AI Warfare

Dong Wang, Tokunbo Makanju and Yunlong Shao
New York Institute of Technology, Vancouver, Canada

dwang52@nyit.edu
amakanju@nyit.edu
yshao12@nyit.edu

Abstract: Artificial intelligence (AI) is rapidly transforming military strategy, yet the idea of AI warfare remains loosely defined. This paper introduces a structured framework that views AI warfare as a form of strategic competition centred on algorithms, training datasets, AI-as-a-Service (AIaaS), and AI-generated content (AIGC). These technological instruments are used to influence adversaries' decision-making and cognitive processes, while simultaneously undermining their defensive and command capacities. The analysis begins with the vulnerabilities of layered architecture of AI ecosystem—infrastructure, algorithm-data, foundation models, and applications—each containing vulnerabilities that can cascade through the system. AI systems also carry inherent limitations, such as unclear problem contexts, lack of creativity or emotional understanding, disembodiment, and weak collaboration between models. These traits distinguish AI warfare from conventional cyber operations and highlight why traditional cybersecurity measures often fall short. The proposed framework operates across three interlinked components: technological infrastructure—semiconductors, high-performance computing, and high-speed networks—sets the capability and flexibility of AI systems, while an opaque algorithm-data core becomes a frontline where adversarial manipulations threaten reliability. At the operational level, AI is deployed through AIGC campaigns and AIaaS platforms, with feedback loops among these components driving continuous escalation and adaptation in AI warfare. Five critical focus areas emerge: (1) proprietary algorithms functioning as opaque "black boxes," examined through a new Wargaming Design (WD) model using iterative red-blue team challenges; (2) vulnerable training datasets, tested for integrity via WD model's consistency checks; (3) AIaaS platforms demanding security-by-design safeguards like role-based access and auditing; (4) AIGC as both an effective offensive tool and a potential source of self-sabotage; and (5) quantitative audience analysis to assess cognitive and behavioural outcomes. Defensive strategies include blockchain-based distributed architectures for resilience, secure interaction design, AI-native firewalls, interdisciplinary user/audience research drawing on sociology and psychology, and global governance structures akin to non-proliferation treaties. Through this approach, the paper explains why conventional defences fail against AI-specific threats. Future work will experimentally apply the WD model to existing AI systems, measuring its effectiveness in identifying algorithmic weaknesses and improving AI security. Ultimately, policymakers must balance the power of AI with the preservation of strategic stability and human agency.

Keywords: AI warfare, Cyber warfare, Cybersecurity, AIaaS, AIGC, Algorithm defence

1. Structural Vulnerabilities of AI Ecosystem

As discussed by Blunden and Cheung (2014), Dyer-Witheford and Matviyenko (2019), Mehan (2014), Carvalho and da Silva (2006), the artificial intelligence (AI) ecosystem can be understood as a set of interdependent layers that together shape how AI systems are developed and deployed.

At its foundation is the infrastructure layer, which includes semiconductors, computing power centres, and high-speed networks; this layer determines computational capacity, scalability, and even geopolitical influence over AI competitions. Built on top of this is the algorithm-data co-evolution layer, where algorithms and training datasets interact continuously to improve model performance. The foundation model layer then provides large, pre-trained models that serve as adaptable bases, allowing developers to reuse general capabilities across tasks while reducing time and cost. These models are further refined in the specialization layer, where fine-tuning, domain-specific data, and retrieval systems produce both general-purpose products and highly specialized services, often applied by sensitive sectors such as government, military and finance. The application layer represents the point of deployment, where AI is integrated into real-world environments and will generate economic, social, or strategic effects.

Across all these layers, governance and security concerns act as a cross-cutting force, shaping how data is accessed, how services are deployed, and how infrastructure is controlled, while also influencing risk, accountability, and competitive dynamics.

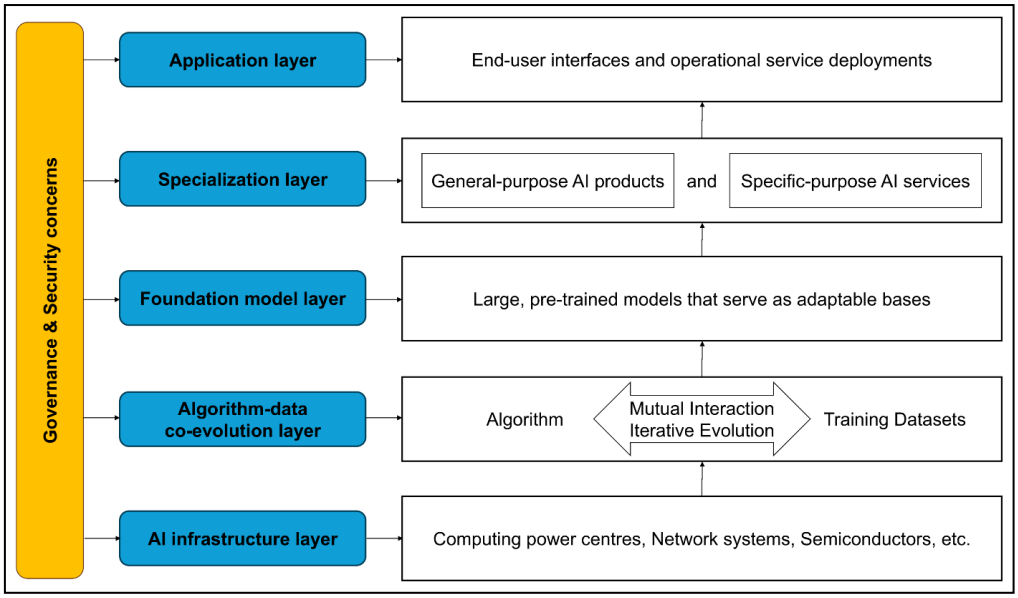

Figure 1: The structural layers of AI ecosystem

However, this stratified architecture creates cascading vulnerabilities where compromise at lower levels propagates upward, undermining algorithmic integrity, model reliability, and operational security.

1.1 Infrastructure Vulnerability

The biggest vulnerability in the AI infrastructure layer is geopolitics, particularly supply chain disruptions from export controls and the risk of kinetic attacks on data centres. U.S. restrictions on advanced AI chips since 2023 have created global dependencies, limiting access to critical hardware like NVIDIA H100 GPUs and amplifying competition with China. Recent 2026 incident, Iranian drone strikes on Gulf data centres, underscore the physical fragility of these facilities amid rising tensions. Technical challenges persist, but geopolitical factors pose the most acute threat to AI capability and scalability.

1.2 Vulnerability of Algorithm and Data

The main vulnerabilities in the algorithm-data co-evolution layer are the black-box nature of algorithms and opaque training datasets quality. Data poisoning attacks pose the greatest risk, as attackers insert malicious samples into datasets, causing models to learn flawed patterns that persist after deployment. Souly, A. et al. (2025) found that just 100-500 poisoned examples can compromise large-scale models, especially in federated learning where detection is difficult. These attacks exploit the iterative algorithm-data feedback loop, creating persistent threats and risks.

1.3 Foundation Model's Vulnerability

The typical vulnerabilities in the foundation model layer are prompt injection and jailbreaking. Attackers craft inputs to override safety alignments, forcing models to produce harmful outputs like malware code or misinformation. Anthropic (2026) detected industrial-scale distillation attacks by DeepSeek, Moonshot AI, and MiniMax, which created ~24,000 fake accounts to systematically extract Claude's reasoning and tool-use capabilities via structured queries.

This threat is critical because foundation models underpin specialized applications across sectors, amplifying systemic risks when guardrails fail. Industry analysis shows these models form chokepoints where a single exploit cascades downstream, unlike narrower attacks on specific adaptations.

1.4 Vulnerability at Specialization Layer

In the specialization layer, AI-generated content (AIGC) and AI-as-a-Service (AIaaS) form key risk surfaces for general-purpose products and specific-purpose services. AIGC introduces vulnerabilities like insecure code generation, Veracode (2025) found that AI-generated code introduced exploitable security vulnerabilities in 45% of tested development tasks across more than 100 large language models, expanding attack surfaces at scale. AIaaS exposes risks through API weaknesses and prompt injections; Cycode (2026) highlights CVE-2025-53773,

where hidden injections in GitHub Copilot enabled remote code execution (CVSS 9.6). These amplify in specialized military or financial systems, where flawed outputs or hijacked workflows cascade to critical failures.

1.5 Application Layer Vulnerability

In the application layer, cognitive manipulation emerges as the primary vulnerability for end-users, where AI subtly influences opinions and behaviours through personalized disinformation. In Ireland's 2025 presidential election, AI-generated deepfakes falsely depicted a leading candidate withdrawing from the race and the election being cancelled, prompting authorities and platforms to remove the content as election-related disinformation. Canadian Centre for Cyber Security (2024) warned that AI-driven narratives erode trust, with chatbots echoing manipulated content from poisoned sources. These risks extend beyond technical flaws, threatening democratic integrity as users increasingly rely on AI for information.

2. Weaponized Vulnerabilities: Why is the Analysis of AI Warfare Needed?

AI is rapidly emerging as a transformative force across economic, social, and security domains. Its speed, scalability, and convenience encourage widespread adoption, often leading users to trade autonomy and control for efficiency. At the same time, it is important to recognize the substantial gap between narrow AI systems and human-level general intelligence. Several structural constraints limit the effectiveness of current AI in military contexts, including the inability to handle undefined problem spaces, lack of genuine insight and creativity, dependence on limited sensory inputs, absence of emotional understanding, and weak cross-system integration. These constraints affect how AI can be deployed, trusted, and countered in operational settings.

In addition to these fundamental gaps, the previously discussed vulnerabilities of AI may also increase the potential for AI weaponization. This can occur through algorithms, training datasets, AIaaS, and AIGC, thereby introducing new attack surfaces and reshaping strategic interactions in conflict environments. These developments demand systematic analysis and the advancement of theoretical frameworks tailored to AI from a warfare perspective.

2.1 Algorithm Weaponization

Algorithm weaponization denotes the intentional design or manipulation of algorithms to produce adverse cognitive and operational effects in conflict environments. In AI warfare context, Lange, L. (2024) described an "algorithmic cognitive warfare" framework, in which AI-driven recommendation systems and social-media algorithms are repurposed to microtarget users with tailored narratives, aiming to polarize public opinion, erode trust in institutions, and fragment collective decision-making under information overload. In particular, Chinese-linked military and security discourse explicitly frames these algorithmic tools as instruments for influencing foreign populations along political, ethnic, and ideological lines. In practice, this can resemble algorithm-driven bursts of misleading content timed to elections or crises, although the specific mechanisms are still being mapped in the literature.

2.2 Data Weaponization

Strategic poisoning, distortion, or exploitation of datasets used to train or operate AI systems is a powerful weapon in warfare. In AI-driven conflicts, adversaries manipulate datasets to degrade reliability, mislead targeting systems, or create synthetic "ground truths" that align with their narrative. Liebgold (2025) researched on U.S. military AI scenarios and highlighted how data poisoning can be used as a covert cyber capability to undermine adversary intelligence, surveillance, and reconnaissance systems without triggering conventional warfare thresholds This case demonstrates that control over data integrity is central to AI warfare, because corrupted or biased inputs systematically warp the outputs of even technically robust models.

2.3 AIGC Weaponization

AIGC weaponization designates the deployment of generative models to produce synthetic audio, video, and text for cognitive-targeting purposes. During the 2026 Israel–Iran hostilities, Brookings Institution (2026) observed that Iranian asymmetric campaigns employed AIGC deepfakes to simulate military leaders' pronouncements and false battlefield footage, which were disseminated through semi-official channels to amplify doubt and reduce coalition cohesion. Guess et al. (2023) conducted empirical studies of synthetic media in information operations show that AIGC disinformation can significantly increase perceived credibility among test audiences, especially when aligned with local cultural or linguistic cues. These dynamics transform content creation into a scalable, adaptive layer of AI warfare rather than a one-off propaganda tactic.

2.4 AIaaS Weaponization

AIaaS weaponization captures how AI services and APIs enable adversaries or non-state actors to execute sophisticated information operations without in-house expertise. In AI warfare, adversaries rent AI services for text generation, voice synthesis, and behavioural prediction, then integrate them into automated attack chains.

Nobles, C. (2024) studied on the weaponization of AI in cybersecurity highlights that AIaaS platforms lower the skill barrier for creating malware-generation tools, sentiment-analysis pipelines, and fake-network-traffic simulators, thereby enabling rapid experimentation and adaptation of offensive toolkits. Consequently, AIaaS weaponization reframes AI warfare as an ecosystem-level problem, where the same commercial platforms used for benign automation become vectors for scalable cognitive attack.

In conclusion, given these multiple, mutually reinforcing layers of AI weaponization, a systematic analysis of AI warfare is essential to anticipate, mitigate, and regulate emerging threats. Without tailored theoretical and operational frameworks, states and other entities risk adopting AI-enabled capabilities reactively, while adversaries exploit structural gaps and cognitive vulnerabilities at scale. Understanding AI warfare is therefore not merely an academic exercise but a strategic necessity.

3. Conceptual Framework of AI Warfare

This section builds on the earlier analysis of the AI ecosystem, its vulnerabilities, and how those weaknesses can be weaponized in warfare, including real-world examples. It will focus on a core question: what is AI warfare? First, we review and evaluate existing definitions. Then, we explain the key components that make up AI warfare and propose a practical definition that can be used for analysis. Finally, we apply this definition to a case study to show how it helps interpret a real-world warfare scenario.

3.1 Literature Review

Since 2020, academic and policy discourse on AI warfare has evolved from a narrow, technology-centric view toward a broader, systemic understanding of cognition, information, and operational speed. Early work treated AI warfare as AI-augmented conventional warfare, focusing on target-recognition systems, predictive logistics, and autonomous platforms.

Around 2022–2023, scholars such as Rickli and Mantellassi (2023) began explicitly framing AI warfare through military uses and international security implications, linking AI-driven capabilities to arms-control, escalation risk, and alliance-level strategic competition. Johnson (2022)'s work on the "AI commander problem" added a strong ethical and psychological dimension, redefining AI warfare as a domain where human–machine interaction creates dilemmas in responsibility, control, and trust.

Around 2024, policy-oriented scholarship emphasized AI's role in information warfare and influence operations, with Hunter et al. (2024) showing how major powers deploy AI for narrative-level contestation, deepfakes, and algorithmic targeting, thereby shifting definitions toward cognitive and information-centric warfare. Mallick (2024) and Nobles (2024) treated AI warfare as a meta-domain that reshapes military capability, doctrine, and national-security thinking, encompassing kinetic autonomy, cyber-information operations, and AI-driven decision-making at scale.

From 2025 onward, conceptualizations have become more systemic, with studies further sharpening definitions of AI warfare by emphasizing AI's role across kinetic, cognitive, and systemic-level conflict. Sticher (2025) framed AI warfare as a domain that reshapes not only weapons and operations but also alliance structures, risk-assessment, and governance. Gallitelli (2025) highlighted AI-driven military decision-making as a core pillar of AI-enabled warfare, underscoring speed, autonomy, and human–machine interaction as defining features of the emerging paradigm.

Table 1: Existing approaches to defining AI warfare

Existing approaches to defining AI warfare		
Definitional basis	Core claim	Main limitation
Technical enablement	AI enhances military effectiveness	Treats AI as tool, obscures strategy
Warfare paradigm	New form of warfare after mechanization, informatization	Underestimates transformation of military organization

Existing approaches to defining AI warfare		
Algorithm-centric	Competition for algorithmic superiority and cognitive dominance	Ignores data dependencies and bias issues
Decision-making support	AI accelerates OODA (Observe-Orient-Decide-Act) cycle	Ignores social/cognitive domains
Hybrid operations	Integration of cyber, kinetic, and information operations under AI coordination	Underspecifies AI's unique contribution

While the literature agrees that AI accelerates military processes and enhances operational effectiveness, it remains divided over whether the concept should be defined narrowly in relation to combat or broadly as a cross-domain transformation of warfare. A key gap in the literature is the lack of a unified, interoperable definition that explicitly integrates kinetic, cognitive, and information-operations dimensions, with most authors focusing on only one or two of these layers.

3.2 Components of Understanding AI Warfare

Understanding AI warfare can be structured around three interrelated components: the technological infrastructure that enables AI systems, the algorithm-data core that generates their behaviour, and the operational practices that deploy AI in conflict. Together they clarify how AI reshapes the speed, scale, and cognitive dimensions of contemporary warfare.

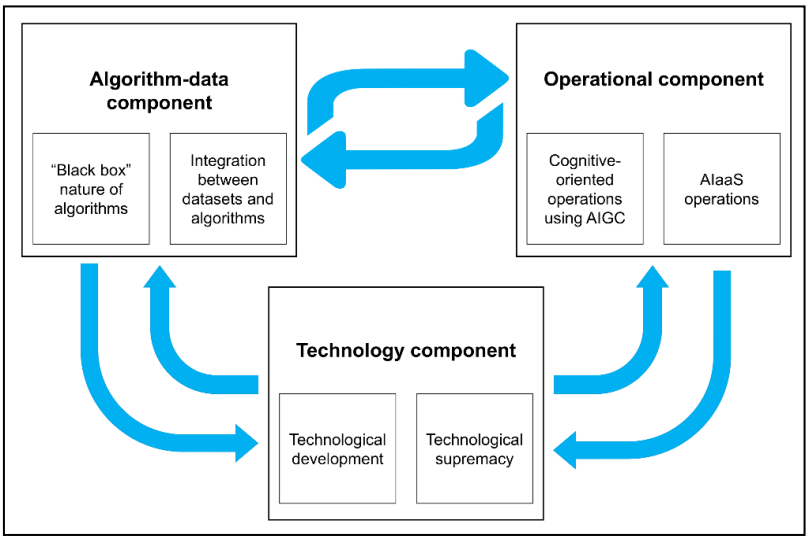

Figure 2: The interdependent components of understanding AI warfare

The technology component rests on semiconductors, high-performance computing, and high-speed networks, which do not decide battles but shape the speed, scale, and adaptability of military operations.

The algorithm-data component reflects the opaque, data-driven logic of AI, where tightly coupled models and datasets make adversarial manipulations difficult to detect and correct. Data-poisoning, model-substitution, and stealthy prompt-based exploitation blur the boundary between normal operation and attack, so the algorithm-data domain has become a critical frontline in AI warfare, where control over data quality and model integrity directly affects system reliability.

The operational component captures how AI is used in practice, mainly through cognitive-oriented and service-oriented operations. Cognitive-oriented operations use AIGC to influence perception, belief, and decision-making, including disinformation and deepfakes. Service-oriented operations apply AIaaS to execution and coordination, such as autonomous systems, swarm tactics, and real-time decision support.

The three components are interdependent. Technological infrastructure provides the foundation that enables algorithmic development, while algorithm–data interactions shape the effectiveness and reliability of operational applications. In turn, operational demands drive the evolution of both algorithms and underlying technologies. This dynamic interplay means that weaknesses at any single component can propagate across the

entire system. Therefore, understanding AI warfare requires a holistic perspective that considers how these levels interact rather than analysing them in isolation.

3.3 Definition of AI Warfare

Based on the preceding discussion, we propose a new conceptual definition of AI warfare：

AI warfare is the conduct of strategic competition through the deliberate manipulation of algorithms, datasets, AIGC or AIaaS to influence the decision-making processes, command-and-control functions, operational behaviours, and cognitive states of target populations (including strategic elites, military organizations, opinion leaders and general publics), thereby advancing the attacker's strategic and tactical objectives while potentially degrading the defender's strategic judgment and operational effectiveness.

This definition emphasizes the intentional character of AI warfare, the multidimensional nature of its effects, and the critical role of human decision-making as both means and target.

3.4 Case Study: China's Spam AIGC Campaign on X.com

In late January 2026, Nikita Bier, X.com's head of product, publicly accused China of deploying 5-10 million fake accounts on the platform. He claimed these accounts flooded Chinese-language searches with pornographic spam during periods of China’s political unrest. This tactic buried dissenting voices and made real-time information nearly impossible to access. The case aligns directly with the proposed AI warfare definition and its analysing framework:

- The technology component formed the backbone of this operation. Attackers relied on scalable computing infrastructure to create and manage millions of accounts rapidly. Cloud services or botnets provided the necessary speed and volume. Such resources ensured the spam campaign could operate undetected at scale. China's heavy investments in computing power made this feasible with minimal cost.
- In the algorithm-data component, AIGC played a key role. Algorithms trained on X.com's user patterns produced realistic spam contents that evaded detection filters. Datasets hoarded from prior activity allowed accounts to mimic authentic behaviour. The black-box nature of these AI systems hid manipulations from platform defences. Bier noted pre-existing account pools, built before X.com tightened signup rules, which added resilience.
- Operationally, the campaign focused on cognitive effects. Spam distorted perceptions among Chinese users, elites, and the general public seeking news on unrest. It overwhelmed searches for sensitive terms, eroding collective judgment at critical moments. Coordination hints at AIaaS too, as bots timed floods to outpace platform algorithms. This advanced Beijing's goals of narrative control and dissent suppression.

These components proved interdependent. Technological scale enabled algorithmic stealth, while data sophistication drove operational impact. Together, they warped user decisions and eroded trust in X.com. Attackers achieved strategic aims by manipulating algorithms, datasets, and AIGC to influence cognition and operations, just as the definition predicts. This degraded defenders' effectiveness in real time.

4. Critical Research Questions on AI Warfare

Effective analysis of concrete AI warfare scenarios requires a systematic framework. We propose five interconnected focus areas that collectively enable comprehensive assessment of AI warfare cases.

4.1 How to Disclose the Black box of Algorithms?

Algorithm is a systematic problem-solving methodological procedure for analysing, modelling, and acting on complex environments. AI infrastructure provides the means to execute algorithms; AIGC and AIaaS represent their observable manifestations in the military and strategy domains.

Due to their strategic value, core algorithms of leading AI providers remain proprietary black boxes. Techniques such as model distillation and exhaustive input-output probing can infer partial behavioural patterns but cannot fully reconstruct underlying design choices.

To address this concern, Dong (2026) proposed the Wargaming Design (WD) model as a viable solution for approximate algorithm transparency.

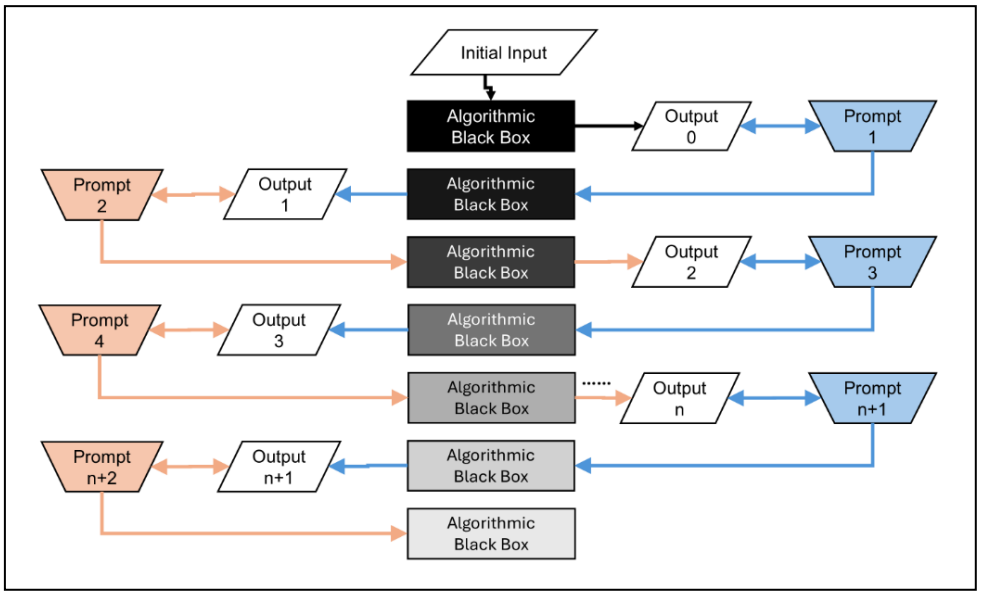

Figure 3: The revised workflow diagram of Wargaming Design model

The core insight of this model applies adversarial reasoning and game theory to algorithm interrogation:

- Blue team proposes input prompts and receives outputs from the target algorithm.
- Red team challenges these outputs through adversarial counter-arguments, alternative interpretations, and critical questioning.
- Through iterative refinement—each team responding to the other's critiques—the algorithm is pressed toward internal logical consistency.
- The critical moment: A black-box algorithm forced to reconcile contradictory outputs while maintaining internal coherence must partially expose its underlying assumptions and constraints.

Rather than achieving perfect transparency, the WD model creates pressure for logical self-consistency that reveals operational characteristics sufficient for strategic assessment. This approach recognizes that perfect transparency may be unattainable but that iterative adversarial dialogue can approximate understanding through forced coherence.

4.2 Why are Training Datasets the Primary Attack Surface?

Thinking of algorithms as the architectural blueprint, datasets provide the building materials through which algorithms learn patterns, correlations, and decision rules. These datasets—whether public, proprietary, real-world, or synthetic—are continuously updated and remain the key constraint on AI performance.

However, they also introduce critical vulnerabilities as discussed. Research shows that data poisoning can implant backdoors using only a small number of malicious samples, making such attacks highly cost-effective regardless of model size. Existing defences rely on pre-training data vetting, but this becomes infeasible at the scale of modern datasets.

We propose applying the idea of WD model for data integrity verification. Instead of exhaustive pre-screening, datasets are iteratively tested for internal consistency using the model itself, allowing unreliable or harmful inputs to be identified and excluded.

4.3 Why is AIaaS a Security-critical Architecture?

Syed et al. (2025) and Lewicki et al. (2023) state that AIaaS serves as the interface between algorithms and end users, acting as the primary channel through which external actors interact with AI systems. Current designs inherit the search-engine model of a single, minimalist input box, which reduces user friction but shifts most security responsibility onto algorithms and training datasets.

As prompt-based interaction remains the dominant communication mode, AIaaS has become a key attack surface, vulnerable to prompt injection, jailbreaking, and context manipulation. To improve resilience, specialized or military AIaaS should embed security into interface design, including role-based access, multi-step

verification workflows, explicit risk acknowledgment, and comprehensive interaction auditing. Although these measures reduce usability, the trade-off is justified in high-stakes and security-sensitive contexts.

4.4 What is the Double-edged Character of AIGC?

AIGC has a fundamentally double-edged character in AI warfare. As an offensive tool, high-quality AIGC enables scalable information operations, psychological campaigns, and ideological influence, especially when integrated into coherent long-term strategies. It can systematically shape public opinion and elite perceptions.

However, AIGC also creates defensive liabilities. Hallucinations, factual errors, and manipulative framing can mislead decision-makers and undermine credibility if exposed. Poorly executed campaigns risk backlash, amplifying negative effects.

This duality highlights a critical principle: public media literacy must keep pace with AIGC capabilities. Widespread access to powerful generative tools without sufficient critical thinking skills is both strategically counterproductive and ethically problematic. Effective defence therefore requires layered responses, including narrative resilience, content verification, and rapid psychological countermeasures.

4.5 Why Quantify AI Warfare Effects?

Quantifying AI warfare effects is essential for assessing both offensive and defensive capabilities. Measuring AI influence requires evaluating how AIaaS and AIGC shape cognitive and behavioural outcomes across different audiences, including decision-makers, operators, and the broader public. Key dimensions include shifts in perception, belief formation, decision processes, and public opinion dynamics.

A robust questionnaire framework must integrate long-term monitoring, targeted analysis, and thematic investigation to capture both persistent and situational effects. This involves tracking audience acceptance of AIaaS, their ability to identify AIGC, changes in issue-specific cognition, and levels of AI security awareness.

Such quantification provides the empirical foundation for designing, calibrating, and validating AI warfare strategies, ensuring that influence operations and defensive measures are both effective and accountable.

5. AI Warfare Defence Strategies

Beyond developing indigenous AI capabilities, states and entities must build resilient, forward-looking defence architectures to address AI-related threats:

- Distributed computing architectures. Reducing reliance on centralized data centres can mitigate single points of failure. While centralized systems maximize computational efficiency, they are vulnerable to disruption if critical nodes are compromised. Distributed architectures—potentially leveraging blockchain—offer enhanced resilience, privacy protection, and traceability, despite current limitations in scalability and coordination. In high-stakes domains such as national defence and finance, these trade-offs may be justified.
- Security-by-design. Prompt-based interaction remains the primary attack vector, yet current AI interfaces inherit the minimalist design of search engines, externalizing security risks. AI systems, especially specialized AIaaS, should embed security into interaction design through structured workflows and explicit thresholds. Although this reduces usability, it significantly strengthens system security.
- AI-native security models. Emerging threats resemble “AI viruses,” which both exploit technical vulnerabilities and manipulate user authorization and cognition. These threats blur the boundary between technical and social attack surfaces. Frameworks such as the WD model illustrate the need for AI-era defences, including advanced firewalls and adaptive security systems.
- Audience-centred research. AI influence extends beyond direct users, shaping cognition and public opinion through pervasive AIGC exposure. Defence strategies must therefore incorporate insights from communication studies, sociology, and psychology to monitor and understand these effects.
- International governance. The weaponization of AI raises urgent legal and ethical challenges, including technological supremacy, privacy, sovereignty, and autonomous weapons control. International frameworks analogous to arms control treaties are needed, alongside long-term national strategies to avoid technological dependence.

6. In Conclusion

AI represents a transformative technology with profound implications for military competition, strategic stability, and international security. Unlike cyber warfare, which operates through traditional networks and exploits system vulnerabilities, AI Warfare operates through algorithms, datasets, and human cognition itself.

In this paper, we have proposed a multidimensional definition of AI warfare emphasizing the integration of technical, operational, and cognitive dimensions. This framework enables systematic analysis of how AI can be weaponized to achieve strategic objectives.

In this study, we identified the key questions that require attention in AI warfare, and we also provide some macro-level solutions for defending against AI warfare. Future research should extend the solutional framework to specific military domains (intelligence analysis, targeting, command-and-control, force protection), examine case studies of AI warfare (actual and hypothetical), and develop defensive doctrine appropriate to military organizations.

The challenge before strategists and policymakers is significant: to harness AI's strategic potential while managing its risks to international stability and human autonomy. This challenge requires sustained investment in research, development of robust defensive capabilities, international coordination, and wisdom to recognize both the power and the limits of AI as a strategic instrument.

Ethics Declaration: This research did not require ethical approval.

AI Declaration: No AI tools were used in the development of this paper.

References

Anthropic (2026) *Detecting and preventing distillation attacks*. Available at: https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks (Accessed: 3 May 2026).

Ashraf, C. (2021) 'Defining cyberwar: towards a definitional framework', *Defense & Security Analysis,* 37(3), pp.274-294.

Bier, N. (2026) *They have a pool of 5-10 million accounts that post about Chinese politics...* X.com, 29 January. Available at: https://x.com/nikitabier/status/2017136051782103050 (Accessed: 2 May 2026).

Blunden, B. and Cheung, V. (2014) *Behold a pale farce: Cyberwar, threat inflation, & the malware industrial complex*. Las Vegas, NV: Trine Day.

Boer, L.J. (2021) *International law as we know it: Cyberwar discourse and the Construction of knowledge in International Legal Scholarship*. Cambridge: Cambridge University Press.

Brookings Institution (2026) *Generative AI as a weapon of war in Iran*. Washington, DC: Brookings Institution. Available at: https://www.brookings.edu/articles/generative-ai-as-a-weapon-of-war-in-iran/ (Accessed: 3 May 2026).

Canadian Centre for Cyber Security (2024) *National Cyber Threat Assessment 2025-2026*. Available at: https://www.cyber.gc.ca/en/guidance/national-cyber-threat-assessment-2025-2026 (Accessed: 28 April 2026).

Carvalho, F.D. and da Silva, E.M. eds. (2006) *CyberWar-Netwar: Security in the information age.* Vol. 4. Amsterdam: IOS Press.

Cycode (2026) *Top AI Security Vulnerabilities to Watch out for in 2026*, 30 March. Available at: https://cycode.com/blog/ai-security-vulnerabilities/ (Accessed: 28 April 2026).

Davis, E.V.W. (2021) *Shadow Warfare: Cyberwar Policy in the United States, Russia and China*. London: Bloomsbury Publishing USA.

Dong, W. (2026) 'Wargaming design for cyber warfare', in *Proceedings of the 21st International Conference on Cyber Warfare and Security (ICCWS 2026)*, Wilmington NC: Academic Conferences and Publishing International Limited, pp. 676–684.

Dyer-Witheford, N. and Matviyenko, S. (2019) *Cyberwar and revolution: Digital subterfuge in global capitalism*. Minneapolis, MN: University of Minnesota Press.

Gallitelli, M.V. (2026) 'Artificial Intelligence-Enabled Military Decision-Making Process: The Forgotten Lessons on the Nature of War', *Journal of Advanced Military Studies*, 16(2), pp. 99–132. Available at: https://research-ebsco-com.nyit.idm.oclc.org/linkprocessor/plink?id=1c3ccec3-7cb3-33e8-be20-4f4b29245a8c (Accessed: 3 May 2026).

Guess, A., Nagler, J. and Tucker, J. (2023) 'Less than you think: prevalence, perceptions, and effects of fake news in the US', *Journal of Communication*, 73(4), pp. 243–261.

Hunter, L.Y., Albert, C.D., Rutland, J., Topping, K. and Hennigan, C. (2024) 'Artificial intelligence and information warfare in major power states: how the US, China, and Russia are using artificial intelligence in their information warfare and influence operations', *Defence & Security Analysis*, 40(2), pp. 235–269.

Hughes, D. and Colarik, A. (2017) 'The hierarchy of cyber war definitions', in *Pacific-Asia workshop on intelligence and security informatics.* Cham: Springer International Publishing, pp. 15–33.

Johnson, J. (2022) 'The AI commander problem: ethical, political, and psychological dilemmas of human–machine interactions in AI-enabled warfare', *Journal of Military Ethics*, 21(3–4), pp. 246–271.

Lange, L. (2024) *Algorithmic cognitive warfare: the next frontier in China's quest for global influence*. Washington, DC: Special Competitive Studies Project (SCSP).

Lewicki, K., Lee, M.S.A., Cobbe, J. and Singh, J. (2023) 'Out of context: investigating the bias and fairness concerns of "artificial intelligence as a service"', in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. New York: ACM, pp. 1–17.

Liebgold, J. (2025) *Data poisoning as a covert weapon: securing U.S. military superiority in AI-driven warfare*. West Point, NY: Lieber Institute.

Mallick, P.K. (2024) 'Artificial intelligence, national security and the future of warfare', in *Artificial intelligence, ethics and the future of warfare*. New Delhi: Routledge India, pp. 30–70.

Mehan, J. (2014) *Cyberwar, Cyberterror, Cybercrime & Cyberactivism: An in-depth guide to the role of standards in the cybersecurity environment*. Palo Alto, CA: IEEE Press.

Nobles, C. (2024) 'The Weaponization of Artificial Intelligence in Cybersecurity: A Systematic Review', *Procedia Computer Science*, 239, pp. 547–555. doi:10.1016/j.procs.2024.06.206.

Rickli, J.M. and Mantellassi, F. (2023) 'Artificial intelligence in warfare: military uses of AI and their international security implications', in *The AI wave in defence innovation*. London: Routledge, pp. 12–36.

Souly, A. *et al.* (2025) 'Poisoning Attacks on LLMs Require a Near-constant Number of Poison Samples'. Available at: https://research-ebsco-com.nyit.idm.oclc.org/linkprocessor/plink?id=c10adda0-caaa-3e6a-b9da-1b014b6aadf4 (Accessed: 3 May 2026).

Sticher, V. (2025) 'War and peace in the age of AI', *Cambridge Review of International Affairs*. Available at: https://journals.sagepub.com/doi/10.1177/13691481241293066 (Accessed: 3 May 2026).

Syed, N., Anwar, A., Baig, Z. and Zeadally, S. (2025) 'Artificial intelligence as a service (AIaaS) for cloud, fog and the edge: state-of-the-art practices', *ACM Computing Surveys*, 57(8), pp. 1–36.

Veracode (2025) *2025 GenAI Code Security Report*. Available at: https://www.veracode.com/resources/analyst-reports/2025-genai-code-security-report/ (Accessed: 3 May 2026).