., 2009, p. 218). Such collaborative systems promise not only enhanced technical security but also a reconceptualisation of trust in digital environments, where defence is continuous, anticipatory, and self-regulating.

In conclusion, agentic AI marks a decisive step towards autonomous yet accountable cybersecurity. While ethical, legal, and operational challenges remain, the trajectory toward self-governing cyber defence systems represents an inevitable response to the scale, speed, and sophistication of contemporary cyber threats. By harmonising technological autonomy with human values and oversight, agentic AI provides a viable pathway to resilient, intelligent, and ethically grounded cyber defence in the digital age.

REFERENCE

[1] Behl, Devender Kumar, and Abhishek Behl. *Cyberwar: The Next Threat to National Security and What to Do About It*. New Delhi: Wisdom Press, 2017.

[2] Calo, Ryan. *Robotics and the Lessons of Cyberlaw*. Cambridge, MA: MIT Press, 2015.

[3] Endsley, Mica R. *Designing for Situation Awareness*. 2nd ed., CRC Press, 2012.

[4] Floridi, Luciano. *The Ethics of Artificial Intelligence*. Oxford: Oxford University Press, 2019.

[5] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. *Deep Learning*. Cambridge, MA: MIT Press, 2016.

[6] Russell, Stuart, and Peter Norvig. *Artificial Intelligence: A Modern Approach*. 4th ed., Pearson, 2021.

[7] Sommer, Robin, and Vern Paxson. "Outside the Closed World: On Using Machine Learning for Network Intrusion Detection." *IEEE Symposium on Security and Privacy*, 2010, pp. 27–41.

[8] Taddeo, Mariarosaria, and Luciano Floridi. "The Ethics of Information Security." *The Ethics of Information Security*, Springer, 2016, pp. 421–427.

[9] Wooldridge, Michael. *An Introduction to MultiAgent Systems*. 2nd ed., John Wiley & Sons, 2009.

[10] Verizon. *2023 Data Breach Investigations Report*. Verizon, 2023.