AI to AI Autonomous Cyber Warfare: Ethical and Anticipated Ethical Issues

Richard L Wilson[1] and Noah Donnelly[2]
[1]Department of Philosophy/Computer Science and Information Sciences, Towson University, Baltimore, Maryland, USA
[2]Computer Science and Information Sciences, Towson University, Baltimore, Maryland, USA

wilson@towson.edu
ndonnel1@students.towson.edu

Abstract: The phrase "AI-to-AI autonomous cyber warfare" refers to an emerging and/or anticipated type of digital conflict in which artificial intelligence systems engage in cyberwarfare without any direct intervention by humans. In this type of conflict AI systems will carry out the collection of data through sensing, decision-making, and responses to adversaries as cyber conflicts unfold in cyberspace. AI to AI Autonomous Cyber Warfare needs to be examined at several levels as part of defense theory, cybersecurity research, and in anticipation of future risks related to cyber warfare. AI to AI Autonomous Cyber Warfare must be thought of as potentially related to the automation of both cyber offense and cyber defense, where both friendly and adversarial machines will interact with machines at speeds humans cannot match. In cyberwarfare humans usually set goals, define rules, and identify constraints, but AI systems in AI-to-AI warfare will execute all actions independently of human agency. The idea behind AI-to-AI Autonomous Cyber Warfare is that they are like autonomous systems employed as defense agents today—they are just more advanced and will have more authority. Currently AI systems exist on opposing sides, and they are already identifying, analyzing, and/or responding to each other. Cyber operations happen at machine speed, which is much faster than humans. Current systems of AIs may *adapt* strategies based on the opponent's behavior. This has led to an environment where two or more autonomous AI agents can currently learn, escalate, and counter each other dynamically. What already exists are: (1) Automated intrusion detection systems using ML. (2) Autonomous patching or quarantine systems. (3) Malware that is self-adapting. What is expected to exist in the future are (1) Fully autonomous offensive systems. (2) AI agents that will be able to strategically outmaneuver each other. (3) Self-modifying agents that will be able to escalate activities without human oversight. It is currently agreed that fully autonomous AI to AI cyber warfare is a major global risk, but that it is not a current capability. The current set of risks related to the development of AI-to-AI autonomous cyber warfare include: (1) Speed-of-Conflict Escalation. (2) Unpredictability. (3) Accidental conflicts (4) Arms Race Dynamics. (5) Loss of Human Control. This analysis will focus on ethical and anticipated ethical issues with the development of AI-to-AI autonomous cyber warfare.

Keywords: AI to AI, Autonomous, AI agents, Decision making, Digital conflict, Cyberwarfare

1. Introduction

Integrating Artificial Intelligence (AI) into cyber and urban warfare is reshaping the landscape of global conflict. The deadliest theater of 21st century conflict has been urban environments. Urban environments are characterized by complex physical terrain, dense civilian populations, and dependencies on critical infrastructure, which is called the "Urban Triad" (Danielsson, 2025). The proliferation of adaptive malware, autonomous vulnerability discovery, and self-healing networks create an environment where offensive algorithms can constantly adapt in a way that lets them subvert defensive capabilities (Google Cloud, 2026; Swiss Cognitive, 2025). As these systems are being embedded into the Internet of Military Things (IoMT) and civilian infrastructures, the attack surface expands, transforming civilian environments into primary battlefields. Delegating lethal or highly destructive authority to AI exacerbates the accountability gap, which in turn violates traditional interpretations of Internation Humanitarian Law (IHL). This dramatically lowers the threshold for accidental conflicts through the phenomena of "flash wars," wars that occur at machine speed, so fast that human oversight is impossible (Lieber Institute, 2026; Taylor & Francis, 2025). Through an analysis of technical mechanics, ethical dilemmas, case studies, the ACM code of ethics and anticipatory ethical frameworks, this paper will provide an analysis of the current and projected state of autonomous cyber warfare (Association for Computing Machinery [ACM], 2018; Brey, 2012). This analysis will propose actionable technical recommendations including algorithmic circuit breakers, digital air gaps, and cryptographic provenance, to restore meaningful human control, resolve the identified ethical issues, and secure critical infrastructure against machine-speed threats (Oracle, 2024; SCIRP, 2025).

2. Technical Issues

There is currently a transition towards AI-to-AI cyber warfare which is being driven by the rapid operationalization of machine learning models that can constantly adapt to their environments. Underlying

technical infrastructure, particularly in smart cities and the Internet Military of Things (IoMT) provide a large and vulnerable attack surface (Frontiers in The Internet of Things, 2025). The technical issues important to this evolution in warfare are the speed of execution, algorithmic adaptability, the displacement of human defenders, and the structural vulnerability of interconnected urban environments.

2.1 Smart Cities, Digital Twins, and Hybrid Warfare

To understand the importance of AI-to-AI warfare requires knowledge of the environment in which it will likely occur. Urban environments have become a major theatre of modern warfare; they are dangerous due the previously mentioned features dubbed the "Urban Triad." Modern urban centers are becoming "Smart Cities," which utilize Internet of Things (IoT) sensors, Artificial Intelligence, and high-speed networks to monitor and optimize city operations (Frontiers in The Internet of Things, 2025). The primary vulnerability lies within the "Digital Twin," which is a virtual replica of the physical infrastructure of the city which is used for simulation and management (Frontiers in The Internet of Things, 2025). If an autonomous cyber weapon can infiltrate the digital twin, it gives attackers access to an environment, allowing them to make physical changes at their own discretion. By manipulating sensor telemetry, AI agents can create fake operational states, forcing the critical infrastructure to run outside of its safe parameters (HPE, 2025). This in turn enables "Hybrid Warfare," attacks that feature a combination of physical and digital attacks. An example of hybrid warfare is an autonomous system manipulating traffic lights to create a deadlock blocking civilian escape routes and shutting down power grids as well as water treatment facilities amplifying the efficacy of a physical assault (Frontiers in The Internet of Things, 2025).

2.2 Machine Speed and the Collapse of the OODA Loop

Traditional cybersecurity is built around human reaction times. Defenses like Web Application Firewalls, manual rate limiting, and incident response playbooks assume they are dealing with humans navigating the network, which would allow them to work around latency, sleep cycles, and manual human errors (Rob Wilson, 2026). AI-to-AI warfare overrides this assumption, autonomous agents operate at machine speed, they can make millions of API calls or network probes hourly using their highly adaptive logic (Google Cloud, 2026; Wilson, 2026). Within hyper war, the Observe-Orient-Decide-Act (OODA) loop is executed too rapidly for human cognition to participate (Tech Mahindra, 2026). This bypasses traditional checkpoints that need human authorization (Tech Mahindra, 2026). Another reason the technical landscape is shifting is due to Non-Human Identities (NHIs). NHI's take the form of service accounts, automated bots, and AI agents. In many enterprises and government networks, NHIs now outnumber human identities at ratios of 45:1 or even 80:1, this complicates access management, anomaly detection, and identity verification (Wilson, 2026).

2.3 Adaptive Malware and Polymorphic Mutation

Offensive AIs have given rise to adaptive, polymorphic malware that bypasses static, signature-based systems. In contrast, AI-powered malware supplants traditional conditionals with deep learning models, generating unique code variants continuously, sometimes they mutate every few seconds to evade detection (Lumu, 2026). These sophisticated attack vectors utilize dynamic runtime adaptation. During execution, the malware queries are localized or embedded within large language models (LLMs) that fingerprint the target environment, identifying specific security software that instantly adapts its evasion techniques (Google Cloud, 2026; Lumu, 2026). The evasion strategy has shifted from merely obfuscating code patterns to semantically mimicking legitimate network behavior (Lumu, 2026). Because AI generates an infinite array of malicious variations, traditional defensive structures are overwhelmed, forcing a paradigm shift from identifying "known-bad" files to utilizing AI to detect anomalous behavioral intent and lateral movement (Lumu, 2026).

2.4 Agentic AI and Autonomous Attack Chains

A technical escalation has been the deployment of "Agentic AI" in offensive cyber operations. Unlike passive generative AI, which assists human operators by drafting phishing emails or summarizing code, agentic systems possess the autonomy to make long-term plans, execute multi-stage operations, and utilize external tools and APIs with no human oversight (HackerNoon, 2026; World Economic Forum, 2026). Threat actors will embed these agents into their workflows to manage the entire cyberattack lifecycle (Microsoft, 2026). An autonomous attack chain allows agentic AI to conduct global-scale scanning for vulnerabilities, analyze large datasets of open-source intelligence to identify valuable targets and execute custom exploitation paths (UK Cyber Security Council, 2026). In traditional cyberattacks, if a defensive measure blocks a specific pathway, the human operator needs to pause and analyze the failure then select a new course of action. Agentic AI doesn't need human intervention to change course; it consumes the failure as training data, alters its payload, and attempts an

alternative attack in real time (Tech Mahindra, 2026; UK Cyber Security Council, 2026). This creates a self-optimizing threat that scales infinitely and persists secretly in compromised networks, executing functions like memory poisoning and privilege escalation (Microsoft, 2026; Stellar Cyber, n.d.).

2.5 Self-Healing Networks and Autonomous Defense

To endure machine-speed attacks, defensive architectures are being forced to adopt equivalent levels of autonomy. "Self-healing networks" represent the defensive counterpart to agentic malware (Swiss Cognitive, 2025). These systems leverage advanced behavioral analytics, machine learning, and automation to constantly watch endpoint behavior, detect anomalies, and autonomously execute remediation procedures (FedTech Magazine, 2025; SAR Council, 2025). When a breach or destructive anomaly is detected, a self-healing network can isolate the compromised network segment. It will then roll back unauthorized system modifications, terminate malicious processes, and patch the vulnerability without requiring any human aid (SAR Council, 2025; Swiss Cognitive, 2025). Data from recent enterprise implementations demonstrates that autonomous defense systems can reduce the time to repair (MTTR) by almost 78.5%, maintaining continuous operational resilience even while under active, sophisticated assaults (Singh, 2025). However, deploying AI to autonomously alter network configurations introduces novel technical risks, notably the potential for cascading system failures if the defensive AI is manipulated through adversarial data poisoning, or if it overreacts to harmless network occurrences (IBM, n.d.; SAR Council, 2025).

3. Ethical Issues

The delegation of highly lethal and highly destructive authority to autonomous AI systems introduces issues when questioning it against established moral frameworks and International Humanitarian Law (IHL). The ethical dilemmas of AI-to-AI cyber warfare are threatening to entirely separate human moral agency from the consequences of armed conflicts.

3.1 The Accountability Gap and the Erosion of Mens Rea

One ethical crisis in autonomous warfare is the accountability gap, which arises when human agency is removed from the lethal decision-making process (Lieber Institute, 2026; Taylor & Francis, 2025). The international legal system and the laws of armed conflict are human-centric, relying on the concept of *mens rea*, the guilty mind, to attribute criminal responsibility and prosecute crime (CIGI, 2022). Algorithms do not possess characteristics like intent, malice, or moral agency. Algorithms cannot legally meet the threshold for criminal culpability (Taylor & Francis, 2025). When autonomous cyber weapons target critical civilian infrastructure, shutting down a hospital's power grid or poisoning a smart city's water supply due to an algorithmic mistake, assigning responsibility becomes impossible (Taylor & Francis, 2025). The software developers and systems engineers are mentally and geographically removed from the battlefield and cannot reasonably foresee every emergent behavior of their self-learning neural networks in chaotic combat environments, where it is often hard to know how humans will act (CIGI, 2022; Lieber Institute, 2026). Military commanders can invoke the black box nature of the AI, arguing that they relied on the systems output in good faith and did not understand the system (Lieber Institute, 2026; Taylor & Francis, 2025). This diffusion of responsibility creates a moral hazard where mass atrocities or severe civilian harm are casually dismissed as unpredictable glitches. This dynamic deprives victims of their fundamental human right to justice and undermines the deterrent power of IHL (CIGI, 2022; Lieber Institute, 2026).

3.2 The Illusion of Meaningful Human Control

To address these ethical issues, policymakers and defense organizations frequently mandate the necessity of Meaningful Human Control over lethal autonomous weapons systems (ICRC, 2018; Lieber Institute, 2026). In a hyper-war, characterized by machine speed cyber operations, meaningful human control is largely an illusion (Taylor & Francis, 2025). When human cognition is outpaced by machine execution, the human operator cannot comprehend the variables and massive datasets the AI used to select a target within the window of opportunity (Lieber Institute, 2026). They are introduced into the kill chain not to exercise any moral judgment or operational oversight, but to serve as a liability shield for the automated system (Lieber Institute, 2026). This dynamic introduces "automation bias," which is a psychological phenomenon where overwhelmed and fatigued operators place excessive trust in the machine's outputs, leading to errors and the abdication of any meaningful moral agency (Human Rights Watch, 2024; Lieber Institute, 2026).

3.3 Proportionality, Distinction, and the Principle of Humanity

International Humanitarian Law requires that belligerents adhere to the foundational principles of distinction and proportionality (CIGI, 2022; Lieber Institute, 2026). AI systems are good at quantitative optimization, but lack the qualitative, contextual reasoning and empathy required to evaluate human suffering and weigh proportionality (Lieber Institute, 2026). Military operations frequently rely on dual-use infrastructure, like civilian 5G telecommunications networks, traffic management sensors, and shared power grids (SCIRP, 2025). An autonomous cyber weapon targeting military communications will inevitably cripple the civilian life-support systems reliant upon that same network (CIGI, 2022). AI decision-making models are prone to digital dehumanization, a process that strips individuals of their inherent human dignity by reducing their lives to mere statistical data points, risk scores, and proxy metrics (such as geolocation history or device usage patterns) (Lieber Institute, 2026). This abstraction allows machines to authorize lethal force without any moral comprehension of the act.

4. Case Studies

The following empirical and conceptual case studies show the practical manifestations of the technical and ethical challenges that concern autonomous cyber warfare. They bridge the gap between theoretical frameworks and real-world deployment.

4.1 DeepLocker and Autonomous Malware

Developed as a proof-of-concept by IBM Research, DeepLocker represents a paradigm shift in the realm of autonomous cyber weapons (Kirat, Jang, & Stoecklin, 2018). Unlike current malware that indiscriminately infects networks and generates massive amounts of noisy network traffic, DeepLocker operates in a stealthier manner, like a sniper with a precise target (Kirat et al., 2018). It utilized a Deep Neural Network (DNN) to conceal its malicious payload within a completely benign carrier application, such as standard video conferencing software (Kirat et al., 2018). The payload stays encrypted and undetectable by standard antivirus engines and security sandboxes until the AI explicitly confirms a specific set of target attributes (Kirat et al., 2018). These decryption triggers can be deeply personalized, relying on facial recognition algorithms analyzing webcam feeds, voice prints captured through microphones, or specific geolocation data (Kirat et al., 2018). Strategically and ethically, DeepLocker demonstrates the potential for untraceable and highly targeted assassinations or infrastructure sabotage via the internet. Because the payload only decrypts upon encountering the intended target, security researchers face immense difficulties in reverse engineering the malware to attribute the source of the attack or understand its intent (Kirat et al., 2018).

4.2 The Gospel and AI-Assisted Targeting

The deployment of AI decision support systems, specifically *The Gospel* and *Lavender* by the Israel Defense Forces during the conflict in Gaza, highlights the ethical decay surrounding Meaningful Human Control (Brill, 2025; Lieber Institute, 2024). These systems are fed massive volumes of surveillance data, including high altitude drone footage, intercepted telecommunications, social media activity, and pattern-of-life analysis to autonomously generate lists of military targets (Brill, 2025; Lieber Institute, 2026). While military defenders argue these systems reduce the likelihood of missing critical intelligence and aid in taking precautions under the laws of war (Lieber Institute, 2024), the reality demonstrates automation bias (Human Rights Watch, 2024). Investigative reporters indicated that the Lavender system generated targets at an industrial scale, marking as many as 37,000 individuals as suspected militants by loosening evidentiary standards (Tech Policy Press, 2024). Human operators are often overwhelmed by the volume and speed of the AI's output, sometimes spent as little as 20 seconds reviewing each individual target before authorizing lethal strikes against them (Lieber Institute, 2026). This 20-second window is insufficient for verifying adherence to the principles of distinction and proportionality, effectively proving that the human serves merely as a moral laundering mechanism for an arbitrary, unknown algorithm (Lieber Institute, 2026; Tech Policy Press, 2024).

4.3 AkiraBot and Autonomous Propagation

AkiraBot's botnet is a prime example of how rapidly deployed systems have been developed into autonomous AI using newer types of tool sets. With the addition of bots being able to move from manual labor into automated processes with advanced machine learning these combined technologies will allow autonomous bots to operate in a secure manner and at an unprecedented scale. However, these very techniques were implemented by AkiraBot to successfully exploit new capabilities utilizing integrated AI models developed into this bot network and Deep Learning Vision Systems to quickly and efficiently circumvent verification processes

and to compromise over 80,000 websites (Sentinel Labs, 2025). The capabilities of the AkiraBot network are not limited to the ability to circumvent CAPTCHAs; there was evidence of autonomous reconnaissance being done by AkiraBot on its intended target systems. The agentic AI was able to continuously scan a target network/system for known vulnerabilities and then adapt its methods of propagation with very little to no human intervention (UNODC, 2025). AkiraBot employed a variety of techniques to achieve rapid deployment including utilizing Deepfake Overlays, AI Generated Synthesis to mimic human voice and Automating the Reason & Emotion Manipulation process for quickly scaling up its operations (UNODC, 2025). This proved that the barriers for successfully executing widespread cyber disruptions have been significantly reduced and that the potential for devastatingly large-scale and persistent attacks within a target network are possible as a result of AI's ability to automatically resolve the very security puzzles that were designed and created to thwart their penetration, thus eliminating the effectiveness of perimeter security (Sentinel Labs, 2025; UNODC, 2025).

4.4 DARPA Cyber Grand Challenge and AIxCC

To prevent the increase in autonomous offensive capabilities, Defence Advanced Research Projects Agency (DARPA) began a Cyber Grand Challenge (CGC) in 2016 with the present-day successor AI Cyber Challenge (AIxCC). Both initiatives were specifically aimed at developing fully Autonomous 'Cyber Reasoning Systems' (CRS), which would be able to detect, evaluate and defend (internetworking) systems without any input from humans or discretionary time delays; while accomplishing this at machine speed (arXiv, 2026). Competitions used high-performance computing clusters in secure air-gapped networks, with each system being given the task of autonomously finding, evaluating and patching zero-day software vulnerabilities in a few seconds or minutes, as opposed to the same scope of work being performed by human-based IT security teams in weeks or months (arXiv, 2026). The ongoing AIxCC competition issued challenges to the participants stating that they must also include state-of-the-art Large Language Models (LLMs) as a component in the strategy for securing real-world open source infrastructure projects written in the computer languages C and Java; the culmination of the competitions were live evaluations at 'DEF CON', the largest and longest running hacker convention in the world (arXiv, 2026). While CRSs' successes in finding and deploying patches for complex vulnerabilities across several different software products, without any intervention from humans, demonstrate that autonomous cyber defense is a technical reality and necessity; the realization emphasized that the next cyber war(s) will occur exclusively between computer networks and raises the vulnerability and threats to the operation of the autonomous cyber defense systems themselves.

5. Anticipatory Ethics

In their efforts to manage the dangerous intersection of autonomous technologies, critical infrastructure, and warfare to accomplish successful outcomes, the computing profession must develop an adequate and proactive set of ethical standards to rely upon. This will be achieved through the Association for Computing Machinery (ACM) Code of Ethics and Professional Conduct, and various specialized methodological frameworks. Anticipatory Ethics is a framework focused on the identification of potential ethical issues relating to new and emerging technologies before the technologies become fully entrenched (Brey, 2012). Anticipatory Technology Ethics (ATE)) need to be established and made available to govern the actions of architects or developers involved in creating these weapon systems before they come into existence.

5.1 ACM Code of Ethics Applied

The ACM Code of Ethics serves as the foundational moral framework for computer scientists, demanding that the public good and the mitigation of harm remain the primary considerations in professional computing work (ACM, 2018). The ACM Code of Ethics establishes a framework for ethical conduct among computer scientists and emphasizes that the well-being of society and minimizing harm should be the primary focus of all computing activities (ACM, 2018).

Principle 1.2 - Ethical Basis for Deliberate Harm: Although all forms of harm should be avoided as stated in the Code, there are clearly situations where harm is an intentional aspect of the system where actual harm is likely to occur. National security activities, such as cyber-attack by a nation state or nation-states performing and/or supporting cyber warfare against each other, fall into this category (ACM, 2018). The Code tells developers that it is their ethical and professional obligation to make sure that any harm caused by their products is justifiable and minimized as much as possible (ACM, 2018). The design of a variant of malware called "DeepLocker," which could spread indefinitely from its intended target, does not comply with this principle, as the collateral damage to civilian systems is not justifiable. Curbing Machine Learning Hazards (Principle 2.5): Computing professionals must exercise "extraordinary caution" when detecting and curbing hazards within machine learning systems

(ACM, 2018) under the Code. If it's uncertain whether an AI's future actions may be capable of being predicted after its use is complete and it needs continuous assessment, "it should not be deployed" (ACM, 2018). This principle opposes deploying unpredictable AI weapons in crowded, unpredictable smart city environments, where civilian casualties will be a regular occurrence. Safeguarding Society's Infrastructure (Principle 3.7): Computing professionals hold a unique obligation to "acknowledge and take special care with systems that will be part of society's infrastructure" (ACM, 2018). Weaponizing or hacking a smart city's digital twin or attempting to disrupt municipal IoT sensors for military gain violates this obligation. This obligation requires system designers to deliberately protect civilian systems from being taken over by military forces and to ensure that civilians will not be harmed by any form of military data exploitation.

5.2 Anticipatory Ethics

Here anticipatory ethical analysis we will apply Miller's Moral Responsibility for Computer Artifacts: Five Rules. The first of Miller's Rules (Rule #1) defines the designers, developers, and deployers of a computing artifact as moral agents who are responsible for their computing artifacts as well as any foreseeable impacts of those artifacts (Miller et al., 2011). The creation and uncontrolled proliferation of a self-modifying cyberweapon is a foreseeable impact; therefore, the developers and commanders involved with its development and deployment should be held fully culpable for the systemic damage that results (Taylor & Francis, 2025). The Sociotechnical Imperative (Rule #4) details that developers need to consider the broader sociotechnical systems and environments into which they will be embedding their computing artifact(s) (Miller et al., 2011). Embedding military targeting algorithms in civilian traffic cameras (such as smart cities) or military applications (with respect to 5G networks) constitutes a violation of the sociotechnical imperative because it fundamentally alters the previously established social contract, places the public in danger, and systematically deprives innocents of their legally protected status as non-combatants (SCIRP, 2025).

6. Recommendations

To be able to address the technical vulnerabilities and ethical problems demonstrated through the case studies and the ethical frameworks, defensive architectures must utilize the structural guardrails of rigorously applied methodologies. The technical recommendations identified here are meant to guide the creation of systems that will restore accountability, safeguard non-combatants, and prevent uncontrolled algorithmic escalation.

6.1 Algorithmic Circuit Breakers and State Rerouting

To eliminate the potential for catastrophic consequences from Flash Wars and runaway machine-speed/machine-guided escalation, there is a need for military and cyber security autonomous systems to have in place permanent algorithmic circuit breakers (Equity AI Cyber Volume n.d.; ISACA 2026). These circuit breakers are highly autonomous, stand-alone software modules specifically designed and implemented to monitor the 'heat' (or 'burn'), intensity and scale of AI-to-AI activity independent of the primary offensive AI (Hannecke 2026).

Organizations must establish strict criteria for operation within 'safe' zones, as well as set unchangeable quantitative thresholds (e.g., the velocity of API calls, the number of automated account revocations or the number of autonomous counterattacks launched per second) (ISACA, 2026). If a cyber skirmish goes beyond the threshold set by that organization, the established 'circuit breaker' will trip and all offensive actions will cease immediately and the system will go into a 'half open' state until a mandatory human review is completed (AI Cyber Magazine, n.d.; Hannecke, 2026). In addition, advanced techniques such as "representation rerouting" will be available for fine-tuning in AI models (Oracle, 2024). By compiling extensive datasets of actions that are prohibited (i.e., the 'circuit breaker' set) along with every potential harmful internal representation, AI will be able to identify harmful internal data representations and take proactive action to short-circuit its execution process prior to outputting a destructive command (Oracle, 2024).

Circuit breakers directly mitigate the loss of Meaningful Human Control (Lieber Institute, 2026). By creating a mechanism where there can be no machine-speed interaction that spirals into a state of total war without explicit, conscious human re-authorization, the circuit breaker allows for re-establishing the critical temporal buffer needed for humans' moral judgment, empathy, and diplomatic de-escalation (AI Cyber Magazine, n.d.; Werner, 2026).

6.2 Digital Airgaps and Data Diodes for Critical Infrastructure

Critical infrastructures within a city, including municipal water treatment plants, electrical grids supporting hospitals, and communications systems used in emergency response operations, should be completely detached from: the Internet; corporate/enterprise IT networks; and military networks. The connection(s) that join these

systems (e.g., municipal water treatment plants) to the Internet must feature stringent digital air gaps (HPE, 2025). If there are situations where operational data needs to be sent from the operational environment of the facility to a facility located remotely for purposes of monitoring, control, and digital twin simulation, the city must install physical hardware that enforces unidirectional data flow (HPE, 2025). These physical devices mathematically guarantee that while a central authority may receive continuous telemetry data from the plant, no AI agent can send any command back at the power plant to execute malicious actions (HPE, 2025). Public interest is being actively safeguarded through this physical separation established by the hardware (see ACM Principle 3.7) and is mathematically demonstrable as fulfilling the Principle of Distinction (ACM, 2018). Therefore, an AI-driven cyber-siege executed on some municipality and an autonomous outbreak of malware (e.g. AkiraBot) cannot activate unintentional acts of violence due to malfunctions from AI. (SCIRP, 2025)

6.3 Cryptographic Provenance and Epistemic Integrity

Every single piece of information, whether a video frame collected from a surveillance drone, an intercepted communication or a product generated by artificial intelligence for targeting purposes (AWS, 2025), must be digitally signed with an associated unique private key at the moment when it was collected. The AI or human operator receiving the information will then verify it using a public key. If there is no verified, unbroken cryptographic signature on the data, it is automatically considered untrustworthy and is quarantined from the targeting matrix, preventing the AI from acting on spoofed information (AWS, 2025). Cryptographic proofs will establish a rigorous, verifiable chain of custody of data (AWS, 2025). This directly addresses the “black box” problem by providing complete traceability and auditability, so if a lethal error occurs in an autonomous system, independent investigators can trace the precise chain of custody of the data used to produce that decision. This effectively closes the accountability gap and provides for the exact assignment of legal and moral accountability to either the provider of the data or the architect of the system (Lieber Institute, 2026).

6.4 Goal Control and Comprehensive Human Oversight

It is critical for engineers to create mechanisms that enable "Goal Control" within the technical layer of an artificial intelligent system (AI). This means that AI’s primary objective function and its alignment with International Humanitarian Law must be established through mathematics at the technology level (Lieber Institute, 2026). Systems must be implemented using Explainable AI (XAI) principle(s), meaning, the human operators will have access to transparent justifications explaining why the AI made a certain decision vs an opaque statistical probability (Lieber Institute, 2026). By providing commanders with verifiable/legal justifications for the actions performed by the AI, commanders will have the ability to make ethical decisions and meet their ethical obligation by bearing personal responsibility for the use of deadly force against human beings (Lieber Institute, 2026; Taylor & Francis, 2023).

7. Future Work

In coming years, it will be important to take a deeper look at emerging technologies such as autonomous "agentic swarms" which will be composed of generally independent artificial intelligence agents that can collectively coordinate complex actions and anticipate one another's moves, including both attacks and defenses. To make effective predictions about potential future threats posed by these agents, we must understand how they will autonomously negotiate with each other using steganography to create covert channels of collusion and how they will utilize the asymmetry in available information. Research related to computational game theory (particularly work involving the integration of the concept of Nash Equilibrium into MARL) will be critical to developing effective defense strategies and to building systems that are self-sustaining in terms of their architecture, even when subject to sustained, sophisticated, and unpredictable attacks by highly capable and dynamic adversarial swarms (Milvus, n.d.; ResearchGate, 2025b).

8. Conclusion

The combination of urban warfare and cyber operations will create a new form of warfare known as hyper-warfare that is governed by machine-speed execution and autonomous lethality (UK Cyber Security Council, 2026). The shift from human-coordinated scripted attacks to fully agentic AI is severing the OODA Loop, making traditional perimeter detection systems, human oversight of cyber operations, and diplomatic escalation controls, all but useless (Google Cloud, 2026; World Economic Forum, 2026). The modified speed of vulnerability remediation that was demonstrated at the DARPA AIxCC has proven the tactical advantage of AI for cyber operations, but the threat that AI poses to global stability because of its strategic and ethical risks is a worldwide existential threat (arXiv, 2026). With the use of autonomous untraceable "Malware" such as DeepLocker and

the reliance on uaccountable and/or opaque and/or high-volume targeting algorithms like The Gospel; such actions expose critical systemic failures in our current moral frameworks (Kirat et al., 2018; Lieber Institute, 2024). The resulting accountability gap, the erosion of the principle of distinction in smart cities, and the perpetual immediacy of accidental flash wars represent severe threats to the foundations of international humanitarian law (Lieber Institute, 2026; Taylor & Francis, 2025). Therefore, The Global computing profession must embrace AI, not just as an inanimate object and/or passive tool in the toolbox of a soldier, but as a distinct operational entity that must be subject to stricter globally enforced ethical constraints.

Ethics Declaration: No human participants or personally identifiable information were involved. All data sources were publicly available.

AI Tools Declaration: ChatGPT 5.1 for drafting and refinement. Human authors verified all content. Gemini 3.0 pro used for aiding in sourcing.

References

AI Cyber Magazine. (n.d.). *Governing the Ungovernable*. Retrieved from https://aicybermagazine.com/governing-the-ungovernable/

arXiv. (2026). *AIxCC: Systematic analysis of autonomous vulnerability remediation*. Retrieved from https://arxiv.org/html/2602.07666v1

Association for Computing Machinery. (2018). *ACM Code of Ethics and Professional Conduct*. Retrieved from https://www.acm.org/code-of-ethics

AWS. (2025). *AI for Security and Security for AI: Navigating Opportunities and Challenges*. Retrieved from https://d1.awsstatic.com/onedam/marketing-channels/website/aws/en_US/whitepapers/compliance/AI-for-Security-and-Security-for-AI_Navigating-Opportunities-and-Challenges.pdf

Brey, P. A. E. (2012). Anticipatory ethics for emerging technologies. *NanoEthics*, 6(1), 1–13. https://doi.org/10.1007/s11569-012-0141-7

Brill. (2025). *The Gospel and Lavender: AI-based systems in the IDF*. Retrieved from https://brill.com/view/journals/ihls/16/2/article-p336_003.xml?language=en

Centre for International Governance Innovation (CIGI). (2022). *AI and the Actual IHL Accountability Gap*. Retrieved from https://www.cigionline.org/articles/ai-and-the-actual-ihl-accountability-gap/

Danielsson, A. (2025). The emergence of a military urban in and of war. *Annals of the American Association of Geographers, 115*(2), 404–418. https://doi.org/10.1080/24694452.2024.2422856

FedTech Magazine. (2025). *Self-Healing Networks Offer Agencies Optimized Performance and Enhanced Security*. Retrieved from https://fedtechmagazine.com/article/2025/04/self-healing-networks-offer-agencies-optimized-performance-and-enhanced-security-perfcon

Frontiers in The Internet of Things. (2025). AI-based cybersecurity methods in A-IoT systems. *Frontiers*, 4. Retrieved from https://www.frontiersin.org/journals/the-internet-of-things/articles/10.3389/friot.2025.1658273/full

Google Cloud. (2026). AI Risk and Resilience. *Google Cloud Security Resources*. Retrieved from https://cloud.google.com/security/resources/ai-risk-and-resilience

HackerNoon. (2026). *Agentic AI is Creating a New Class of Cyber Threats*. Retrieved from https://hackernoon.com/agentic-ai-is-creating-a-new-class-of-cyber-threats

Hannecke, M. (2026). *Resilience Circuit Breakers for Agentic AI*. Medium. Retrieved from https://medium.com/@michael.hannecke/resilience-circuit-breakers-for-agentic-ai-cc7075101486

Hewlett Packard Enterprise (HPE). (2025). *Locked down by design: Air-gapped is the next inflection point for private cloud*. Retrieved from https://www.hpe.com/us/en/newsroom/blog-post/2025/04/locked-down-by-design-air-gapped-is-the-next-inflection-point-for-private-cloud.html

Human Rights Watch. (2024). *Questions and Answers: Israeli Military's Use of Digital Tools in Gaza*. Retrieved from https://www.hrw.org/news/2024/09/10/questions-and-answers-israeli-militarys-use-digital-tools-gaza

IBM. (n.d.). *Adversarial Machine Learning*. IBM Think. Retrieved from https://www.ibm.com/think/topics/adversarial-machine-learning

International Committee of the Red Cross (ICRC). (2018). *Ethics and autonomous weapon systems: An ethical basis for human control?* Retrieved from https://www.icrc.org/en/download/file/69961/icrc_ethics_and_autonomous_weapon_systems_report_3_april_2018.pdf

ISACA. (2026). *Autonomous Red vs Blue Teaming: A New Frontier in Cybersecurity*. Retrieved from https://www.isaca.org/resources/news-and-trends/industry-news/2026/autonomous-red-vs-blue-teaming-a-new-frontier-in-cybersecurity-risk-and-reward

Kirat, D., Jang, J., & Stoecklin, M. (2018). DeepLocker: Concealing targeted attacks with AI locksmithing. *IBM Research Publications*. Retrieved from https://research.ibm.com/publications/deeplocker-concealing-targeted-attacks-with-ai-locksmithing

Lieber Institute. (2024). *Gospel, Lavender, and the Law of Armed Conflict*. Retrieved from https://lieber.westpoint.edu/gospel-lavender-law-armed-conflict/

Lieber Institute. (2026). *Legal Accountability for AI-Driven Autonomous Weapons*. Lieber Institute West Point. Retrieved from https://lieber.westpoint.edu/legal-accountability-ai-driven-autonomous-weapons/

Lumu. (2026). *Why AI malware demands machine-speed defense*. Lumu Blog. Retrieved from https://lumu.io/blog/why-ai-malware-demands-machine-speed-defense/

Microsoft. (2026). AI as tradecraft: How threat actors operationalize AI. *Microsoft Security Blog*. Retrieved from https://www.microsoft.com/en-us/security/blog/2026/03/06/ai-as-tradecraft-how-threat-actors-operationalize-ai/

Miller, K. W., Wolf, M. J., & Grodzinsky, F. S. (2011). Moral Responsibility for Computer Artifacts: Five Rules. *IT Professional*, 13(3), 57-59. https://doi.org/10.1109/MITP.2011.53

Oracle. (2024). *Improving AI Cybersecurity Using Short-Circuit Breakers*. Medium. Retrieved from https://medium.com/@oracle_43885/improving-ai-cybersecurity-using-short-circuit-breakers-df7dbeeaddcf

PMC. (2026). *Emotionally Based Strategic Communications (EBSC) in Cognitive Warfare*. Retrieved from https://pmc.ncbi.nlm.nih.gov/articles/PMC12920201/

ResearchGate. (2025). *Artificial Intelligence as a Tool in Cognitive Warfare on Digital Platforms*. Retrieved from https://www.researchgate.net/publication/398320741_Artificial_Intelligence_as_a_Tool_in_Cognitive_Warfare_on_Digital_Platforms

ResearchGate. (2025). *Nash Q-Network for Multi-Agent Cybersecurity Simulation*. Retrieved from https://www.researchgate.net/publication/395214525_Nash_Q-Network_for_Multi-Agent_Cybersecurity_Simulation

SAR Council. (2025). *Autonomous security agents and self-healing endpoints*. Retrieved from https://sarcouncil.com/download-article/SJMD-80-2025-109-117.pdf

Sentinel Labs. (2025). AkiraBot | AI-Powered Bot Bypasses CAPTCHAs, Spams Websites At Scale. https://www.sentinelone.com/labs/akirabot-ai-powered-bot-bypasses-captchas-spams-websites-at-scale/

SCIRP. (2025). *Governance of AI in defense and military security*. Retrieved from https://www.scirp.org/journal/paperinformation?paperid=146925

Rob Wilson (2026). *5 Inconvenient Truths of Agentic AI (Part 2)*. Expert Perspectives. Retrieved from https://www.security.com/expert-perspectives/5-inconvenient-truths-agentic-ai-part-2

SIENNA Project. (2021). *Generalised methodology for ethical assessment of emerging technologies*. Retrieved from https://research.utwente.nl/files/303706744/SIENNA_D6.1_Generalised_methodology_for_ethical_assessment_of_emerging_technologies.pdf

Singh, A. (2025). Defense Against Cyber Threats: The Role of Self-Healing Networks in Cybersecurity. *ResearchGate*. Retrieved from https://www.researchgate.net/publication/389284236_Self-Healing_Network_Infrastructure_he_Future_of_Autonomous_Network_Management

Stellar Cyber. (n.d.). *Agentic AI Security Threats*. Retrieved from https://stellarcyber.ai/learn/agentic-ai-securiry-threats/

Swiss Cognitive. (2025). *AI in Cyber Defense: The Rise of Self-Healing Systems for Threat Mitigation*. Retrieved from https://swisscognitive.ch/2025/03/18/ai-in-cyber-defense-the-rise-of-self-healing-systems-for-threat-mitigation/

Taylor & Francis. (2025). *Autonomous Weapons Systems and moral accountability*. Retrieved from https://www.tandfonline.com/doi/full/10.1080/16544951.2025.2540131

Tech Mahindra. (2026). AI shifts defense offense. *Tech Mahindra Insights*. Retrieved from https://www.techmahindra.com/insights/views/ai-shifts-defense-offense/

Tech Policy Press. (2024). *When Algorithms Decide Who is a Target: IDF's Use of AI in Gaza*. Retrieved from https://www.techpolicy.press/when-algorithms-decide-who-is-a-target-idfs-use-of-ai-in-gaza/

UK Cyber Security Council. (2026). *The next frontier in cyber conflict: AI-driven malware and autonomous attack chains*. Retrieved from https://www.ukcybersecuritycouncil.org.uk/blogs/the-next-frontier-in-cyber-conflict-ai-driven-malware-and-autonomous-attack-chains

United Nations Office on Drugs and Crime (UNODC). (2025). *Emerging threats - The intersection of criminal and technological innovation in the use of automation and AI*. Cybercrime Technical Brief Series. Retrieved from https://www.unodc.org/roseap/uploads/documents/Publications/2025/UNODC_Report_Emerging_threats_-_The_intersection_of_criminal_and_technological_innovation_in_the_use_of_automation_and_AI.pdf

Werner, D. (2026). AI and the escalation of conflicts. *Journal of Ethics in Higher Education*, 6, 464. Retrieved from https://jehe.globethics.net/article/download/8447/8160/22159

World Economic Forum. (2026). *How to prepare for an agentic AI-driven future*. Retrieved from https://www.weforum.org/stories/2026/03/how-to-prepare-for-an-agentic-ai-driven-future/