Code, Command, and Conflict: Charting the Future of Military AI

Priyesh Mishra
Prakhar Pandey
Leah Cole
Seungmi Hong
Brandon Tran
Kathy Huang
Isaac Andrew Bangura

NOVEMBER 2025

HARVARD Kennedy School
BELFER CENTER
for Science and International Affairs

Table of Contents

How will the advent of artificial intelligence (AI) shape the logic of geopolitical competition, especially among the great powers? States have already begun to incorporate sophisticated AI systems into their military postures, diplomatic toolkits, and decision-making processes. At the Harvard Kennedy School, a study group, organized by Anatoly Levshin under the auspices of the Belfer Center's Program on Emerging Technology, Scientific Advancement, and Global Policy, endeavored to make sense of these transformations by investigating alternative uses of AI in militarized bargaining and crisis diplomacy. This white paper reports their findings and policy recommendations.

Overview

> *"Whoever becomes the leader in this [AI] sphere will become the ruler of the world."*
> — Vladimir Putin, President of Russia

> *"We must have complete self-reliance, and comprehensively advance technological innovation, industrial development, and all AI-empowered applications, with mastery across all core AI technologies."*
> — Xi Jinping, President of China

> *"From this day forward it'll be a policy of the United States to do whatever it takes to lead the world in artificial intelligence."*
> — Donald Trump, President of the United States

The integration of artificial intelligence into defense and national security is becoming a global priority. For the purposes of this report, we define AI as digital or physical systems that can perform tasks that normally require human intelligence, such as perception, learning, reasoning, and decision-making, and can operate in dynamic environments with varying degrees of autonomy.[1]

Global military spending on AI is estimated to have doubled from $4.6 billion to $9.2 billion USD between 2022 and 2023, and is expected to reach $38.8 billion by 2028.[2] Militaries are using AI to achieve greater operational efficiency and accuracy for strategic advantage against adversaries. These efforts, however, are not new. Between 1966 and 1972, Stanford Research Institute developed Shakey, a robot that used computer vision and language processing to perceive and reason about its surroundings, to make decisions, and perform tasks.[3] In the late 1980s, the US developed the Dynamic Analysis and Replanning Tool (DART) — an AI software designed to optimize transportation of personnel and supplies.[4] At the same time, unmanned weapon systems like autonomous robots and drones were also being developed.

Prior to 2017, however, AI was far from being a strategic military priority. In April 2017, the US Department of Defense (DoD) established the Algorithmic Warfare Cross-Functional Team (aka Project Maven) to accelerate the DoD's integration of big-data and machine learning.[5] In July of that same year, China unveiled its New Generation Artificial Intelligence Development Plan, laying out a high-level design of integrating AI into the country's social, economic, and national security paradigm.[6] A few months later, Russian President Vladimir Putin prophesized that leadership in AI will be key to global dominance in the future.

Perhaps the most significant development was not political. In June 2017 Google released a seminal paper

titled “Attention Is All You Need,” introducing transformers, a deep learning architecture that revolutionized the AI industry. It enabled the creation of general-purpose linguistic models called foundation models. Unlike traditional AI models that are designed and trained for a single, specific task (like image recognition or predicting stock prices), foundation models are built to be general-purpose. This means they learn a broad understanding of patterns, structures, and relationships within the data on which they are trained. The most prominent examples of foundation models are Anthropic’s Claude, Google’s Gemini, and OpenAI’s ChatGPT. Foundation models can be customized for a variety of use cases, including military applications.

This development has had three broad implications. First, the ability to process vast amounts of data with speed and accuracy unlocked a world of novel military applications. Less than ten years after the publication of “Attention is All You Need,” AI-powered solutions like Israel’s Lavender and Gospel systems or Palantir’s MetaConstellation platform have been deployed in active conflicts for precision targeting. Other use cases of AI include war-gaming and generation of plausible strategies, forecasting requirements for equipment maintenance, and logistical planning. Second, foundation models lowered the entry barrier for military integration of AI for states with smaller economies. Crafting usable AI models requires substantial human capital, research computing, and quality data. This high upfront cost, coupled with export controls on cutting-edge hardware that advanced models require for training, makes it prohibitive for smaller states to build their own national AI models. While leading private players like Google, Meta, and OpenAI initially banned military use of their models, that restriction has since been removed.[7] This created opportunities for smaller states to license proprietary models for military applications.

Third, the increasing integration of sophisticated AI systems into national militaries has given technology companies a seat at the high table. The private sector has significant advantages over the public sector throughout the AI production chain. From designing and producing cutting-edge graphic processing units (GPUs) — the essential hardware that supercharges the training of deep neural networks — to assembling and training frontier AI models, several large players such as Alphabet, Anthropic, Meta, Microsoft, NVIDIA, and OpenAI Microsoft dominate the AI industry. Governments across the world are increasingly relying on the private sector for infrastructure and expertise in modernizing their militaries. The appointment of senior executives from Meta, OpenAI, and Palantir into the newly formed Detachment 201 Executive Innovation Corps of the US Army is indicative of the private sector’s growing stature in the defense and national security space in the United States.[8] Consequently, national security increasingly depends on the innovation, ethical standards, and business decisions of these powerful technology firms, creating new institutional dependencies and governance challenges for states.

It is plausible to suspect that AI will yield an array of tangible benefits in warfare. Advances in deep learning promise to improve targeting accuracy, enhance situational awareness, and accelerate decision-making. They can potentially lower the incidence of target misidentification and reduce the likelihood of unintended harm to civilians. Several AI applications have been specifically designed to mitigate collateral damage. These systems use sensor swarms, behavioral analysis, and pattern-of-life assessments to detect transient civilians and identify designated visual symbols such as those used on humanitarian sites. Such automated warning systems can pause engagements whenever risk factors are detected.[9]

However, military uses of AI remain fraught with ethical and legal challenges. Legal guardrails around autonomous weapons systems and AI-based decision support systems are a work in progress and must sit on top of well-developed legal and ethical frameworks, policies, conventions, and traditions. This leaves a regulatory vacuum that exacerbates the risk of violations of core principles of the laws of war, such as the accountability of combatants, proportional use of force, and immunity of noncombatants. Furthermore, excessive augmentation of decision-making with AI courts the dangers of unintended military engagements

and, therefore, unpredictable escalation; this, in turn, limits the room for meaningful and timely diplomatic resolutions to spiraling crises. Finally, the democratization of AI through the distribution of sophisticated open-source models allows non-state actors, including terrorist groups and armed militias, to acquire greater capability to inflict damage.

Given the inevitable integration of AI into the military domain, establishing strong regulatory and ethical safeguards is crucial for mitigating negative consequences. This report will explore the diverse military applications of AI, evaluate the current regulatory landscape, examine the legal and ethical dilemmas, and propose strategies for ensuring its responsible implementation.

Military applications of AI

The growing integration of AI into the national security apparatus will undoubtedly alter the calculus of militarized bargaining, as defined by the use, or threat of use, of military force to achieve desired objectives in disputes. The Russia-Ukraine conflict offers the most visible example of how modern war is conducted with the inclusion of AI and autonomous weapons on the battlefield. The rest of the world is taking note of the rapid technological change in both the lethality and logistical character of warfare. In an era where precision weapons reigned supreme during the 1990s due to their improved targeting, the use of AI with autonomous weapons flipped the paradigm on its head, reflecting that quantity was a quality of its own.

Militaries have been researching and experimenting with this technology since long before Russia's invasion of Ukraine in 2022. The US military invested decades of research to field the first significant use of AI in the 1990s through the Dynamic Analysis and Replanning Tool, representing the first wave of AI that helped optimize the logistics and planning during the Iraq War. Its success fueled the military to invest more in the "enabling technology" and seek different applications within the military. In 2017, the DoD, in partnership with Google, launched Project Maven, an initiative aimed at enhancing the military's surveillance and intelligence capabilities with AI. The project used sophisticated convolutional neural networks to process vast quantities of battlefield data in order to automate recognition and classification of various types of targets for more efficient engagement. Since Palantir took over the project, the Maven Smart System has become a flagship element in the NATO Intelligence, Surveillance, Reconnaissance program.[10]

The ability to process data at unprecedented speed and scale has given military commanders a new level of situational awareness, which will have significant implications for military strategy. For instance, Ukraine used Palantir's data integration tools like MetaConstellation to combine commercial satellite imagery with classified intelligence. This integration was notably impactful during operations such as the liberation of Kherson, where accurate, real-time intelligence on Russian troop positions enabled the execution of precise long-range strikes.[11] Similarly, Israel has deployed Lavender and Gospel, two AI-powered data-processing systems that facilitate high-precision targeting using satellite imagery, intercepted communications, and drone footage.

These examples demonstrate how AI can more accurately and quickly interpret information, potentially leading to better decision-making. The rapid integration of information from a suite of sensors accelerates the decision-making timeline and prompts a reevaluation of the norms regarding where humans fit in the decision-making loop, whether it is outside, inside, or on the loop.

Countries are devising distinct approaches to AI adoption. The US prioritizes the development of domain-specific models developed through initiatives like Project Maven and Overmatch, the US Navy's initiative to

create a data-driven and AI-enhanced fleet that connects every sensor to every shooter. As these models gain access to larger volumes of data, a future hypothetical capability that the military could employ lies in the domain of predictive AI. Such hypothetical systems could eventually predict the likelihood of military engagements and their outcomes with extreme precision. Prototypes of such models can already be used in wargaming, a domain where sophisticated AI models can produce meaningful returns by reducing the minimal number of analysts needed to craft scenarios, increasing the speed of development of wargame mechanics, improving participant immersion, accelerating exercise execution, and identifying innovative strategies and actions.

Much like the US, China has rapidly adapted AI technologies into strategic planning, training scenarios, and professional military education. Recently, researchers at Xi'an Technological University used DeepSeek's foundation model to autonomously generate military simulations, providing a digital testing ground for future competitive confrontations.[12] China's "military-civil fusion" model integrates the civilian AI sector with the military establishment. This synergy positions China to leverage AI in military applications with significant long-term strategic advantages. In 2020, "intelligentization" — the adoption of AI and advanced technologies — was officially affirmed as the third initiative under Xi Jinping's goal of modernizing the People's Liberation Army by 2035.[13] This integration is sustained by laws that compel domestic companies to cooperate with state security organs when requested.[14] China's model could plausibly facilitate efficient incorporation of frontier research into military applications. In any case, however, it also implies that China perceives its exports of AI technologies as one additional instrument of geopolitical leverage.

As the US grapples with effectively countering China's increasingly militarized presence in the Indo-Pacific, the US Navy has initiated its Third Offset Strategy. This strategy aims to counter China's military power by developing and deploying AI technologies, robotics, directed energy weapons such as lasers and railguns, hypersonic weapons, and capabilities for scaling the manufacturing of unmanned weapons systems. The Third Offset Strategy points to yet another domain where machine learning has long facilitated the augmentation of human expertise with AI: autonomous weapons systems (AWS). Presently, the introduction of uncrewed surface vehicles (USVs) and global autonomous reconnaissance crafts (GARCs) can significantly expand the service's power projection into strategically contested areas without increasing risk to servicemembers' lives or valuable surface assets. This approach is both politically and operationally sustainable, demonstrating how AI may be used to counter military threats. In the air domain, initiatives such as the Autonomous Collaborative Teaming effort and the Air Force's Loyal Wingman concept enable individual operations to command and coordinate multiple autonomous platforms. These platforms can execute missions in coordination, share data, and adapt in real-time to dynamic threats. This creates a networked swarm capability that can deceive and overwhelm even state-of-the-art enemy defensive systems. As AWS becomes more proficient in operating in complex environments, its integration into military operations may become increasingly central to mission success. Future operations may consist of swarming drones providing suppression, underwater patrolling of vital strategic chokepoints, or loitering munitions that can reprioritize targets based on changing tactical conditions.

These evolving capabilities present both opportunities and risks that necessitate deeper consideration by military strategists. On the one hand, AWS enables militaries to project power with precision and at low cost, thereby enhancing the credibility of threats without requiring large-scale mobilization. On the other hand, AWS invariably introduces emergent opportunities for unintended escalation. Both the opportunities and the risks must be given due attention as states consider alternative pathways for regulating military applications of AI.

Regulatory Landscape

Domestic Regulations

China, India, the US, and the UK have all affirmed legal principles governing military applications of AI domestically. There is some convergence among these national sets of principles, which reflects ongoing efforts by the great powers to reconcile distinct cultural traditions and ethical imperatives with shared geopolitical exigencies.

In 2020, the US Department of Defense adopted five ethical principles to guide the development and deployment of AI capabilities: responsibility, equitability, traceability, reliability, and governability. Responsibility requires the exercise of adequate care in the development and deployment of AI capabilities. Equitability requires minimization of unintended bias in applications. Traceability requires the training of operating personnel in the suite of skills and methodologies needed to deploy AI capabilities. Reliability requires delineation of potential use cases for AI applications and engineering of testing and monitoring protocols that can ensure adequate performance across those use cases. Finally, governability requires proactive identification of unintended consequences across potential use cases and incorporation of deactivation protocols into AI capabilities. These principles were reinforced in 2023 by the Presidential Executive Order 14110 on "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence."[15] The UK's "Defence Artificial Intelligence Strategy," published in 2022, and India's "Evaluating Trustworthy Intelligence Framework," published in 2024, both articulated similar principles.[16] By contrast, China has not yet formalized any such governance principles. However, China's "Position Paper on the Military Application of Artificial Intelligence," presented at the United Nations in 2022, affirms more generic imperatives, such as safety and control in use, and emphasizes the need for subordinating national uses of AI to the shared interests of humanity.[17]

The similarities in the regulatory nomenclatures used by the four states suggest at least a shared acknowledgment of the need for domestic regulation of military applications of AI. However, this thin stratum of agreement obscures more than it reveals. Due to differences in domestic politics, socio-economic legacies, and international commitments, China, India, the US, and the UK are likely to operationalize regulatory imperatives such as responsibility, equitability, traceability, reliability, and governability quite differently. In turn, these differences may further intensify geopolitical competition, especially between the United States and China. In 2025, President Donald Trump revoked Executive Order 14110, which imposed ethical guidance on cross-domain applications of AI, and instead enacted Executive Order 14179, "Removing Barriers to American Leadership in Artificial Intelligence."[18] The new order shows a clear shift toward deregulation in the quest for "global AI dominance in order to promote human flourishing, economic competitiveness, and national security." The US "AI Action Plan," which was unveiled in July 2025, underlined deregulation as a preferred approach to accelerate AI innovation.[19] The plan, among other things, directs the Office of Management and Budget to "identify, revise, or repeal regulations, rules, memoranda, administrative orders, guidance documents, policy statements, and interagency agreements that unnecessarily hinder AI development or deployment." By contrast, China continues to pursue its strategy of military-civil fusion. A notable example of this is the adoption, in 2022, by the Cyberspace Administration of China of the Internet Information Service Algorithm Recommendation Management Regulations requiring companies to register algorithms with "public opinion characteristics" and "social mobilization capabilities" to a central national filing system.[20] This trend in domestic regulations — shared ethical principles but intensifying competition in security — creates both opportunities and obstacles for international governance.

Fragmented Governance Among Nation States

There is, at present, no global agreement regulating military applications of AI. Instead, this issue area is characterized by the overlapping presence of several nonbinding regulatory frameworks. In 2019, the G20, under Japan's presidency, adopted the "AI Principles" framework identifying five groups of governance principles: inclusivity and sustainability, human centricity, transparency and explainability, security and safety, and accountability.[21] In 2023, the G7 adopted the Hiroshima AI Process (HAIP), a policy framework comprising two sets of high-level guidelines: the "International Guiding Principles for All AI Actors and for Organizations Developing Advanced AI Systems" and the "International Code of Conduct for Organizations Developing Advanced AI Systems."[22] The guidelines are intended to create a risk-cantered framework that can steward frontier AI development toward responsible innovation. Many leading AI developers, including Anthropic, Google, Microsoft, and OpenAI, are voluntarily complying with HAIP. Adoption is monitored through a voluntary reporting and certification mechanism implemented by the OECD.[23] Beyond the G7, HAIP has garnered larger support through the HAIP Friends Group, presently comprising fifty-six countries and twenty-three nongovernment stakeholders including large technology companies and United Nations agencies.[24]

Beyond multilateral clubs such as the G7 and the G20, military alliances and regional organizations have also begun articulating frameworks for the regulation of the use of AI among their member-states. In 2021, NATO defined responsible use of AI in terms of compliance with six broad principles: lawfulness, responsibility and accountability, explainability and traceability, reliability, governability, and bias mitigation.[25] The European Union's AI Act represents arguably the most significant multilateral regulatory effort in this space to date.[26] However, it is crucial to note that Article 2(3) of the AI Act explicitly excludes technology intended exclusively for military use.[27] This exclusion is particularly salient given that, on January 20, 2021, the European Parliament acknowledged the military threat posed by AI and emphasized the importance of coordinating member-states' policies regulating military applications of AI at the European level.[28]

Much of this regulatory effort has centered on challenges presented by Lethal Autonomous Weapon Systems (LAWS).[29] LAWS can be defined as weapons systems that are designed to identify, select, and engage targets without the need for real-time supervision by a human operator. Work in this direction began in 2013, when state parties to the Convention on Certain Conventional Weapons established an expert group to study the impact of emerging technologies on plausible uses of LAWS.[30] In 2016, the group was issued a mandate to propose recommendations for the multilateral regulation of LAWS.[31] In 2019, this Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems endorsed core guiding principles, including the primacy of international humanitarian law, importance of human supervision, urgency of incorporating risk assessment and mitigation protocols into the design of autonomous weapons, and warning against anthropomorphizing LAWS.[32] The Group considered four alternative pathways for incorporating these principles into international practice. First, signatory states could develop legal instruments for the authoritative regulation of the use of LAWS. Second, they could issue a nonbinding declaration outlining core regulatory principles. Third, they could adapt existing legal instruments. And fourth, they could simply affirm the sufficiency of existing legal instruments.[33] Signatory states have not yet chosen either pathway. Presently, 129 nations, including Brazil, China, Italy, and South Africa, advocate for a legally binding agreement. Only twelve countries, including India, Russia, the UK and the US oppose such an agreement, while fifty-four countries have yet to declare their stance.[34] The UN Secretary General and expert organizations like the International Committee of the Red Cross have also advocated for a legally binding instrument restricting the use of LAWS.[35]

This lack of tangible progress has catalyzed efforts by the transnational civil society, including organizations such as Stop Killer Robots, and other multilateral fora, such as the General Assembly of the United Nations, to develop meaningful proposals for potential regulatory frameworks. For example, in October 2023, Austria proposed a resolution to the UN General Assembly's Disarmament and International Security Committee requesting the Secretary-General to survey the official positions of the member-states and prepare a report on the matter for the General Assembly. This resolution, co-sponsored by more than forty states, was adopted by the General Assembly in December 2023.[36]

Furthermore, states are increasingly undertaking efforts to broaden the scope of multilateral regulation of military applications of AI beyond LAWS. For example, the 2023 Global Summit on Responsible AI in the Military Domain, cohosted by the Netherlands and South Korea, sought to bring all conceivable use cases within the regulatory ambit. The summit also established a dedicated global commission to improve understanding of potential military uses of AI and issue guidance on requirements for responsible use.[37] During the summit, the United States unveiled its Political Declaration on Responsible Military Use of AI and Autonomy. The declaration outlines governing principles for the responsible development, deployment, and use of military AI, emphasizing, in particular, the importance of transparency, adequate human training, and rigorous testing of new systems. As of November 2024, fifty-eight countries have endorsed the declaration. Similarly, the General Assembly's Resolution 79/239, passed in December 2024, acknowledged that international humanitarian law should apply across all stages of the lifecycle of any military application of AI, including "systems enabled by AI."[38]

Strategic Risks and Regulatory Challenges

The kaleidoscopic complexity of disparate multilateral initiatives aimed at regulating military uses of AI reflects a shared awareness of strategic risks that military applications of AI create for all states, especially the great powers. These include diminished thresholds for conflict, unpredictable escalation trajectories, unforeseen operational vulnerabilities, and interstate arms races. First, the deployment of autonomous AI systems may reduce political and moral thresholds for engaging in military conflict as the human costs of warfare, at least with the exclusion of civilian casualties, diminish. This diminution may, in turn, contribute to volatile escalation. For example, LAWS are designed to compress sensor-to-shooter decision timelines from minutes to seconds. This reduces the window for human deliberation and increases the likelihood of unintended escalation. Because AI can easily misinterpret adversary activities — say, due to incomplete or misleading data — it may also recommend premature or disproportionate responses. The escalation bias induced by undue automation can also undermine established diplomatic channels and methods for timely de-escalation. This risk is further compounded by the capacity of sophisticated AI systems for emergent behaviors. Such behaviors can deviate from operational plans or rules of engagement. For example, a wargaming exercise conducted by the *Stanford Institute for Human-Centered AI* revealed that, in the absence of human supervision, surveyed LLMs exhibited unpredictable escalatory behavior.[39]

Third, AI-powered systems suffer from two types of unforeseen operational vulnerabilities: unintentional failures and specification gaming. Unintentional failures usually result from overfitting at the training stage, leading the trained system to produce incongruous responses upon encounter with real-world data that statistically differs from its training data.[40] Military conflicts are deeply complex and, therefore, models trained on inadequately rich data may well overfit to action-response patterns which could be irrelevant or damaging in specific real-world scenarios. By contrast, specification gaming refers to an AI model's capacity to achieve a minimal realization of its prefigured objective without, however, accomplishing the qualitative

outcome that model was designed to produce. As one hypothetical scenario of such reward-hacking in the military domain, consider an AI model, optimized for maximum enemy neutralization, resolving to prioritize low-value targets with a higher number of enemies over high-value targets. Both unintentional failures and specification gaming can be exacerbated by efforts on the part of malicious actors to manipulate and degrade AI systems. Such attacks include data poisoning, where adversaries corrupt training data to embed flaws in the model's behavior, and spoofing, where adversaries manipulate real-time inputs to disable or distort AI functionality. State militaries that adapt commercial foundation models face additional risks due to reliance on vast, unverified data sources and code, leaving them vulnerable to backdoor hacking.[41]

Finally, the proliferation of AI capabilities among state and nonstate actors can contribute to arms races and strategic instability. Open-source and commercial foundation models can be repurposed easily for military uses. This lowers the barriers to entry for nonstate actors such as terrorist organizations.[42] In turn, this ease of access to sophisticated AI capabilities promising tactical or strategic advantages to early adopters incentivizes a "race to the bottom" on safety and reliability standards among states and nonstate actors alike. Such competitive pressures may foster rapid deployments of insufficiently vetted systems, amplifying both risks posed by operational vulnerabilities and risks posed by inadvertent escalation.

Nonetheless, shared awareness of these dangers among states has not yet translated into substantive agreement on binding principles that can help them mutually regulate those risks. This regulatory failure reflects, in part, divergences in their national-security interests and socioeconomic legacies. But it also reflects structural challenges that inhere in the very nature of military uses of AI considered as a regulatory problem. One such challenge deserves special attention: opaque accountability. International law requires clear attribution of responsibility for actions during conflict. AI systems complicate this requirement because they can undertake actions without human oversight. This possibility creates a moral buffer that distances operators, commanders, and developers from the real-world consequences of actions taken by specific systems. Without clear pathways of accountability, the deterrent effect of legal punishment weakens and, more worryingly, the very concept of international responsibility for violations of the laws of armed conflict risks losing applicability. In order for states to agree on binding multilateral regulations to govern military uses of AI, it is first necessary for the great powers especially to devise a mutually agreeable solution to this problem of opaque accountability.

Recommendations for Responsible Military AI

The integration of AI into defense and national security is likely to prove pervasive. The sheer complexity of its expected impact on the international order demands a comprehensive policy response. To articulate such a response, we propose a novel forward-oriented framework: preventive security governance. We propose to understand "preventive security governance" as the proactive codification of specific international norms intended to regulate patterns of cooperation and conflict among states and nonstate actors in the national-security domain. We have previously noted that military uses of AI are not presently regulated by any single multilateral framework. We have also argued that this absence of binding regulation creates abiding strategic risks for states and nonstate actors alike. Our focus on articulating a new governance framework for this issue area, then, comports with the precautionary principle of international law, which advocates for action to be taken before harm is done. To improve the safety, trustworthiness, and standard compliance of military applications of AI, we propose eleven concrete recommendations grouped into four actionable categories.

I. Prioritizing Compliance-by-Design and Ethical Engineering

Integrating international legal obligations and ethical principles directly into the technical specifications and development processes of AI systems is a foundational element of preventive security governance.

1. **Mandate "Compliance-by-Design" in AI Development:** States should establish clear legal requirements that AI systems are designed from the outset to comply with international law. This means embedding international norms directly into prospective models' architectures and decision-making parameters rather than attempting to retroactively incentivize compliance.[43] If a system cannot be engineered to ensure compliance with international norms, its development should be halted.

2. **Implement Robust Ethical Review Panels for AI Development:** Beyond legal compliance, multidisciplinary ethical review panels, hosting ethicists, legal scholars, military strategists, and AI developers, should be integrated into the development lifecycle of AI models. These panels would vet AI projects for potential ethical concerns and ensure that deployed AI systems cannot engage in behaviors that can violate peremptory requirements of human dignity.

3. **Integrate International Law and Ethics Training for AI Developers and Military Personnel:** To support compliance-by-design, AI developers and military personnel must receive clear training in relevant aspects of ethics and international law.

II. Crafting Adequate Testing and Certification Standards

Current testing and evaluation (T&E) methodologies, designed for traditional weapons systems, are inadequate for the complexities of AI. This inadequacy necessitates the development of AI-specific standards.

1. **Develop AI-Specific Verification and Validation (V&V) Frameworks:** New V&V frameworks are urgently needed to account for dynamic learning behaviors, opacity of accountability, and operational vulnerabilities inherent in frontier models.[44] These frameworks must include rigorous stress-testing under simulated combat conditions designed to predict performance especially in novel environments.[45]

2. **Mandate Continuous Performance Monitoring and Auditing Post-Deployment:** Given that AI systems can exhibit unpredictable behavior in real-world scenarios, continuous performance monitoring and auditing are essential after deployment. This involves ongoing tracking, regular reevaluations, retraining, retesting, and reapproval of the system.

3. **Establish International Standards for AI Testing and Reliability:** International cooperation is vital to develop and promote universal standards for effective T&E practices governing military applications of AI.[46] These standards should aim to guarantee a common baseline of robustness and reliability for all systems.

III. Ensuring Transparency and Explainability

The black box nature of many AI systems and the inherent biases in data pose significant challenges to human understanding and operator trust.

1. **Prioritize Explainable AI (XAI) in Military Systems:** Research and development should prioritize XAI techniques that can render algorithmic decisions understandable to human operators, legal reviewers, and other stakeholders. Unilateral adoption of XAI might be perceived as a strategic disadvantage in a nascent AI arms race; therefore, robust adoption of XAI principles should be a cornerstone of multilateral arms control agreements governing military development and deployment of sophisticated AI systems.
2. **Increase Transparency Regarding AI's Role and Limitations to Counter Automation Bias:** Operators and commanders are susceptible to automation bias, which denotes undue reliance on AI guidance even in situations when that guidance is factually incorrect or otherwise, say morally or legally, inadequate. To counter automation bias, there must be greater transparency about functional limitations of deployed systems and quality of their training data.

IV. Strengthening Legal and Policy Frameworks for Accountability

The integration of AI into military operations complicates the attribution of responsibility and calls for clear legal and policy frameworks to ensure accountability from design to deployment.

1. **Clarify the Locus of Human Conduct for State Responsibility:** Policies should explicitly outline how individual and state responsibility under international law applies across the lifecycle of deployed AI systems to ensure their autonomy, speed, and unpredictability do not create legal responsibility gaps.

2. **Develop Due Diligence Criteria for AI Acquisition and Use:** States acquiring AI technologies from third parties, such as private companies, should be obligated to ensure those systems comply with international norms. This requirement involves diligently seeking information from developers and testing acquired systems for legal compliance.

3. **Operationalize International Legal Obligations Into Technical Specifications:** Research is needed to translate abstract legal and ethical principles into concrete operational guidelines for AI developers and military planners. This involves adapting existing legal review processes, such as, most notably, Article 36 evaluations under Additional Protocol I of the 1949 Geneva Conventions, to explicitly consider distinctive characteristics of proposed AI models.

Responsible deployment of AI in military contexts is not merely a technical challenge but a governance imperative. Adopting a preventive security governance approach, characterized by compliance-by-design, rigorous testing, transparency, and clear accountability, is essential for negotiating the complex opportunities and risks presented by the many actual and potential military applications of AI. This comprehensive strategy aims to ensure AI systems are inherently safe, trustworthy, and lawful before they ever enter a conflict scenario. By adopting this framework, states can ensure the deployment of military AI contributes to, rather than detracts from, international peace. The urgency of this task cannot be overstated: Policy choices made today will profoundly shape the future of international stability and, indeed, the very character of warfare.

Endnotes

1 This definition builds upon the description of AI in UN Secretary General's Report on *"Current developments in science and technology and their potential impact on international security and disarmament efforts,"* A/78/268, August 1, 2023

2 Kurtis H. Simpson, Samuel Paquette, Raphael Racicot, Samuel Villanove, *Militarizing AI: How to Catch the Digital Dragon?,* Centre for International Governance and Innovation, February 2025. https://www.cigionline.org/articles/militarizing-ai-how-to-catch-the-digital-dragon/#:~:text=Worldwide%20estimates%20of%20military%20spending,testing%20of%20military%20AI%20applications.

3 Shakey, Computer History Museum, https://www.computerhistory.org/revolution/artificial-intelligence-robotics/13/289

4 S. Cross and R. Estrada, "DART: an example of accelerated evolutionary development," *Proceedings of IEEE 5th International Workshop on Rapid System Prototyping*, Grenoble, France, 1994, pp. 177-183, doi: 10.1109/IWRSP.1994.315895

5 US Department of Defense Memorandum, April 26, 2017. https://www.govexec.com/media/gbc/docs/pdfs_edit/establishment_of_the_awcft_project_maven.pdf

6 State Council Notice on the Issuance of the New Generation Artificial Intelligence Development Plan, Translated version, DIGICHINA, Stanford University. July 2017. https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/

7 Sam Biddle, "OpenAI Quietly Deletes Ban on Using ChatGPT for 'Military and Warfare,'" *The Intercept*, January 12, 2024. https://theintercept.com/2024/01/12/open-ai-military-ban-chatgpt/#:~:-text=%E2%80%9CGiven%20the%20use%20of%20AI,at%20the%20Federal%20Trade%20Commission

8 Heather Somerville, "The Army's Newest Recruits: Tech Execs From Meta, OpenAI and More," *The Wall Street Journal*, June 13, 2025. https://www.wsj.com/tech/army-reserve-tech-executives-meta-palantir-796f5360

9 Larry Lewis and Andrew Ilachinski, *"Leveraging AI to Mitigate Civilian Harm,"* Center for Naval Analyses, February 2022. https://apps.dtic.mil/sti/trecms/pdf/AD1181578.pdf

10 Henry Foy and Tim Bradshaw, "Nato acquires AI military system from Palantir." *Financial Times*, April 14, 2025. https://www.ft.com/content/7f80b1bc-114c-4a00-ad06-6863fb435822

11 Goncharuk, V. (2024). Survival of the Smartest? Defense AI in Ukraine. In *The Very Long Game* (pp. 375-395). Springer

12 Xi, Yu, and Liu Xuanzun. "Chinese University Unveils New DeepSeek-Based Simulated Military Scenario Generator." *Global Times*, May 15, 2025. https://www.globaltimes.cn/page/202505/1334151.shtml

13 Wang, Zichen. "Once-in-a-generation change in PLA guidelines: intelligentization added, mechanization declared 'basically accomplished.'" *Pekingnology*, December 8, 2020. https://www.pekingnology.com/p/once-in-a-generation-change-in-pla

14 Jili, Bulelani. "China's surveillance ecosystem and the global spread of its tools." *Atlantic*

Council, October 17, 2022. https://www.atlanticcouncil.org/in-depth-research-reports/issue-brief/chinese-surveillance-ecosystem-and-the-global-spread-of-its-tools/

15 Executive Order 14110, "Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." The White House, October 30, 2023. https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence.

16 Defence Artificial Intelligence Strategy, Ministry of Defence, Government of UK, June 2022. https://assets.publishing.service.gov.uk/media/62a7543ee90e070396c9f7d2/Defence_Artificial_Intelligence_Strate.pdf and Framework & Guidelines to integrate Trustworthy AI into critical infrastructure sectors released, Press Information Bureau, Government of India, October 17, 2024. https://www.pib.gov.in/PressReleasePage.aspx?PRID=2065847

17 Position Paper of the People's Republic of China on Strengthening Ethical Governance of Artificial Intelligence (AI), Ministry of Foreign Affairs, People's Republic of China, 2022. https://www.fmprc.gov.cn/eng/zy/wjzc/202405/t20240531_11367525.html

18 Removing Barriers to American Leadership in Artificial Intelligence, The White House, January 23, 2025. https://www.whitehouse.gov/presidential-actions/2025/01/removing-barriers-to-american-leadership-in-artificial-intelligence/

19 Winning the Race: America's Action Plan, The White House, July 2025. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

20 Original CSET translation of "The '13th Five-Year' Special Plan for S&T Military-Civil Fusion Development," PRC Ministry of Science and Technology, August 24, 2017. https://cset.georgetown.edu/research/the-13th-five-year-special-plan-for-st-military-civil-fusion-development/

21 G20 AI Principles, Japan 2019. https://www.mofa.go.jp/policy/economy/g20_summit/osaka19/pdf/documents/en/annex_08.pdf

22 G7 Reporting Framework – Hiroshima AI Process. https://transparency.oecd.ai/

23 Ibid.

24 Members of the Hiroshima AI Process Friends Group. https://www.soumu.go.jp/hiroshimaaiprocess/en/supporters.html

25 NATO, "Summary of the NATO Artificial Intelligence Strategy," October 22, 2021. https://www.nato.int/cps/en/natohq/official_texts_187617.htm; NATO Archives, PO(2021)0350-ANNEX2, https://archives.nato.int/po-2021-0350-annex2-eng

26 European Union, Regulation (EU) 2024/1689 (AI Act), OJ L 202 (July 12, 2024), art. 2(3), https://eur-lex.europa.eu/eli/reg/2024/1689/oj

27 Treaty on European Union, art. 4(2), Consolidated Version at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012M/TXT

28 Artificial intelligence: questions of interpretation and application of international law, (2020/2013(INI)), OJ C 456/34, January 20, 2021. https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52021IP0009

29 UN Office for Disarmament Affairs (UNODA), *"Background on LAWS in the CCW,"* retrieved August 18, 2025. https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/background-on-laws-in-the-ccw/

30 UN Office for Disarmament Affairs (UNODA), *"Timeline of LAWS in the CCW,"* retrieved August 18, 2025. https://disarmament.unoda.org/timeline-of-laws-in-the-ccw/?utm_source=chatgpt.com

31 Ibid.

32 https://dig.watch/processes/gge-laws

33 Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/GGE.1/2018/3, October 23, 2018. https://documents.un.org/doc/undoc/gen/g18/323/29/pdf/g1832329.pdf

34 Positions of UN members on the issue that states adopt new legally binding rules on autonomous weapons systems can be found consolidated here: https://automatedresearch.org/state-positions/

35 International Committee of the Red Cross, "Position on Autonomous Weapon Systems and Background Paper," May 12, 2021. https://www.icrc.org/en/download/file/166330/icrc_position_on_aws_and_background_paper.pdf

36 UN General Assembly Resolution A/RES/78/241, December 28, 2023. https://docs.un.org/en/A/RES/78/241

37 Responsible AI in the Military Domain: Call to Action," February 16, 2023, retrieved August 18, 2025. https://www.government.nl/documents/publications/2023/02/16/reaim-2023-call-to-action

38 United Nations General Assembly Resolution 79/239, *"Artificial intelligence in the military domain and its implications for international peace and security,"* December 24, 2024. https://digitallibrary.un.org/record/4071348?v=pdf

39 Rivera, Juan-Pablo, Gabriel Mukobi, Anka Reuel, Max Lamparth, Chandler Smith, and Jacquelyn Schneider, "Escalation Risks from LLMs in Military and Diplomatic Contexts," Stanford HAI, May 2, 2024. https://hai.stanford.edu/assets/files/2024-05/Escalation-Risks-Policy-Brief-LLMs-Military-Diplomatic-Contexts.pdf.

40 Quinonero-Candela, Joaquin, Masashi Sugiyama, Anton Schwaighofer, and Neil D. Lawrence, eds. *Dataset Shift in Machine Learning*. Cambridge, MA: MIT Press, 2022

41 Heidy Khlaaf and Sarah Myers West (AI Now Institute), Meredith Whittaker (Signal), *Mind the Gap: Foundation Models and the Covert Proliferation of Military Intelligence, Surveillance, and Targeting, ArXiv*, October 2024. https://doi.org/10.48550/arXiv.2410.14831

42 Simmons-Edler, Riley, Ryan P. Badman, Shayne Longpre, and Kanaka Rajan, "AI-Powered Autonomous Weapons Risk Geopolitical Instability and Threaten AI Research," *ArXiv*, May 31, 2024

43 Boutin, Bérénice. "State Responsibility in Relation to Military Applications of Artificial Intelligence," *Leiden Journal of International Law* 36, no. 1 (2023): 133–50. https://doi.org/10.1017/S0922156522000607

44 Cooper, Shannon, Damian Copeland, and Lauren Sanders. "Methods to Mitigate Risks Associated

With the Use of AI in the Military Domain," in *Chapman and Hall/CRC eBooks*, 127–52, 2024. https://doi.org/10.1201/9781003410379-9

45 Greipl, Anna, Geneva Academy of International Humanitarian Law and Human Rights, and International Committee of the Red Cross. "Expert Consultation Report on AI and Related Technologies in Military Decision-Making on the Use of Force in Armed Conflicts," *Geneva*, March 2024. https://www.geneva-academy.ch/joomlatools-files/docman-files/Artificial%20Intelligence%20And%20Related%20Technologies%20In%20Military%20Decision-Making.pdf

46 Michael C. Horowitz, and Paul Scharre. "AI and International Stability: Risks and Confidence-Building Measures," *Technology & National Security*, Center for a New American Security, 2021. https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/AI-and-International-Stability-Risks-and-Confidence-Building-Measures.pdf