CYBERSECURITY RISKS FROM AI–PHYSICAL CONVERGENCE

Vladimir Babanov
South-West University "Neofit Rilski", Blagoevgrad, Bulgaria, v.babanov@law.swu.bg

Abstract: The primary goal of this research is to identify the extent to which ongoing artificial intelligence integration into physical systems creates new types of multidomain cyber threats. It also aims to illustrate how the combination of AI and physical systems creates unique and unpredictable security risks with the potential to cause physical harm. In order to accomplish these objectives, a conceptual framework based on literature, defense analysis and threat modeling was applied to identify and categorize the major vulnerability areas in AI-physical systems such as systemic vulnerabilities, adverse input vulnerabilities, network intrusion vulnerabilities, and supply chain vulnerabilities. Data collected through case studies and documented events such as Stuxnet and hacking of autonomous vehicle systems, indicated that AI-physical system convergence significantly increases uncertainty and therefore allows for sensor spoofing, model corruption and stealthy adversarial attacks. Additionally, these data indicate that AI-physical system convergence has the potential for causing physical damage and operational failure. Current Information Technology-centric cybersecurity paradigms are inadequate for protecting AI-physical systems, particularly when those systems are used within critical infrastructure or defense applications due to the rapid escalation of attacks and the difficulty of attributing them to a specific party. The convergence of AI and physical systems blurs the distinctions between digital and physical and forces a reevaluation of traditional responses to cyber threats. Therefore, it is recommended that the development of defensive architectures that include multiple layers, real-time anomaly detection, cross-discipline risk assessments, and the design of secure AI models be employed to mitigate the threats associated with AI-physical system convergence. Further, policymakers and industry stakeholders should recognize that cybersecurity cannot be separated from physical safety when dealing with AI-physical systems. Additional recommendations in the study include the development of updated governance models and the inclusion of cyber-physical resilience in military and civilian plans to prevent catastrophic exploitation of AI-based systems.
Keywords: cybersecurity, artificial intelligence, infrastructure security, autonomous systems, defence

1. INTRODUCTION

The integration of Artificial Intelligence (AI) into physical systems such as robots, drones, intelligent structures and self-driving automobiles is rapidly developing. The process creates unforeseen possibilities for AI adoption into solving persistent problems, but it also brings entirely new forms of threats for the security of the adopted systems. The development of interconnected, AI-controlled physical procedures creates another large attack surface in addition to the existing vulnerabilities. Robots and autonomous machines are increasingly being implemented in critical infrastructure sectors, such as transportation and power grids, making cybersecurity a major priority and risk mitigation factor for the functioning of society. Autonomous Vehicles (AVs) represent a major component of the U.S. transportation system, which is defined by the Department of Homeland Security as a "Critical Infrastructure" (Artusy, 2025). The advancements in AI in real-time sensing and sensor fusion enable AVs and robots to be operated at high levels of autonomy bringing serious cyber, legal and ethical risks. Additionally, the widespread proliferation of the Internet of Things (IoT) encompassing drones, autonomous machines and robots, has created a large number of low-cost devices that are often vulnerable to distributed denial-of-service (DDoS) attacks, botnet inclusion and data breaches that pose a threat to the integrity of critical systems (Radanliev et al., 2024). Ultimately, the increasing interdependency of the digital and physical domains provides the potential for cyber-attacks to inflict direct physical harm, thus making cybersecurity resilience an imperative for both civilian infrastructures and military systems.

2. MATERIALS AND METHODS

This analysis follows a comprehensive review of the current literature from academic journals, defense whitepapers and official reports. The emphasis of the study falls on articles focused on Cyber-Physical Systems (CPS), robotics security, autonomous vehicles, and critical infrastructure. Government and industry analyses were also examined. A conceptual methodology was employed, including the synthesis of threat modeling frameworks and comparative analyses to categorize and identify risk vectors. Specifically, a multidisciplinary approach was undertaken to evaluate vulnerabilities in system architectures of networked sensors, AI models, and control units, to determine how an adversary might exploit them. Threat modeling techniques and taxonomy frameworks were utilized to identify vulnerabilities and determine possible types of attacks. The resulting output included a categorization of

major cybersecurity risk classes for AI-physical systems, supported by representative cases from recent research and operations.

3. RESULTS

Systemic Vulnerabilities.

AI-physical systems include numerous sensor inputs, controllers, and actuators which together create a systemic vulnerability due to their complexity. Increased connectivity and the enhanced sophistication of CPS also increase the number of possible attack vectors across different public and private sectors (Verma et al., 2025). Examples of the types of systems that are increasingly using wireless connections to AI cloud services and other devices include robotic systems. Because many robotic systems have remote access capabilities that are not secure, they could fall victim to unauthorized control or theft of sensitive information (Tanimu & Abada, 2025). The combination of IoT devices with the existing infrastructure, form a tightly coupled network with dependencies, where a breach in one device can lead to a cascading effect across multiple devices (Radanliev et al., 2024). Studies have shown that vulnerabilities in one IoT device can be spread throughout the entire system and magnify the negative impacts of a successful attack. Additionally, reviews indicate that because there is no human oversight in CPS networks, autonomous errors or attacks may go undetected and without mitigation (Olthuis, 2025). Therefore, the inherent nature of AI-augmented physical systems, with software defined control over important physical processes, means that any cyber compromise has the potential to be disastrous in the physical world too. History provides evidence of these as malicious cyber-attacks on CPS have caused significant disruptions to both physical and digital operations, ranging from power outages to physical damage to equipment (Ghafouri et al., 2018).

Attack Vectors.

AI-physical systems have several different paths that adversaries can take to compromise them. One of the most common categories of attack vectors is sensor and perception-based attacks. For example, adversaries can spoof or jam the sensors that feed data into AI controllers. Autonomous vehicles rely on Light Detection and Raging (LiDAR), a radar, and camera inputs to navigate. Researchers have demonstrated that spoofing or injecting false data into these input sources can alter the vehicle's situational awareness, or even cause the vehicle to steer or brake improperly (George et al., 2025). Similar to autonomous vehicles, robots typically require remote access capabilities for various reasons. If the remote access interface is insecure, or weakly authenticated, then an adversary can assume control of the robot or inject malicious code (Tanimu & Abada, 2025). Network and communication-based attacks are another path. DDoS attacks, man-in-the-middle type intrusions, and protocol exploitation can all be used to disrupt communication between individual components within an AI-physical system. Attacks on the network can prevent the delivery of critical updates or sensor data necessary for proper functioning of the AI-physical system. For example, a GPS or V2X (vehicle-to-everything) communication disruption to an autonomous vehicle would essentially render it blind (George et al., 2025). Industrial control systems, such as smart grid components, have already been compromised via network attacks resulting in manipulation of control signals and ultimately leading to a loss of electrical power (Bouslimani et al., 2025).

Another way to consider an attack vector is through the lens of AI model and software attacks. The addition of AI to a system brings with it the same known machine-learning based vulnerabilities that exist in all machine-learning systems. An adversary can poison a training dataset or a model through the supply-chain or insider type threats, to result in incorrect behavior from the deployed system (Deloitte, 2025). Adversarial examples, as specifically crafted inputs that intentionally deceive an AI's perception system, are a tangible risk in physical contexts. In this context, a slight alteration to a traffic stop sign can cause an autonomous vehicle's vision system to misclassify it. Additionally, software exploits against controllers and operating systems are still viable in the current cybersecurity environment. Studies have demonstrated that robotic systems are vulnerable to the same malware families as PCs. Supply-chain based attacks are particularly concerning as well. Third party libraries, pre-trained models, or cloud services can be exploited as hidden backdoors or vulnerabilities in the AI model (Deloitte, 2025). Direct physical sabotage or insider threats should also be included in this list as an adversary may physically damage a device by installing a rogue module or damaging a sensor to compromise its function.

AI Unpredictability and Adversarial Risks.

The behavior of AI driven systems can be unpredictable, making security assessments challenging. Traditional rule-based machines make predictable decisions whereas AI models learn from data and therefore the results of an AI decision should be non-intuitive. The opacity of AI decision-making processes makes it difficult to predict how a system would react to either an attack or an error. AI models also possess unique vulnerabilities. For example, adversarial machine learning allows for attacks designed to be stealthy through modifying sensor data in ways that the system cannot identify as an attack, but still producing harmful control outputs (Ghafouri et. al., 2018). Stealthy attacks that corrupt sensor readings can produce "enormous physical damage" and evade detection by anomaly

detection algorithms (Ghafouri et al., 2018). Defense communities find the unpredictability associated with AI driven systems alarming as well. Autonomous weapons or decision support systems may react in unexpected ways when an AI model is corrupted or provided with unfamiliar stimuli. Additionally, the autonomous nature of AI systems increases the risk of miscalculation and unintended escalation. A recent policy review identified the risks associated with AI enabled rapid decision loops in military systems, stating that these systems may produce responses that human operators cannot anticipate, increasing the likelihood of accidental conflict (UNIDIR, 2025). The machine-learning layer adds an additional degree of uncertainty and an additional attack vector, through poisoning or evasion, to cyber-physical risk.

Cyber-Physical Attack Cases.

Documented cases of cyber-physical attacks demonstrate these risks in action. The 2010 Stuxnet attack was the first to demonstrate a cyber intrusion resulting in physical destruction. Malware installed on an industrial control network altered sensor data and motor speed in an Iranian nuclear facility, yet the operators were viewing normal readings. This demonstrates that altering the inputs to the controllers of a system can drive the system into unsafe operational states. More recently, researchers have discovered vulnerabilities in autonomous vehicles. Experiments demonstrated that providing an autonomous vehicle with spoofed GPS signals or replaying a previously recorded LiDAR feed can confuse or lock up the vehicle, placing the passengers at risk (Artusy, 2025). Research has also demonstrated that hijacking drones or manipulator arms through remote exploits can result in immediate physical harm. In the past year, distributed IoT botnets, in that case the botnet known as Mirai, used poorly secured smart devices to launch massive DDoS attacks that disrupted online services. Although Mirai was not a direct AI-based attack, it demonstrates how a large-scale number of ubiquitous embedded devices can be commandeered en-mass. In each of the cases mentioned above, the link between a cyber breach and a physical consequence is evident: a breach of software, sensors, or communications resulted in a real-world outcome. From stopping electrical power flow to taking control of vehicles it is evident that the integration of AI and physical systems enhances cyber threat magnitude.

4. DISCUSSIONS

The above discoveries have significant implications for both defense and strategic security. Firstly, protection of critical infrastructure must be transformed. A growing number of energy grids, transportation networks, and manufacturing sites are expected to have a degree of control through AI-driven management systems. Therefore, the cybersecurity of those systems inherently encompasses physical safety. Attacks on such systems could impose catastrophic consequences, like large-scale power outages, transportation collapse, or possibly loss of life, which would be in addition to the expected financial and business disruptions created by a typical cyber-attack. However, the weaknesses discovered clearly demonstrate that current IT security strategies are insufficient. Operators of critical infrastructure should implement layered defenses that include real-time observation of physical behavior, redundant sensing, and fail-safe defaults. Additionally, a "disproportionate focus on defense" has been seen in studies about CPS sectors, and there appears to be a lack of understanding of the complete attack surface (Olthuis, 2025).

This indicates a need for interdisciplinary research and integrated security frameworks for the private and public sectors. In terms of national defense, the stakes are much higher. Military services are using AI-enabled autonomous platforms integrated in unmanned aerial vehicles, unmanned ground vehicles, decision support systems, to improve situational awareness and force multiplication. However, like the other cyber-physical systems, these are susceptible to the same cyber-physical attacks. A review of autonomous vehicles in military contexts points out that as militaries continue to rely on autonomous, connected systems, "ensuring cyber and operational resiliency becomes a national security imperative" (Artusy, 2025). In a combat environment, an enemy could try to spoof or jam sensors belonging to friendly forces or insert malicious input into those systems, potentially leading to failure of friendly robots or even turning them against their own side. The unpredictable behavior of AI presents the challenge of hostile spoofing potentially creating unforeseen ways to trick systems. As noted by RAND, U.S. defense decision-makers need to act quickly to adjust to the implications of AI's cybersecurity issues, lest they find themselves without a plan when faced with AI-related cybersecurity issues (Danzig, 2025).

Thus, defense planners must include robust AI security, such as hardened ML models, secure supply chains, and real-time anomaly detection in military modernization efforts. Strategically and internationally, AI-physical convergence provides new escalation dynamics. Cyberattacks that cause physical damage to critical systems may be viewed as acts of aggression. However, attributing the source of such attacks is notoriously difficult in cyberspace. One survey states that sophisticated adversaries use supply chain manipulation and AI-based obfuscation to make it "critical and elusive" to determine who is responsible for cyberattacks (Prasad et al., 2025). Failure to determine the origin of an attack can undermine deterrence and trust. For instance, if a grid disruption caused by AI-controlled

drones cannot be attributed to a nation-state and the nation-state does not deny responsibility, it could lead to misinterpreted responses. The increased likelihood of rapid, autonomous responses increases the risk of unintended escalation as AI-enabled systems may respond to a cyber incident before humans can intervene (UNIDIR, 2025), increasing the possibility of misinterpretation of intent and unintended escalation. Therefore, policymakers must examine how AI-driven warfare blurs the line between digital and kinetic conflict. Emerging AI and ML tools present opportunities to improve the accuracy of determining the source of attacks (Prasad et al., 2025). However, these solutions are still developing.

AI-physical risks far surpass traditional cybersecurity in both scope and form. Traditional cybersecurity typically addresses protecting data and business IT systems. Conversely, AI-physical security must address safety, reliability, and the ability to withstand physical harm. Controllers in CPS make decisions based on data from sensors and as demonstrated, tampering with that data can lead a system to "an unsafe condition," a phenomenon that is not applicable in purely digital environments.

Therefore, models of defense must be updated in addition to confidentiality and integrity, availability and real-time correctness as safety concerns. This necessitates the development of cross-disciplinary standards that combine cybersecurity principles with control-theoretic safety margins. Research in CPS security, such as developing anomaly detection methods for sensor feeds (Ghafouri et al., 2018) and secure hardware modules (George, 2025) illustrates how AI-physical systems require multi-faceted protections. The convergence of AI and the physical world require reevaluation of assessing risk as threats are no longer hypothetical breaches of data, but potential life-threatening manipulations of machines and infrastructure.

5. CONCLUSIONS

Physical-AI convergence offers advantages as well as adversaries with new opportunities to threaten both national defense and global stability through a number of emerging cybersecurity threats. The research has identified five primary categories of these threats from sensor spoofing and network attacks, to sensor attacks and adversarial machine learning, all of which contribute to a large increase in the attack surface of current systems. Evidence indicates that an adversary can use AI-enhanced autonomy to create unpredictable conversions of digital assaults into physically destructive damage. Autonomous vehicle, robotic, and grid attacks are examples of how an AI-based system can be used to create devastating effects. As the military and many industries increasingly begin to deploy AI-based systems in their respective environments, the possibility for rapid escalation and misattribution increases. To protect against the risks associated with physical-AI convergence, there must be coordinated efforts across multiple domains including threat modeling, secure design of AI systems by default, collaborative work among cross-functional teams and organizations, and adaptive policy development. In addition to coordinating the necessary responses, national security and industry must recognize that cybersecurity is no longer separate from physical security. It is clear that if proactive measures are not taken to protect the physical-AI based systems from adversaries' attempts to disrupt them and cause potentially catastrophic damage, they will utilize the physical-AI convergence as a mechanism to create disruption and chaos.

REFERENCES

Artusy, D. (2025). Autonomous Vehicles in Critical Infrastructure: Technologies, Vulnerabilities, and Implications. The Cyber Defense Review, 10(2), 69–78. DOI: 10.55682/cdr/m9d9-5ee5.

Bouslimani, M., Benbouzid-Si Tayeb, F., Amirat, Y., & Benbouzid, M. (2025). Cyber-Physical Security in Smart Grids: A Comprehensive Guide to Key Research Areas, Threats, and Countermeasures. Applied Sciences, 15(23), 12367. DOI: 10.3390/app152312367.

Danzig, R. (2025). Artificial Intelligence, Cybersecurity, and National Security: The Fierce Urgency of Now. RAND Corporation Expert Insights. DOI: 10.7249/PEA4079-1.

Deloitte Insights, 2025, Tech Trends 2026, Deloitte, https://mkto.deloitte.com/rs/712-CNF-326/images/DI_Tech-trends-2026.pdf

George, D., Pavithra, S., & Das, J. (2025). Cyber-resilient autonomous vehicles: Securing networks and enhancing decision-making with next-gen security measures. Results in Engineering, 28, 107179. DOI: 10.1016/j.rineng.2025.107179.

Ghafouri, A., Vorobeychik, Y., & Koutsoukos, X. (2018). Adversarial Regression for Detecting Attacks in Cyber-Physical Systems. In Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI 2018) (pp. 3769–3775). DOI: 10.24963/ijcai.2018/524.

Olthuis, J. J., Sciancalepore, S., & Zannone, N. (2025). Cyberattacks and defenses for Autonomous Navigation Systems: A systematic literature review. Computer Networks, 267, 111331. DOI: 10.1016/j.comnet.2025.111331.

Prasad, N., Diro, A., Warren, M., & Fernando, M. (2025). A survey of cyber threat attribution: Challenges, techniques, and future directions. Computers & Security, 157, 104606. DOI: 10.1016/j.cose.2025.104606.

Radanliev, P., De Roure, D., Maple, C., Nurse, J. R., Nicolescu, R., & Ani, U. (2024). AI security and cyber risk in IoT systems. Frontiers in Big Data, 7, 1402745. DOI: 10.3389/fdata.2024.1402745.

Tanimu, J. A., & Abada, W. (2024). Addressing cybersecurity challenges in robotics: A comprehensive overview. Cyber Security and Applications, 3, 100074. DOI: 10.1016/j.csa.2024.100074.

UNIDIR, 2025, UNIDIR's Security and Technology Programme, Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security: An Evidence-Based Road Map for Future Policy Action, (Geneva: UNIDIR, 2025). https://unidir.org/wp-content/uploads/2025/07/UNIDIR_AI_military_domain_implications_international_peace_security.pdf

Verma, N., Kumar, N., Sheikh, Z. A., Koul, N., & Ashish, A. (2025). Machine Learning for the Cybersecurity of Robotic Cyber-Physical Systems: A Review. Procedia Computer Science, 259, 1817–1826. DOI: 10.1016/j.procs.2025.04.137.