GC REAIM

GLOBAL COMMISSION ON RESPONSIBLE AI IN THE MILITARY DOMAIN

Foreign, Commonwealth & Development Office

GC REAIM Expert Policy Note Series

The Risks of Integrating Generative AI into Weapon Systems

Vincent Boulanin

April 2025

POWERED BY The Hague Centre for Strategic Studies

GC REAIM Expert Policy Note Series

The Risks of Integrating Generative AI into Weapon Systems

Authors: Vincent Boulanin

April 2025

Cover photo: Unsplash

The Global Commission on Responsible Artificial Intelligence in the Military Domain (GC REAIM) is an initiative of the Government of the Netherlands that was launched during the 2023 REAIM Summit on Responsible Artificial Intelligence in the Military Domain in The Hague. Upon request of the Dutch Ministry of Foreign Affairs, the Hague Centre for Strategic Studies acts as the Secretariat of the Commission.

The GC REAIM Expert Policy Note Series was funded by the Foreign, Commonwealth and Development Office (FCDO) of the United Kingdom. GC REAIM Experts maintained full discretion over the topics covered by the Policy Notes. The contents of the GC REAIM Expert Policy Note series do not represent the views of the Global Commission as a whole. The Policy Notes are intended to highlight key issues related to the governance of AI in the military domain and provide policy recommendations.

HCSS
Lange Voorhout 1
2514 EA The Hague

Follow us on social media:
@hcssnl

The Hague Centre for Strategic Studies
Email: info@hcss.nl
Website: www.hcss.nl

1. Introduction

Generative artificial intelligence (AI) may no longer need an introduction, given that it is the technology that powers the general-purpose AI systems (e.g. ChatGPT, Gemini, Claude, DeepSeek) that have taken the world by storm.[1] Simply put, generative AI refers to a category of artificial intelligence algorithms that can generate new content, such as text, images, audio, and even code, based on the data they have been trained on.[2] The performance of generative AI models, particularly those trained on the text (so-called large language models) has drastically improved over the past five years – to the point that some experts believe that generative AI is the technical paradigm that could lead to artificial general intelligence.[3]

While the integration of generative AI into the military domain is still in its nascent stages, militaries worldwide are actively exploring its potential applications. For instance, shortly after the launch of ChatGPT, a powerful language model developed by OpenAI, the United States Department of Defence (DoD) established Task Force LIMA to comprehensively assess the opportunities presented by generative AI for the DoD.[4] Furthermore, companies at the forefront of developing the most potent AI models have begun strategically positioning themselves to supply generative AI solutions tailored specifically for military purposes. Anthropic, for example, has partnered with Palantir and Amazon Web Services to provide US intelligence and defence agencies with access to Claude, its advanced AI model.[5] OpenAI has also entered into a strategic partnership with Anduril Industries to collaborate on counter-unmanned aircraft systems.[6] Meanwhile, Meta and Google have amended their internal policies to facilitate their involvement in military contracts. These developments underscore a growing willingness on both the demand and supply sides to foster the adoption of generative AI capabilities in the military domain.

---

[1] Miles Brundage, 'Time's Up for AI Policy', Substack newsletter, *Miles's Substack* (blog), 20 December 2024, https://milesbrundage.substack.com/p/times-up-for-ai-policy.

[2] Adam Zewe, 'Explained: Generative AI', MIT News | Massachusetts Institute of Technology, 9 November 2023, https://news.mit.edu/2023/explained-generative-ai-1109.

[3] These progress of generative AI are primarily attributed to scaling. Models got better because they were trained on more training data and have access to more computing power. Whether that approach will continue to deliver major improvements is debated. Some experts believe scaling will hit a limit (not least because there is so much data to train systems on AI) and will require breakthroughs at algorithmic level to make generative AI more accurate and better formal reasoning; Gary Marcus, 'The New AI Scaling Law Shell Game', Substack newsletter, *Marcus on AI* (blog), 24 November 2024, https://garymarcus.substack.com/p/a-new-ai-scaling-law-shell-game; Gary Marcus, 'AGI versus "Broad, Shallow Intelligence"', Substack newsletter, *Marcus on AI* (blog), 13 January 2025, https://garymarcus.substack.com/p/agi-versus-broad-shallow-intelligence.

[4] Office of the Chief Digital and Artificial Intelligence Officer, Department of Defense, *Task Force Lima Executive Summary* (August 2023), https://www.ai.mil/Portals/137/Documents/Resources%20Page/2024-12-TF%20Lima-ExecSum-TAB-A.pdf.

[5] Kyle Wiggers, 'Anthropic Teams up with Palantir and AWS to Sell AI to Defense Customers', *TechCrunch* (blog), 7 November 2024, https://techcrunch.com/2024/11/07/anthropic-teams-up-with-palantir-and-aws-to-sell-its-ai-to-defense-customers/.

[6] Anduril Industries, 'Anduril Partners with OpenAI to Advance U.S. Artificial Intelligence Leadership and Protect U.S. and Allied Forces', 12 April 2024, https://www.anduril.com/anduril-partners-with-openai-to-advance-u-s-artificial-intelligence-leadership-and-protect-u-s/.

However, it is crucial to acknowledge that generative AI remains, in many respects, an immature and brittle technology, exhibiting limitations in aspects that are critical in military contexts, such as correctness, reliability, and predictability.[7] While these limitations may not pose significant challenges for all military applications of AI, they warrant careful consideration for weapon systems and other means of warfare that are used in combination with kinetic capabilities, including decision support systems for battle management and targeting.[8]

This policy note delves into the risks associated with integrating generative AI into weapon systems following a structured risk assessment approach. It begins by mapping out potential applications of generative AI in weapon systems, followed by a comprehensive identification of the risks that could arise from such use. It then evaluates the likelihood of generative AI integration in weapon systems and concludes by analysing the options available to states and industry actors to mitigate these risks.

[7] Emily M. Bender et al., 'On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜', in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21 (New York, NY, USA: Association for Computing Machinery, 2021), 610–623, https://doi.org/10.1145/3442188.3445922.
[8] Heidy Khlaaf, Sarah Myers West, and Meredith Whittaker, 'Mind the Gap: Foundation Models and the Covert Proliferation of Military Intelligence, Surveillance, and Targeting', *ResearchGate*, 2024, https://doi.org/10.48550/arXiv.2410.14831.

2. Potential Military Benefits of Generative AI for Weapons Systems

In the realm of weapon systems, generative AI holds the promise of significantly enhancing autonomy by improving capabilities in three key areas: perception, human-machine interaction, and adaptiveness.

2.1 Perception

Computer vision models and other models that allow robots to perceive the world are typically trained with vast amounts of labelled data. Collecting and curating such data is resource-intensive. In the military context, finding a sufficient volume of good data can also be difficult. Generative AI's ability to create new content could present an opportunity in that context. Generative AI can generate synthetic data that closely resembles real-world sensor input and can augment existing datasets. This synthetic data could be utilized to train weapon systems, enabling them to recognize objects with greater accuracy, even in challenging environmental conditions that may impair sensor performance. For instance, weapon systems could be trained to identify targets effectively in situations characterized by poor lighting, or other adverse environmental factors.[9]

The value proposition of generative AI extends beyond its ability to provide valuable training data. The process through which generative AI algorithms learn to create new data also enables them to discern intricate patterns in data, which in turn can help computers and robots make sense of the world they interact with. Generative AI leverages large language models (LLMs) and, increasingly, multimodal models (MMs), which are trained on multiple types of data, such as text, images, and video. MM can generate images and videos from text but also describe in natural language what is represented in an image and video. That latter capability is particularly relevant for weapon systems, and robots in general, because it could increase their "scene understanding" ability. MM could enable robots to transition from a paradigm where they could only detect or recognize discrete objects to one where they can understand the relationships between these objects and people in specific contexts. For instance, if a robot detects people running towards an object, generative AI can assist the robot in generating possible explanations for this behaviour. In the context of weapon systems, this capability could have multiple applications. It could be utilized to improve the ability of weapons to distinguish between combatants and non-combatants, including not only civilians but also military personnel who are hors de combat. Additionally, the ability to

[9] It should be noted that this possibility remains debated among computer vision experts. Some researchers have shown that relying on synthetic training data can actually degrade the performance of AI models; Ilia Shumailov et al., 'AI Models Collapse When Trained on Recursively Generated Data', *Nature* 631, no. 8022 (July 2024): 755–59, https://doi.org/10.1038/s41586-024-07566-y; M.L. Cummings and Ben Bauchwitz, 'Identifying Research Gaps through Self-Driving Car Data Analysis', *IEEE Transactions on Intelligent Vehicles*, 2024, 1–10, https://doi.org/10.1109/TIV.2024.3506936.

understand scenes could, in theory, be leveraged to assist with collateral damage assessment and the selection of precautionary measures to minimize harm to civilian populations.

2.2 Human-Machine Interaction

Drawing inspiration from voice assistants in the civilian domain, generative AI has the potential to facilitate more natural and effective communication between humans and robots. Generative AI can enable robots not only to engage in conversations but also to perform task planning based on natural language instructions. The ability to command robots to execute tasks through voice commands has been a long-standing aspiration for military planners.[10] However, until the advent of LLMs, speech recognition technology remained too unreliable to be employed for anything beyond simple commands and non-life-critical tasks.

Generative AI has ushered in a paradigm shift by enabling robots to not only process human speech more accurately but also to translate these instructions into executable computer code. This capability is an active area of research for organizations developing legged robots, such as Agility Robotics and Boston Dynamics. Furthermore, this capability could be utilized to enable robots to communicate with enemy combatants who are wounded or show signs of surrender, as well as with civilians who may be present.

2.3 Adaptiveness

In the context of robotics, adaptiveness refers to the ability of robots to autonomously adjust their actions in response to unexpected obstacles or changes in the environment. Adaptiveness has long been a major challenge in the design of autonomous systems. Traditionally, developing adaptive systems required anticipating, at the design stage, the various situations or changes the systems might encounter. Situations that were not foreseen during the design stage could lead to suboptimal performance or even system failure. This model made the design of autonomous systems for complex, dynamic, and adversarial environments historically very difficult.[11]

Generative AI has the potential to transform this paradigm in two ways. Firstly, it could enable robots to rely on more general knowledge of the world, facilitating the transfer of knowledge and skills from one context to another. Secondly, it could assist robots in simulating different scenarios, predicting their associated outcomes, and selecting the most appropriate course of action accordingly. In the context of autonomous vehicles, generative models are already being employed to forecast the actions of other vehicles and pedestrians. In the context of weapon systems, this capability could be leveraged for various purposes, including navigation and mission (re)planning.

---

[10] Ruth David and Paul Nielsen, 'Final Report of the Defense Science Board Summer Study on Autonomy. Publicly-Releasable Version', June 2016, https://www.researchgate.net/publication/306286423_Final_Report_of_the_Defense_Science_Board_Summer_Study_on_Autonomy_Publicly-Releasable_Version.

[11] Vincent Boulanin and Maaike Verbruggen, 'Mapping the Development of Autonomy in Weapon Systems' (SIPRI, November 2017), https://www.sipri.org/publications/2017/policy-reports/mapping-development-autonomy-weapon-systems.

3. Risks Associated with Integrating Generative AI into Weapon Systems

Despite the potential benefits, it is essential to acknowledge and address the significant risks associated with integrating generative AI into weapon systems. These risks can be broadly categorized as accidental risks, misuse and adversarial risks, and structural risks.[12]

3.1 Accidental Risks

Accidental risks encompass the possibility that the intended use of an AI technology could lead to unintended negative consequences. These risks typically stem from three main sources:

1. **Technical malfunctions or poor technical performance**. This includes instances where the AI system malfunctions or performs poorly, leading to unintended outcomes. For example, an autonomous weapon system might misidentify a civilian bus as a military vehicle, resulting in civilian casualties. Such incidents may be caused by multiple factors, including algorithmic bias.
2. **Human error**. This involves mistakes made by human operators in deploying or operating generative AI-enabled weapon systems. For instance, humans might decide to deploy an autonomous system in a context for which it was not designed, leading to unforeseen consequences.
3. **Negative externalities**. This refers to situations where the technology functions as intended but generates unintended negative consequences at a broader more systemic level. For example, increasing reliance on autonomous systems could accelerate the pace of warfare, potentially escalating conflicts and increasing the risk of unintended harm (more on this in section 3.3).

The risk of technical malfunctions or poor technical performance is particularly relevant in the case of generative AI. Despite the rapid advancements made since the introduction of ChatGPT in 2022, including the development of so-called "reasoning models" such as OpenAI's O1, generative AI systems continue to exhibit significant limitations in terms of correctness and reliability.[13] The inner workings of generative AI systems and their capacity for reasoning are still subjects of debate, but their reliability issues are undeniable.[14] Generative AI systems frequently "hallucinate," generating incorrect answers or representations. A compounding challenge is the lack of robust methods to evaluate their reliability and the likelihood of generating untrustworthy

[12] Remco Zwetsloot and Allan Dafoe, 'Thinking About Risks From AI: Accidents, Misuse and Structure', *Lawfare*, 2019, https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure.
[13] Melanie Mitchell, 'The LLM Reasoning Debate Heats Up', Substack newsletter, *AI: A Guide for Thinking Humans* (blog), 21 October 2024, https://aiguide.substack.com/p/the-llm-reasoning-debate-heats-up.
[14] Subbarao Kambhampati et al., 'Position: 41st International Conference on Machine Learning, ICML 2024', *Proceedings of Machine Learning Research* 235 (2024): 22895–907.

outputs.[15] In the context of weapon systems, the limitations of generative AI in terms of correctness and reliability could pose significant challenges and potentially lead to harmful consequences. To illustrate this, let's revisit the capabilities discussed earlier: machine perception, human-machine interaction, and adaptiveness.

In the case of machine perception, the generation of inaccurate or biased synthetic data, if left unchecked, could adversely affect the weapon's perceptual capabilities in unpredictable ways. This raises the risk of the weapon misidentifying civilians or civilian objects as legitimate targets. Similarly, when applied to situations requiring scene understanding, such as assisting with collateral damage assessment, generative AI systems could recommend or pursue courses of action that are inappropriate, potentially exposing civilians and military personnel to unintended harm.

Using generative AI to translate high-level verbal commands into executable plans for weapons could also lead to unintended consequences. The risk here is that the systems might misinterpret the instructions and pursue a course of action that deviates from the commander's intent. In the context of targeting and battle management, this could result in unintended harm to individuals and objects protected under international humanitarian law or trigger escalation dynamics, such as weapons venturing into enemy territory they were not supposed to enter.

Using generative AI to enable weapons to communicate with civilians and enemy combatants could also be problematic if the technology performs poorly on certain languages or accents, as is currently the case for many languages. If the systems have been trained on datasets that lack cultural sensitivity, there are also the risks of not recognizing expressions of surrender and misinterpreting certain behaviours as signs of hostile intent.

Employing generative AI to enhance the adaptiveness of weapon systems introduces an alignment problem for developers and users. They would need to ensure that the systems do not develop subgoals or action plans that are:

- Not aligned with the intentions of commanders.
- In contravention of international law and safety considerations applicable to armed forces using the systems.
- Likely to cause unintended or disproportionate harm.

3.2 Misuse and Adversarial Risks

Misuse risks typically refer to the possibility that an AI technology could be intentionally used in ways not intended by its creators, with or without malicious intent.[16] It is important to note that any weapon or capability within a weapon can be misused. However, discussing misuse in this context is particularly relevant from the perspective of adversarial attacks.

---

[15] Maribeth Rauh et al., 'Gaps in the Safety Evaluation of Generative AI', *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7, no. 1 (16 October 2024): 1200–1217, https://doi.org/10.1609/aies.v7i1.31717.

[16] Miles Brundage et al., 'The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation', 20 February 2018, https://doi.org/10.17863/CAM.22520.

Adversaries could seek to exploit the cyber vulnerabilities introduced by generative AI to cause weapon systems to malfunction or behave in ways that endanger protected individuals and objects. While the risk of cyberattacks is not new, generative AI exacerbates existing cybersecurity concerns and introduces new attack vectors. Adversaries could employ techniques such as data poisoning or jailbreaking to manipulate or gain control of generative AI-enabled weapon systems.[17] The fact that existing models are largely trained on data from the internet represents from that perspective a challenge.

3.3 Structural Risks

Accidental and misuse risks typically focus on the final steps in the causal chain leading to harm, such as an AI system behaving in an unintended way or the possibility of misuse by an actor. The concept of structural risks offers a broader perspective, considering how technology shapes the broader environment in ways that can be disruptive and ultimately lead to harm.[18]

From this perspective, the integration of generative AI into weapon systems could exacerbate three commonly discussed concerns surrounding the use of AI in the military domain: the erosion of human control, the acceleration of warfare, and accountability.

The potential for generative AI to enhance the perception, adaptiveness, and human-machine interaction capabilities of weapons suggests a potential transformation in the relationship between military personnel and the weapons they use. Futuristic visions of "human-machine teaming" advocated by some military planners in countries like the USA and the UK could become a reality.[19] However, from a legal and policy standpoint, this raises questions about how military personnel would exercise their responsibility over the use of force, as agreed upon at the UN Convention on Certain Conventional Weapons (CCW) through the adoption of the 2019 Guiding Principles on Autonomous Weapons Systems.[20]

A key concern is that advances in perception, adaptiveness, and human-machine interaction could lead to greater autonomy in weapon systems. This, in turn, could result in a greater disconnect between human commanders and the battlefield, not only in

---

[17] Yutong Zhang et al., 'A Review of Adversarial Attacks in Computer Vision', 2023, https://doi.org/10.48550/ARXIV.2308.07673; Alexander Robey et al., 'Jailbreaking LLM-Controlled Robots', 2024, https://doi.org/10.48550/ARXIV.2410.13691.

[18] Remco Zwetsloot and Allan Dafoe, 'Thinking About Risks From AI: Accidents, Misuse and Structure', *Lawfare*, 2019, https://www.lawfaremedia.org/article/thinking-about-risks-ai-accidents-misuse-and-structure.

[19] Defense Science Board Washington DC, 'Defense Science Board Task Force Report: The Role of Autonomy in DoD Systems', 2012, https://apps.dtic.mil/sti/citations/ADA566864; United Kingdom. *Human Machine Touchpoints: The United Kingdom's Perspective on Human Control over Weapon Development and Targeting Cycles.* Working paper CCW/GGE.2/2018/WP.1, presented at the *Group of Governmental Experts on Lethal Autonomous Weapons Systems, Convention on Certain Conventional Weapons*, Geneva, 8 August 2018. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-*Group_of_Governmental_Experts*(2018)/2018_GGE%2BLAWS_August_Working%2BPaper_UK.pdf.

[20] United Nations. Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems. CCW/GGE.1/2019/3, 25 September 2019. Annex IV, "Guiding Principles." https://documents.unoda.org/wp-content/uploads/2020/09/CCW_GGE.1_2019_3_E.pdf.

terms of physical distance but also in terms of cognitive distancing. The potential consequence of this disconnect, unless appropriate risk reduction measures are implemented, could be an increased risk of unintended civilian harm and accidental or inadvertent escalation.[21] While these concerns have been extensively discussed in the literature on military AI, it is important to recognize that the integration of generative AI could further amplify these risks.[22]

Another related concern is that the integration of generative AI into weapons and military decision support systems could contribute to an acceleration of warfare. This accelerated pace of operations could increase the risk of escalation as human commanders would have less time to make informed decisions and would need to rely more heavily on technology to guide their actions at all levels of warfare: strategic, operational, and tactical.[23]

Furthermore, the affordances of generative AI in weapon systems could exacerbate concerns regarding legal compliance and accountability. A study conducted by the Stockholm International Peace Research Institute (SIPRI) concluded that ensuring respect for international humanitarian law (IHL) in the use of autonomous weapon systems (AWS) hinges on the fulfilment of three conditions:[24]

1. The ability to reliably foresee whether the effects of the AWS would, in some or all circumstances, contravene specific and/or general prohibitions and restrictions on weapons, means, and methods of warfare.
2. The ability to administer the operation of AWS in a manner consistent with the rules governing the conduct of hostilities.
3. The ability to trace the operation, performance, and effects of AWS back to the relevant human agent(s).

The nature of generative AI could make the fulfilment of these conditions challenging. The models underpinning generative AI, whether LLMs or MMs, are inherently opaque. Even the creators of these systems often struggle to explain what the AI has learned during the training process and why they behave in certain ways. Moreover, there is a lack of formal methodologies to evaluate the reliability of this technology. While

---

[21] Vincent Boulanin et al., 'Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control', 2020, https://www.sipri.org/sites/default/files/2020-06/2006_limits_of_autonomy.pdf.
[22] Vincent Boulanin, 'Risks and Benefits of AI-Enabled Military Decision-Making', in *Research Handbook on Warfare and Artificial Intelligence* (Edward Elgar Publishing, 2024), 99–115, https://www.elgaronline.com/edcollchap/book/9781800377400/book-part-9781800377400-11.xml.
[23] Gregory Allen et al., 'Strategic Competition in an Era of Artificial Intelligence', CNAS, 2018, https://www.cnas.org/publications/reports/strategic-competition-in-an-era-of-artificial-intelligence; Michael C. Horowitz, 'When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability', *Journal of Strategic Studies* 42, no. 6 (19 September 2019): 764–88, https://doi.org/10.1080/01402390.2019.1621174; Michael Horowitz and Paul Scharre, 'AI and International Stability: Risks and Confidence-Building Measures', CNAS, 2021, https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.
[24] Vincent Boulanin, Netta Goussac, and Laura Bruun, 'Autonomous Weapon Systems and International Humanitarian Law: Identifying Limits and the Required Type and Degree of Human–Machine Interaction' (SIPRI, June 2021), https://www.sipri.org/publications/2021/policy-reports/autonomous-weapon-systems-and-international-humanitarian-law-identifying-limits-and-required-type.

empirical tests can be conducted, they cannot account for all possible situations the systems might encounter once deployed in real-world scenarios.

These challenges make it potentially difficult to predict, control, and investigate the effects of generative AI-enabled weapon systems, especially in complex and dynamic environments. This concern is not unique to generative AI, as it applies to any AI-enabled weapon system that relies on machine learning for critical functions. However, generative AI exacerbates existing concerns about the reliability, predictability, control, and explainability of AI-enabled military systems that states have been grappling with for years.

4. How Likely Is the Integration of Generative AI into Weapons?

How likely is the integration of generative AI into weapon systems? Ultimately, the answer depends on the progress of technology, use cases and how military actors worldwide weigh the perceived benefits against the associated risks. On the technical front, given the current state of technology, the risks appear to outweigh the benefits. Therefore, it would be logical for militaries to adopt a cautious approach, potentially limiting the use of generative AI to non-critical functions. However, the prevailing geopolitical landscape and market pressures exerted by certain actors could shift this balance.

4.1 Hurdles to Adoption

Historically, the military sector has been slow to adopt new technologies due to cultural, institutional, and, perhaps more importantly, safety and operational considerations.[25] Militaries typically prioritize technologies they can trust in challenging and unpredictable situations. Guarantees of reliability are paramount for the formal adoption of new capabilities. History is full of examples of R&D projects that showcased impressive capabilities but ultimately stalled at the prototype stage because militaries could not obtain assurances of reliability and predictability, and therefore safety and effectiveness, once deployed.[26]

Modern militaries adhere to stringent safety requirements and robust testing and evaluation procedures for military systems in general and weapon systems in particular.[27] Some countries, such as the US and the UK, have also adopted specific policies to guide the procurement of AI-enabled systems.[28] These policies articulate specific limits and requirements on the design and adoption of such systems, which could make the integration of generative AI into weapon systems unlikely, at least in the near term. The level of correctness, reliability, and predictability exhibited by current generative AI systems is simply not sufficient to meet the stringent standards that

[25] Paul Scharre, 'Autonomous Weapons and Operational Risk', CNAS, 2016, https://www.cnas.org/publications/reports/autonomous-weapons-and-operational-risk.
[26] Vincent Boulanin and Maaike Verbruggen, 'Mapping the Development of Autonomy in Weapon Systems' (SIPRI, November 2017), https://www.sipri.org/publications/2017/policy-reports/mapping-development-autonomy-weapon-systems.
[27] UK Ministry of Defence, 'An Introduction to System Safety Management in the MOD Part 2', 2018, https://assets.publishing.service.gov.uk/media/5f1ff88bd3bf7f5969fa0f62/SSM_Whitebook_PART_II_v5.pdf.
[28] US Department of Defense, 'DoD Announces Update to DoD Directive 3000.09, "Autonomy In Weapon Systems"', U.S. Department of Defense, 2023, https://www.defense.gov/News/Releases/Release/article/3278076/dod-announces-update-to-dod-directive-300009-autonomy-in-weapon-systems/https%3A%2F%2Fwww.defense.gov%2FNews%2FReleases%2FRelease%2FArticle%2F3278076%2Fdod-announces-update-to-dod-directive-300009-autonomy-in-weapon-systems%2F; UK Ministry of Defence, 'JSP 936: Dependable Artificial Intelligence (AI) in Defence (Part 1: Directive)', GOV.UK, 2024, https://www.gov.uk/government/publications/jsp-936-dependable-artificial-intelligence-ai-in-defence-part-1-directive.

militaries typically demand for critical software in weapon systems. This does not preclude the use of generative AI in other military applications. Generative AI will probably be utilized in information operations and to improve various support functions, such as maintenance, logistics, and intelligence analysis. There are reports of this already happening; the US and Israeli militaries have reportedly been using OpenAI's GPT-4 model to process intelligence data.[29]

Standards of reliability and safety are not the only roadblocks to the adoption of generative AI into weapon systems. There are also several developmental challenges that companies developing generative AI for weapon systems would have to overcome. These challenges primarily relate to the data used to train generative AI models.

Concerns about data poisoning would likely make military procurement agencies wary of acquiring generative AI based on general-purpose models trained on the entire internet. They would either want to see guarantees that data poisoning attempts can be detected or would be inconsequential regarding how the weapon systems would perform. Moreover, the military would want – and need – the system to be trained on military-specific data (e.g., data collected from intelligence, surveillance, or reconnaissance operations), given that certain tasks or capabilities demand information that is not available in the public domain. This, in turn, presents two very practical difficulties. The first is that the data would need to be labelled. Data labelling for model training is, in general, a labour-intensive endeavour. Large AI labs rely on a vast pool of remote workers to label the data used to train their general-purpose models. Data labelling in the military context is not a task that can be easily outsourced for obvious security reasons. Data labellers would likely need some form of security clearance, which can take a long time (e.g., weeks or even months) to obtain. Second, the labelling process may also require military expertise, which further restricts the pool of people who can label the data. These practical limitations represent a barrier for AI labs that want to develop generative AI for weapon systems, as well as a hurdle for military procurement agencies that want to accelerate the acquisition of generative AI capabilities by the armed forces.

4.2 Accelerating Factors

The technical and developmental hurdles listed above suggest that the integration of generative AI into weapons may take some time and will likely be done slowly and incrementally, where generative AI would be integrated into support functions like maintenance or navigation. However, one should not exclude the possibility that some actors will make different calculations and conclude that the benefits of the technology outweigh the limitations and will push for its adoption, even though the technology remains immature and does not meet the desired standards of safety and reliability. This risk scenario is not unique to generative AI; in fact, it applies to most emerging technologies. History provides several examples where states deployed brittle or insufficiently tested technologies for strategic or operational reasons.

[29] Sam Mednick et al., 'Israel's Use of Microsoft and OpenAI Raises Questions about What Could Go Wrong with the Powerful Tech', Fortune, 2025, https://fortune.com/2025/02/19/israel-microsoft-openai-raises-questions-powerful-tech/.

During the Cold War, the Soviet Union deployed some satellites that it knew had a high failure rate, but it was deemed strategically important to demonstrate that it possessed such a capability.[30] More recently, the war in Ukraine has shown how considerations for the responsible adoption of technology can be set aside. Ukrainian armed forces have had a relaxed approach to system certification since the war began, encouraging defence startups to propose AI innovations that could be quickly fielded.[31]

One should not exclude, given the current geopolitical context, that some actors would prematurely adopt generative AI capabilities either to showcase their position at the forefront of defence innovation or because they perceive real strategic or operational benefits.

Market pressure could be another factor that could lead to the adoption of generative AI in weapon systems. Nearly all the AI labs behind the most powerful AI models (Anthropic, OpenAI, Google) have recently indicated their willingness to market their technology for national security purposes.[32] OpenAI, through its partnership with Anduril, even signalled its openness to supporting the development of combat capabilities in weapon systems. This shift has garnered significant media attention, partly because these companies had long refrained from entering the defence market. The reasons behind this shift are debated, but one explanation often cited is that general-purpose models are expensive to train and operate. From this perspective, defence contracts represent a valuable source of revenue. If this is true, companies will likely seek to directly or indirectly promote the use of generative AI in the military domain, including in weapon systems. One should also not rule out the possibility that some companies will lobby for military procurement agencies to relax their requirements around safety and system certification.[33]

[30] David Hoffman, *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy* (Knopf Doubleday Publishing Group, 2009).
[31] Kateryna Bondar, 'Understanding the Military AI Ecosystem of Ukraine', 11 December 2024, https://www.csis.org/analysis/understanding-military-ai-ecosystem-ukraine.
[32] Frank Holmes, 'Silicon Valley Conquers The Pentagon As Defense Tech Explodes', Forbes, 2025, https://www.forbes.com/sites/greatspeculations/2025/02/20/silicon-valley-conquers-the-pentagon-as-defense-tech-explodes/.
[33] Maggie Gray and Max Dauber, 'Simplifying AI Deployment for Defense', 2024, https://maggiegray.us/p/simplifying-ai-deployment-for-defense.

5. What Should and Can Be Done

The outcome of a risk assessment process typically involves choosing between three options:

- **Option A: Accepting the risk fully**, essentially taking no action to address it.
- **Option B: Avoiding the risk**. This may require pausing or abandoning the development or deployment of the technology.
- **Option C: Reducing the risk through risk prevention and mitigation measures.**

Given the well-documented limitations of generative AI, it seems evident that Option A is not a viable choice. One could argue that Option B would be a reasonable path to pursue until methods for properly evaluating the reliability of generative AI have been developed. However, it is reasonable to assume that major military powers (such as the USA, China, and Russia) will deem Option B politically undesirable and unviable in the current geopolitical context. Even if they were to agree on a pause, they would likely harbour doubts about each other's commitments, given the "trust deficit disorder" (to use the words of the UN Secretary-General) that characterizes their relationships. For this reason, it seems most appropriate to focus on Option C, which entails the identification and implementation of risk mitigation measures.

In many regards, the integration of generative AI into weapons does not create fundamentally new risks. Rather, it exacerbates concerns that have long been discussed in the policy debates on autonomous weapon systems and the responsible military use of AI. Therefore, most, if not all, of the risk reduction measures discussed in the context of the Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE on LAWS) and the Responsible AI in the Military Domain (REAIM) summits remain relevant. These measures can be broadly categorized as technical, institutional, and policy measures (see Table 1).

Type	Measures

Technical	• Rigorous testing and evaluations to determine how the weapon systems may perform in the anticipated circumstances of its use
	• Place limits on the parameters of use of the weapons including
	  ○ The type of targets the systems can engage
	  ○ The duration, geographical scope and scale of operations
	  ○ Ensure that the mission parameters cannot be modified by the systems without appropriate human control and judgement
	• Conduct reviews to detect possible unwanted bias in datasets
	• Implement measures to reduce automation bias

Institutional	• Conduct legal reviews • Ensure appropriate training of human operators • Provide guidance to commanders and legal advisers regarding the circumstances under which the tool may or may not be used lawfully • Ensure internal mechanisms for the reporting of incidents that may involve violations of IHL
Policy	• Do not deploy or use the weapon system if the effects in attack cannot be anticipated and controlled, as required by international humanitarian law in the circumstances of use • Adopt and make publicly available national policy on the adoption of military AI

Table 1: Risk reduction measures proposed at the UN GGE on LAWS and REAIM.[34]

In addition to these general measures, several AI-specific measures could be considered, including:

- Ensuring that the integration of AI in weapon systems is associated with a well-defined use case and that thorough testing and evaluation are conducted for that specific use case.
- Establishing a repeatable validation process to ensure that the system remains fit for purpose, safe, and secure every time it undergoes a significant update.

[34] United Nations. *Rolling Text Draft of the Group of Governmental Experts on Lethal Autonomous Weapons Systems*. Status as of 26 July 2024. https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2024)/Rolling_text_draft.pdf; REAIM, 'REAIM Summit 2024', 2024, https://reaim2024.kr/home/reaimeng/board/reaim2024.kr.

6. Conclusion

The integration of generative AI into weapon systems presents both opportunities and risks. While generative AI has the potential to enhance the capabilities of weapon systems, there are significant concerns about its accuracy, reliability, and potential impact on human control, escalation, and accountability. A prudent approach would involve careful risk assessment, mitigation measures, and adherence to international norms and standards for the responsible development and use of AI in the military domain. This includes ongoing international policy discussions on responsible military use of AI, which provide an opportunity to reaffirm established good practices around safety, security, legal compliance, and accountability. These discussions have already led to the formulation of risk measures that could contribute to mitigating the risks associated with integrating generative AI into weapon systems and military systems more broadly.

About the Author

Dr. Vincent Boulanin

Dr Vincent Boulanin is Senior Researcher and Director of the Governance of Artificial Intelligence Programme at the Stockholm International Peace Research Institute (SIPRI). He leads SIPRI's research on the impact of artificial intelligence on peace and security. He has written extensively on issues related to the development, use and control of autonomy in weapon systems and military applications of artificial intelligence.

HCSS
Lange Voorhout 1
2514 EA The Hague

Follow us on social media:
@hcssnl

The Hague Centre for Strategic Studies