EU Non-Proliferation and Disarmament Consortium

Promoting the European network of independent non-proliferation and disarmament think tanks

NON-PROLIFERATION AND DISARMAMENT PAPERS

No. 92 January 2025

ARTIFICIAL INTELLIGENCE, NON-PROLIFERATION AND DISARMAMENT: A COMPENDIUM ON THE STATE OF THE ART

THOMAS REINHOLD, ELISABETH HOFFBERGER-PIPPAN, ALEXANDER BLANCHARD, MARC-MICHAEL BLUM, FILIPPA LENTZOS AND ALICE SALTINI

I. ARTIFICIAL INTELLIGENCE IN THE MILITARY DOMAIN: TECHNICAL, LEGAL AND ETHICAL PERSPECTIVES

THOMAS REINHOLD, ELISABETH HOFFBERGER-PIPPAN AND ALEXANDER BLANCHARD

Overview

Defence organizations are increasingly turning to artificial intelligence (AI) for achieving tactical and strategic advantages over their adversaries. The growing sophistication of AI technologies has accelerated their adoption for a number of tasks and functions, allowing different stakeholders to plan and accommodate their respective military operations. This adoption includes the direct integration of AI into weapon systems; decision-support systems, intelligence analysis or target recommendation systems; and external communication platforms. However, the development and deployment of military AI systems raise a number of significant legal and ethical challenges that will have to be met at various levels of responsibility as well as across the technology life cycle. This briefing paper summarizes these challenges. Developed for stakeholders within the military AI policy debate, the paper first outlines the current state of the art in AI technologies and their military uses, particularly with respect to conventional weapons. This includes the integration of AI into weapon systems and to facilitate the use of conventional weapons. Second, it sketches out key legal challenges and legal frameworks associated with the use of military AI in battlefield settings. Third, it looks at relevant ethical considerations and considers some initiatives undertaken by states to address these challenges.

SUMMARY

This multiauthored compendium offers a state-of-the-art summary of the artificial intelligence (AI) issues facing non-proliferation and disarmament. It pulls together four topics—Artificial Intelligence in the Military Domain: Technical, Legal and Ethical Perspectives by Thomas Reinhold, Elisabeth Hoffberger-Pippan and Alexander Blanchard (section I); Artificial Intelligence and Chemical Weapons by Marc-Michael Blum (section II); Artificial Intelligence and Biological Weapons by Filippa Lentzos (section III); and Assessing the Implications of Integrating AI in Nuclear Decision-making Systems by Alice Saltini (section IV)—that, taken together, offer a concise overview of the proliferation- and disarmament-related challenges and opportunities that AI presents.

Section I describes how military organizations increasingly use AI to enhance operational effectiveness in weapon systems, decision support and intelligence, and illuminates some of the critical technological, legal and ethical challenges posed by AI's integration into military organizations. Section II examines AI's impacts on chemical weapons, highlighting emerging risks from state and non-state actors, the need for regulation to prevent misuse and the importance of global collaboration to uphold norms against chemical warfare. Section III explores security concerns raised by the intersection of AI and biology with a specific focus on the risk that AI could facilitate the deliberate use of bacteria and viruses to inflict harm, emphasizing the need for a nuanced and evidence-based understanding of these risks. Finally, section IV examines AI integration into nuclear command, control and communications systems, noting its potential to enhance intelligence and situational awareness alongside significant risks of unreliability, cyber threats, and misaligned decision making, while calling for international dialogue and regulatory measures to avert catastrophic escalation.

The texts compiled in this compendium were originally prepared as briefs in support of four ad hoc seminars on AI and arms control for the European Union and its member states.

INTRODUCTION TO ARTIFICIAL INTELLIGENCE AND ITS MILITARY APPLICATIONS[1]

The different definitions of AI reflect the varying ways that intelligence itself is understood. This multifaceted characterization of AI aligns with the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, which was launched by the United States in 2023. According to this declaration, AI 'may be understood to refer to the ability of machines to perform tasks that would otherwise require human intelligence. This could include recognizing patterns, learning from experience, drawing conclusions, making predictions, or generating recommendations. An AI application could guide or change the behaviour of an autonomous physical system or perform tasks that remain purely in the digital realm.'[2]

From a technical perspective, AI covers a multitude of approaches (see figure 1). In the broadest sense, AI encompasses the set of all processes that mimic human capacities. The subset category *machine learning* refers to algorithms that can be used to derive knowledge-like relations or patterns within information. The further subset *deep learning* refers to algorithms that build on machine-learning concepts to simulate human brain cells (neurons) and their interconnections in *artificial neural networks* (ANN). The development over the past two decades of relatively cheap consumer electronics capable of performing enormous amounts of computing tasks simultaneously has enabled advances in ANNs. ANNs are now able to simulate thousands to millions of artificial neurons in extremely large and complex networks. While this is still far from the billions of neurons in human brains, these so-called *deep neural networks* (DNN) and the related technology of deep learning have enabled recent technological leaps in AI.

Current-generation ANNs must be trained towards the desired capability. This process generally consists of five steps:

1. Collecting vast amounts of specific data containing the information from which to learn.

2. Curating the data to reach statistical balances (e.g. to avoid biases).

3. Processing the data by specific deep-learning algorithms to create the *model* that represents the learned 'knowledge'.

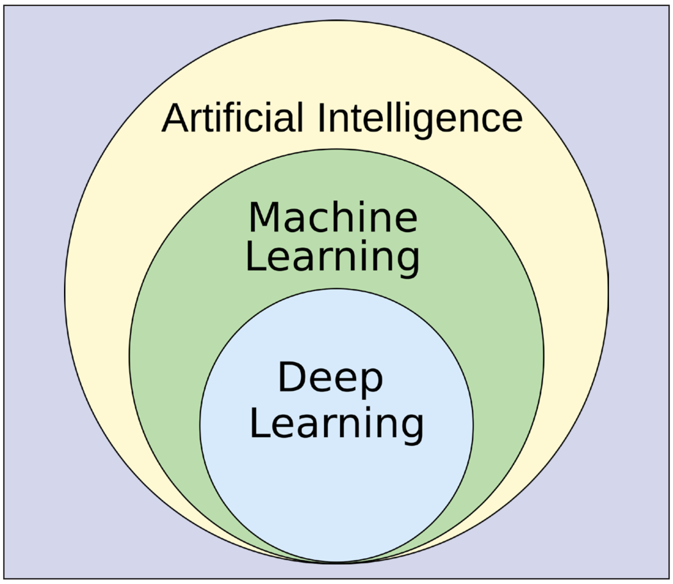

Figure 1. The hierarchy of artificial intelligence technologies

Source: 'AI hierarchy', Wikimedia Commons.

4. Testing the model with some of the originally collected training data to check its performance and accuracy and to correct possible learning errors.

5. Optimizing the overall performance; typically, this is done manually by human operators checking the response of the AI to specific inputs and then fine-tuning the model through feedback.

When put into actual use, the model itself is usually not altered further; however, for some applications, user feedback is used to retrain the model in a process known as *in situ training.* The model itself is usually considered a *black box,* and it is not possible to explain how the model has stored specific 'knowledge'. Some approaches to AI, like explainable AI (XAI),[3] try to extend the model to follow the input–output–processing step. However, approaches that seek to extend the model downgrade AI performance and cannot reach the explanatory potential of human accounts of decision making.

Examples of current-generation AI

While all current-generation AI relies on the technology of deep neural networks, it can serve very different application scenarios depending on training and optimization. Popular examples of current-generation AI can be grouped into four categories—generative AI, large language models (LLMs), large

[1] This section was authored by Thomas Reinhold.

[2] US Department of State, 'Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy', 9 Nov. 2023.

[3] IBM, 'What is explainable AI?', 2024.

multimodal models (LMMs) and foundation models—as described here:

Generative AI. AI that learns from digital images how to compose and create new images from a user's textual description (e.g. Midjourney[4] or DALL-e[5]) or short video snippets (e.g. Sora[6]).

LLMs. AI that is trained with textual input to learn the structure and sequence of written texts (e.g. ChatGPT-3.5[7] or CoPilot). From a user's instructions they can perform conversations in written form as well as create text of different styles, length and purpose.

LMMs. Trained to interact with users based on different media and create different media output, such as textual descriptions of images (e.g. ChatGPT-4). In contrast to former AI generations, LMMs are directly connected to information sources like the internet to collect additional input for the processing and generation of outputs.

Foundation models. Different from the above-mentioned AI systems, these foundation models are pre-trained for tasks like interpreting text or recognizing images. Customers then use their own data sets to finish training the model in a second step towards their specific application scenario.

Outlook for next-generation AI

A current trend is the integration of different kinds of media into LMMs to diversify the abilities of the AI model. Another emerging trend is the development of a market for AI whereby suppliers rent out or sell pre-trained foundation models and the required computing infrastructure, network technology and power supply. This allows customers to tailor the AI application to their needs, such as using sensitive training data. The next technological step that AI suppliers are envisioning is the arrival of so-called artificial general intelligence (AGI). Rather than being limited to one specific task, AGI could have capability to develop strategies motivated by 'goals' and 'reasons'. While still only a vision, AGI could also have knowledge about the world and its implicit rules, connections and relations. Generally, any steps in the extension of AI capabilities—like increasing the size of the artificial neural network or using more complex input training data—exponentially increases the cost of training and running the AI system and the availability of computing power. The demand for AI-enabled applications and systems is, therefore, strongly connected to geopolitical tensions about the microprocessor industry, the availability and restrictions of the necessary manufacturing materials, and the skills and technologies required for their production.[8]

The military applications of AI

Military actors primarily expect two things from AI. The first is a tactical advantage from the management and pre-processing of vast data sets (from surveillance and weapon systems, drones, satellite images, etc.) to enable human operators to achieve speedier and better decisions. For example, the US research project 'Convergence'[9] aims to reduce the 'sensor-to-shooter' time in battlefield management systems from 20 minutes to 20 seconds.

The Russia–Ukraine war has become a testbed for military AI applications that are used to monitor military manoeuvres, intercept and translate communications, and take decisions. AI decision-support systems are also being used by Israel in its war in Gaza, including the 'The Gospel',[10] a program for the automated aggregation and analysis of surveillance intelligence information used to preselect military targets. Beside AI applications on the battlefield, other processes like logistics or the maintenance of equipment via predictive maintenance management could be enhanced.

The second major driver of AI military applications goes hand in hand with the increasing use of autonomous (weapon) systems. These systems often need to operate in uncertain environments and be adaptable to different operating conditions. Additionally, AI enables autonomous systems to potentially operate over greater distances and time

[4] Midjourney, 'Midjourney', 2024.

[5] OpenAI, 'DALL·E 3', 2024.

[6] OpenAI, 'Creating video from text', 2024.

[7] OpenAI, 'GPT Chat 3.5', 2024.

[8] Kleinhans, J.-P. and Baisakova, N., 'The global semiconductor value chain: A technology primer for policy makers', Stiftung Neue Verantwortung, Oct. 2020.

[9] Strout, N., 'At second Project Convergence, US Army experiments with joint operations in the Arizona desert', C4ISRNET, 10 Nov. 2021.

[10] Davies, H., McKernan, B. and Sabbagh, D., '"The Gospel": How Israel uses AI to select bombing targets in Gaza', *The Guardian*, 1 Dec. 2023.

spans, or in communication-denied environments. This drive to develop AI extends to complex, large systems (e.g. the US project 'Loyal wingman'[11] or the European Future Combat Air System, FCAS[12]) that aim to include autonomous aerial fighting support. Also, the broad application of cheap, off-the-shelf consumer drones or loitering munitions increases the demand for autonomous navigation and image recognition capabilities, thus raising concerns of an unregulated application of AI in military systems.

INTERNATIONAL HUMANITARIAN LAW, HUMAN RIGHTS LAW AND THE IDEA OF MEANINGFUL HUMAN CONTROL[13]

Military AI raises a number of legal challenges, especially from the perspective of international humanitarian and human rights law. The obligation of states parties to the Additional Protocol I to the Geneva Conventions (AP I GC) to undertake weapon reviews is one of the most central legal norms related to military AI. However, the exact scope of a weapon review is far from clear, especially in cases where AI-enabled technology is used. In addition, targeting law (see below) shapes and restrains how military AI can be used responsibly and in line with legal but also ethical challenges. The USA's 2023 political declaration on AI and autonomy might also have repercussions on the relationship between humans and machines. Although it is not legally binding, the declaration reflects the state practice of those states that have endorsed it. This, in turn, might ultimately contribute to the formation of new norms of customary law if combined with *opinio juris* (i.e. the belief that an action was carried out as a legal obligation). Most importantly, the declaration calls on states to establish (technical) safeguards in order to ensure that military AI does not exhibit unintended behaviour. And, last but not least, the European Union's Artificial Intelligence Act (EU AI Act) could play a role regarding military AI, at least when it comes to dual-use technology.

The role of Article 36 AP I GC weapon reviews

States parties to AP I GC are obliged to undertake a legal review of all new weapons, means or methods of warfare. Furthermore, it is the prevailing view that, in general, Article 36 AP I GC is not reflective of customary law.

Are decision-support systems covered by Article 36 AP I GC?

The use of AI-enabled weapon systems in warfare pose a number of legal challenges. A legal review of an AI-enabled weapon system will most certainly fall under the purview of Article 36 in the case of autonomous target identification and engagement. The review itself would most probably have to include the software that was designed to perform the task of both target selection and engagement as well as the relevant hardware components, such as the relevant weapons platform, as well as sensors.[14] However, the legal situation of decision-support systems (DSS) in the context of Article 36 reviews is unclear. Some commentators argue that DSS should also fall under the purview of Article 36 AP I GC in case a DSS, inter alia, forms an integral part of military decision making, and in case the DSS 'poses a challenge to the application of humanitarian law'.[15] Other commentators argue that DSS are only covered by Article 36 AP I GC if states parties have explicitly decided to extend the scope of their reviews to DSS. Whenever states parties to AP I GC place the emphasis of their review on weapons only, DSS are, in their opinion, not covered by Article 36 AP I GC.[16] The Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy does not focus on *weapon* reviews explicitly, but it calls on states to undertake *legal* reviews in order to ensure that military AI-capabilities are employed in compliance with humanitarian law. The terminology used in the political declaration takes account of the fact that AI-enabled technology can play a role

[11] Fish, T., 'Uncrewed ambitions of the Loyal Wingman', Airforce Technology, 1 Nov. 2022.

[12] AIRBUS, 'Future Combat Air System (FCAS): Shaping the future of air power', 2024.

[13] This section was authored by Elisabeth Hoffberger-Pippan.

[14] Mimran, T. and Weinstein, L., 'The IDF introduces artificial intelligence to the battlefield: A new frontier?', Articles of War, Lieber Institute for Law & Warfare, 1 Mar. 2023.

[15] Klonowska, K., 'Article 36: Review of AI decision-support systems and other emerging technologies of warfare', *Yearbook of International Humanitarian Law*, vol. 23 (TMC Asser Press: The Hague, 2020), pp. 123–24. See also Copeland, D., Livoja, R. and Sanders, L., 'The utility of weapons reviews in addressing concerns raised by autonomous weapon systems', *Journal of Conflict and Security Law*, 2022, vol. 28, no. 2 (summer 2023), pp. 285–316.

[16] Meier, M. W., 'Responsible AI Symposium: Responsible AI and legal review of weapons', Articles of War, Lieber Institute for Law & Warfare, 27 Dec. 2022.

at different echelons in the military domain. AI can form an integral part of a weapon system and perform critical functions (such as target selection and engagement without the need for further human input), but it could also function as a DSS. The broad terminology employed in the political declaration clearly suggests that AI-enabled technology in the military domain should be reviewed whenever its use might have repercussions in terms of compliance with humanitarian law (and other international law as applicable).

In situ *learning algorithms and Article 36 AP I GC weapon reviews*

The use of *in situ* machine-learning mechanisms poses another legal challenge. One suggestion would be to continuously monitor such algorithms and/or undertake periodic post-deployment reviews.[17] However, *in situ* machine-learning algorithms pose a number of risks, especially with regard to predictability and explainability.[18] While predicting the modus operandi of *in situ* learning algorithms might work under 'laboratory conditions', dynamic environments would pose significant challenges—for example, potentially leading to unintended behaviour of AI-enabled weapon systems. By the same token, account should be taken of the fact that hitherto deep neural networks[19] cannot be fully understood from a technical perspective, which makes it even more difficult to determine whether *in situ* learning algorithms will exhibit unintended behaviour. In the light of these challenges, it is arguable that the use of *in situ* learning algorithms should be prohibited, at least when it comes to AI-enabled weapon systems that can identify and engage targets without the need for further human input.

Human rights and Article 36 AP I GC weapon reviews

When undertaking legal reviews within the meaning of Article 36 AP I GC, account should be taken of the fact that states must determine whether the use of a weapon system, means or methods of warfare would 'in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party'. It has been argued that 'any other rule of international law' refers, inter alia, to human rights law. Even though the extraterritorial application of human rights law in armed conflict is highly complex,[20] some commentators argue that especially the right to life within the meaning of Article 6 of the International Covenant on Civil and Political Rights (ICCPR)[21] does play a role when it comes to weapon reviews. In its General Comment No. 36[22] on the right to life, the United Nations Human Rights Committee (HRC) addressed the right to life in a military context. According to the HRC, Article 6 of the ICCPR 'invites preventive impact assessment measures, including legal reviews for new weapons'.[23]

Private industry and Article 36 weapon reviews

Private industry plays a substantial role in the development of new weapons, means or methods of warfare, including AI-enabled technology. Some commentators contend that states parties sponsoring private industry investigating the military use of technology are obliged to undertake reviews.[24] By the same token, it has been argued that states must review weapons that were produced by private industry for the purpose of being ultimately exported to other countries.[25] Article 36 AP I GC should, in the opinion of these commentators, be read in conjunction with Common Article 1 of the GC, according to which 'the High Contracting Parties undertake to respect and to ensure respect for the present Convention in all circumstances'.[26]

Targeting law: The central role of the obligation to take precautions in attack

Targeting law—especially the principle of distinction according to Articles 48, 51(2) and 52(2) AP I GC,

[17] McFarland T. and Assaad, A., 'Legal reviews of *in situ* learning in autonomous weapons', *Ethics and Information Technology*, vol. 25, no. 9 (2023), pp. 1–10.

[18] McFarland and Assaad (note 17).

[19] According to S. J. Pawan and Jeny Rajan, 'deep neural networks (DNNs) comprise multiple non-linear computational units or neurons organized in a layer-wise fashion to extract high-level, deeper, robust, and discriminative features from the underlying data'. See Pawan, S. J. and Rajan, J., 'Capsule networks for image classification: A review', *Neurocomputing*, vol. 509 (14 Oct. 2022), pp. 102–20.

[20] For a comprehensive overview see, inter alia, Oberleitner, G., *Human Rights in Armed Conflict: Law, Practice, Policy* (Cambridge University Press: Cambridge, 2015).

[21] International Covenant on Civil and Political Rights, New York, 16 Dec. 1966, in force 23 Mar. 1976, United Nations Treaty Collection, vol. 999, no. 14668.

[22] United Nations, Human Rights Committee, General comment no. 36, CCPR/C/GC/36, 3 Sep. 2019.

[23] Mimran and Weinstein (note 14).

[24] Copeland, Livoja and Sanders (note 15), pp. 306–307.

[25] Copeland, Livoja and Sanders (note 15), pp. 306–307.

[26] Copeland, Livoja and Sanders (note 15), pp. 306–307.

and the related prohibition of indiscriminate attacks according to Article 51(4) AP I GC, as well as the principle of proportionality pursuant to Article 51(5) b AP I GC, and the obligation to take precautions in attack pursuant to Article 57 AP I GC—regulates the use of force, including scenarios where military AI is used. The provisions within targeting law are reflective of customary law and apply in both international armed conflict as well as non-international armed conflict.[27]

When it comes to military AI, the obligation to take precautions in attack plays a particularly central role. Parties to an armed conflict are obliged to take constant care and, as far as possible, to spare civilians from the dangers arising from military activities. This includes intelligence gathering but also data collection and management activities 'as long as these activities are intended to advance combat'.[28]

The political declaration on AI: From meaningful human control to technical safeguards

The aforementioned Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy has, thus far, been endorsed by 52 states.[29] The declaration consists of 10 principles and is not legally binding. However, its endorsement certainly is reflective of state practice of those states that have endorsed it. This, in turn, could theoretically contribute to the formation of customary law if combined with the respective element of *opinio juris*.[30] As such, Principle J—which declares that 'States should implement appropriate safeguards to mitigate risks of failures in military AI capabilities, such as the ability to detect and avoid unintended consequences and the ability to respond, for example by disengaging or deactivating deployed systems, when such systems demonstrate unintended behaviour'—deserves greater attention.[31] The principle's reference to 'safeguards' in order to mitigate risks is telling. Until now, the term 'meaningful human control' in the context of autonomous weapon systems (AWS)[32] was used to describe the relationship between humans and machines. The reference to safeguards in the declaration seems to herald a shift in terminology. The term 'safeguards' provides more tangible guidance on how to guarantee (also from a technical perspective) that military AI is used responsibly and in line with legal and ethical considerations.[33]

The role of the EU AI Act: Applicability to dual-use technology?

Another issue deserving greater attention is the so-called EU AI Act.[34] On 13 March 2024, the European Parliament approved the AI Act with 523 votes in favour, while 46 members of the parliament voted against and 49 abstained. The EU AI Act is the first legal framework within the EU that addresses the various risks associated with AI. The AI Act identifies four levels of risk: minimal risk, limited risk, high risk and unacceptable risk. It is questionable whether the EU AI Act would apply to dual-use technology.[35] While the EU AI Act would not apply to goods exclusively for the military, there are, in fact, no convincing legal arguments that would exclude dual-use technology from the EU AI Act's scope of application.[36] It is very likely that regulatory efforts by the EU in the civilian realm will, at least indirectly, also influence military AI. It should also be noted that the European Defence Fund only supports defence projects where meaningful human control can be maintained over AWS in case such systems carry out strikes against humans.

[27] International Committee of the Red Cross, Customary IHL Database.

[28] Lubin, A., 'Lieber studies big data volume—Algorithms of care: Military AI, digital rights and the duty of constant care', Articles of War, Lieber Institute for Law & Warfare, 13 Feb. 2024.

[29] US Department of State (note 2).

[30] Nasu, N., 'Nova 2, Legion X, and the AI Political Declaration', Articles of War, Lieber Institute for Law & Warfare, 27 Nov. 2023

[31] US Department of State (note 2).

[32] Other terms that have also been used are, inter alia, 'human judgment or control' or 'appropriate levels of human judgment'. See e.g. US Department of Defense, Directive 3000.09, Autonomy in weapon systems, updated version 25 Jan. 2023.

[33] Nasu (note 30).

[34] Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts, 2021/0106(COD).

[35] Regulation (EU) 2021/821 of the European Parliament and of the Council of 20 May 2021, setting up a Union regime for the control of exports, brokering, technical assistance, transit and transfer of dual-use items, 11 June 2021, L 206/1.

[36] See, most importantly, preambular paragraph 12(a) of the EU AI Act Proposal.

THE ETHICS OF MILITARY ARTIFICIAL INTELLIGENCE[37]

Military AI raises a number of ethical concerns, including concerns about harm to civilians, difficulties for exercising human judgement, desensitization to the act of killing, and concerns by some that AWS have the potential to violate human dignity. However, the place and scope of ethics in the international policy debate have been unclear, including uncertainty about the relevance of ethics-based argumentation to relevant legal frameworks. Moving forward, more research is required to address this matter. Many states have undertaken a number of initiatives to account for military AI ethical considerations, including the adoption of military AI ethics principles. As a result, a number of militaries are now turning their attention to translating these principles into practice. However, to do this, governance frameworks must be established.

What is ethics?

Ethics is the study of moral phenomena. It investigates what people ought or ought not to do, and what justifications can be given for such claims. Since AI is a digital technology, it falls under the branch ‘digital ethics’, which studies and evaluates moral problems related to information and data, algorithms, and corresponding practices and infrastructures—so as to formulate morally good solutions for digital technologies.

Digital ethics, digital governance and digital regulation are complementary but not the same. Digital governance is the practice of establishing and implementing policies, procedures and standards for the proper development, use and management of digital technologies. Digital regulation is a system of rules elaborated and enforced through social or governmental institutions to regulate use of technology. Ethics is complementary to regulation because it informs ‘what ought or ought not to be done over and above existing regulation’.[38]

The ethics of military AI draws on digital ethics, and ethics more broadly, making use of perspectives that have been debated for centuries. When considering adversarial kinetic uses of AI, the ethics of military AI is informed by different traditions for the ethics of war, particularly ‘just war theory’, which also provides the philosophical basis for international humanitarian law (IHL).

Ethics in international policy debate: Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Execution

Ethics gained prominence in the international policy debate around military AI with the 2013 publication of the Report of the UN Special Rapporteur on Extrajudicial, Summary or Arbitrary Execution by Christof Heyns.[39] While the context of the report was the use of drones for targeted killing, Heyns focused on the increasing integration of algorithms into this activity. The report highlighted a number of ethical concerns about such ‘automated killing’, including the potential to detach and desensitize humans from the decision to kill; blurring the distinction between ‘weapons and warriors’; reducing the cost threshold (broadly conceived) of going to war; the lack of contextual understanding required to respect ethical (and legal) principles; and that the delegation of life or death decisions to automated processing represents a form of arbitrary execution.

The ethical concerns in Heyns’s report drew on a mix of ethics and legal-based considerations, illustrating the interrelation of the two. However, while Heyns’s report provided the contours of much of the ensuing policy debate on AWS, the role and scope that ethics ought to have in this debate remains unclear. As stakeholders return to ethics to advance the international governance of AI, returning to Heyns’ recommendations—in particular, the recommendation to convene a panel of ‘ethics and philosophy’ experts to explore ethical and policy issues—could pay dividends. In addition, international policy debate could benefit from research on the relationship between ethics and legal frameworks such as IHL and international human rights law (IHRL).

Ethics at the UN CCW GGE: An uneasy home

In 2014, the High Contracting Parties (HCPs) of the 1981 Convention on Certain Conventional Weapons (CCW) convened an informal meeting of experts to

[37] This section was authored by Alexander Blanchard.

[38] Floridi, L., ‘Soft ethics, the governance of the digital and the General Data Protection Regulation’, *Philosophical Transactions of the Royal Society*, vol. 376, no. 2133 (28 Nov. 2018).

[39] United Nations, Human Rights Council, Report of the Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions, Christof Heyns, A/HRC/23/47, 9 Apr. 2013.

discuss questions related to emerging technologies in the areas of lethal autonomous weapon systems. There, the idea of 'meaningful human control' emerged as a possible point of common ground for addressing ethical and legal issues associated with AWS and, with the 2016 establishment of the group of governmental experts (GGE) by the CCW HCPs, 'human control' came to act as a proxy for discussions on ethics.[40] However, with the CCW being an instrument of IHL, the GGE has not always been a sympathetic forum for ethics-based concerns. For instance, the GGE's mandate directs delegations to draw on legal, military and technological expertise, but not ethics. Moreover, because for many years AWS has been the centre of gravity for international policy debate on military AI, and because that debate is based in the framework of the GGE, ethics-based argumentation is underdeveloped relative to IHL-based argumentation.

One enduring ethics-based objection to AWS is that they violate human dignity. The objection entails a number of interrelated but distinct claims. Two are key. First, that machines cannot deliberate about the gravity of taking life, and so the delegation of such decisions to machines represents a form of arbitrary execution. Second, that AWS use generalized target profiles in attacks, and thereby dehumanize people by reducing them to data points. The human dignity argument applies regardless of (*a*) whether combatants or non-combatants are targeted by AWS and (*b*) the technological maturity of AWS. Providing a constant, that AWS violate human dignity is a mainstay argument among some campaign groups opposed to AWS. It has also been invoked by state delegates at the GGE and international organizations as a basis to limit certain uses of AWS. For instance, in its position paper on AWS, the International Committee of the Red Cross (ICRC) ruled out anti-personnel AWS based on the human dignity argument.[41] However, some stakeholders have expressed scepticism at the human dignity argument, particularly at the assumption that AWS violate human dignity in a way other weapon systems do not. Debate on this issue would benefit from further research, particularly for clarifying the way different legal and ethical elements inform the argument.

State approaches to the ethics of military AI: A principles-based approach

States have sought to account for ethical concerns about military AI through a number of initiatives. One approach has been the development of sets of principles to guide the development and deployment of military AI in conformity with moral frameworks. This follows an approach that has been taken more generally to address concerns raised by AI in the civil domain. As of 2023, 26 states have adopted national sets of civilian AI ethics principles.[42] The number of states adopting specifically national military AI ethics principles is far smaller than this. Currently only the US Department of Defense (DOD) and the British Ministry of Defence (MOD) have published officially adopted principles. At the intergovernmental level, the North Atlantic Treaty Organization (NATO) presented its 'principles of responsible use' in its AI strategy, which was formally adopted by the allied defence ministers in 2021. In adopting these principles, the allied states committed to ensuring that their development and deployment of military AI accords with the principles of lawfulness, responsibility and accountability, explainability and traceability, reliability, governability, and bias mitigation.[43]

Principles to practice: The ethical governance of military AI

The principles-based approach to the ethical challenges presented by AI has been criticized as too high-level to provide directives for morally correct action. However, high-level principles can be seen as having a status much like constitutional principles and, as with constitutional principles, they require interpretation to provide directives for morally correct action in specific concrete cases. As NATO recognizes in its AI strategy, 'the task now turns to translating them [the principles of responsible use] into principled action.'

[40] Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have indiscriminate Effects, Report of the 2014 informal meeting of experts on Lethal Autonomous Weapons Systems (LAWS), CCW/MSP/2014/3, 11 June 2014.

[41] International Committee of the Red Cross, 'ICRC position on autonomous weapons systems', 12 May 2021.

[42] Anand, A. and Deng, H., 'Towards responsible AI in defence: A mapping and comparative analysis of AI principles adopted by states', Research Brief, UNIDIR, 13 Feb. 2023.

[43] Stanley-Lockman, Z. and Christie, E. H. 'An Artificial Intelligence Strategy for NATO', NATO Review, 25 Oct. 2021.

So far, very little practical guidance exists for translating military AI ethics principles into practice, or indeed for enabling the ethical development and deployment of military AI more broadly. But there have been some efforts. In 2021 the US DOD's Defense Innovation Unit (DIU) published 'Responsible AI Guidelines in Practice', which provides a flow-diagram of questions that relevant personnel ought to address to ensure the development and deployment of AI systems align with the DOD's responsible AI principles. Australia's Defence Science Technology Group (DSTG) published 'A Method for Ethical AI in Defence', which similarly provides a flow-diagram of questions and prompts, as well as specific tasks, to guide personnel involved in the development and deployment of military AI systems. Such guidance contributes to delineating an ethical institutional attitude towards the adoption of military AI. However, some consider question-based flow-diagrams to be limited because responsibility for making complex ethical assessments is devolved to practitioners who may lack necessary expertise. This means that decisions about what is ethically acceptable are left to local decision makers (e.g. procurement personnel, developers, operators), with the burden for the responsible use of AI shifted from institutions to sole individuals or groups of individuals.[44] To complement these approaches, the (successful) ethical development and deployment of military AI requires institutional frameworks for ethical governance, comprising clear responsibilities, tools, mechanisms, shared standards, processes and, some have argued, ethics review panels with AI expertise.[45]

II. ARTIFICIAL INTELLIGENCE AND CHEMICAL WEAPONS

MARC-MICHAEL BLUM

Overview

This brief explores the likely impacts of artificial intelligence (AI) systems on the development, production and use of chemical weapons. While this topic is not currently the main focus of the discussions about AI risks related to weapons of mass destruction (WMD), the brief explains why an in-depth assessment of emerging chemical weapons risks is essential and should be based on a reappraisal of the potential uses of chemicals in warfare, at a time when the global norm against such use appears to be eroding.

The brief outlines the varying risks posed by state and non-state actors, the mitigation of which will require different approaches from policymakers. Policymakers will also have to strike a fine balance when designing and implementing regulation on AI so as not to disrupt the desired progress of the technology for benign purposes. The brief calls for close monitoring of the industry, especially in terms of AI use in chemistry, with particular attention placed on small and medium-sized enterprises, which are perhaps at a higher risk from malign actors than larger, established companies. It recommends preparing an amended regulatory framework while putting any immediate implementation on hold until a more complete picture of the risks and opportunities stemming from AI has emerged. In addition, it proposes that the European Union (EU) and its member states should take direct action to support the Chemical Weapons Convention and other control regimes such as United Nations Security Council Resolution 1540 and the Australia Group.

AN INTRODUCTION TO AI AND CHEMICAL WEAPONS RISKS

Among the different classes of WMD, chemical weapons have been used, at different scales, most often in recent years. Notable examples include the use of nerve agents for targeted assassinations, as in the attempted poisonings of Sergei Skripal in 2018 and Alexei Navalny in 2020; the use of sulfur mustard by the Islamic State (IS) in Syria in 2015; the use of sarin and chlorine by the Syrian government on numerous occasions during the Syrian civil war; and the alleged

[44] Blanchard, A., Thomas, C. and Taddeo, M., 'Ethical governance of artificial intelligence for defence: Normative tradeoffs for principle to practice guidance', *AI and Society* (21 July 2023), pp. 1–14.

[45] Taddeo, M., Blanchard, A. and Thomas, C., 'From AI ethics principles to practices: A teleological methodology to apply AI ethics principles in the defence domain', *Philosophy and Technology* (13 Mar. 2024), pp. 1–21.

use of irritants such as tear gas and chloropicrin by Russia in the war in Ukraine.[1] Important contributing factors for this are the relatively low technological barriers to production and the easy accessibility of some chemical agents and toxic industrial chemicals. In addition, some states still retain knowledge and expertise from chemical weapons programmes that were only recently terminated or might even continue to exist at a clandestine level to this day.[2]

Despite the comparatively frequent use of chemical weapons during the past few years, they have not been the main focus of recent public discussions about the risks of AI with respect to WMD. These discussions have instead largely focused on biological weapons.[3] For example, in early 2024 OpenAI, the creator of the chatbot ChatGPT, published a study evaluating the risk that a large language model (LLM) could aid experts and non-experts to produce a biological weapon and carry out a biological attack.[4] A separate study, conducted by the RAND Corporation and published in 2024, also explored the feasibility of exploiting LLMs for biological attacks.[5] Furthermore, in its most recent system card, released in September 2024, OpenAI evaluated only the potential biological risks related to its o1 LLM, stating that it focuses its 'CBRN [chemical, biological, radiological and nuclear] work on biological threat creation because this is the area of catastrophic risk with the lowest barriers to entry'.[6] This statement is highly debateable and probably wrong.

Notably, to date, there are no publicly available studies aimed at assessing the risks posed by AI with regard to chemical weapons. However, one recent study did at least touch on this area and received considerable attention from experts, the media and the general public. The study was conducted by the drug development company Collaborations Pharmaceuticals in cooperation with the Spiez Laboratory (a Swiss federal institute that develops measures for protection against chemical, biological and nuclear threats) and King's College, London.[7]

In this study, as part of a thought experiment, the generative AI model used by the pharmaceutical company to develop new drug molecules with low toxicity was instead used to design new highly toxic molecules based on a template for the chemical nerve agent VX. The AI model identified 40 000 virtual molecules, most of which had not been previously identified and some of which had a predicted toxicity level higher than VX itself. The study did not publish the generated data set of chemicals for security reasons (making it impossible for others to conduct further work on the actual toxicity of the generated compounds), but the speed with which the model identified such a large number of compounds was a clear cause for concern.

While the study conducted by Collaborations Pharmaceuticals indicates that new, potentially toxic molecules could be readily identified using AI, some authors have suggested that the development, production and testing of such molecules would be unreliable and indeed unnecessary, probably amounting to a waste of resources since currently available chemical agents are already sufficiently toxic and relatively easy to synthesize.[8] These likely drawbacks would deter many actors, and especially non-state actors, from pursuing this path to producing a chemical weapon, which highlights why there is currently a debate as to the relevance of AI risks with respect to chemical weapons. However, it is important to consider that although the identification of new toxic chemicals would perhaps be the most obvious application of generative AI in this area, there are many other possible applications that could support the successful use of chemical weapons. Examples include the development of methods for the dissemination and dispersal of toxic chemicals and of formulations and mixtures to optimize physical properties that affect

[1] See e.g. Dewey, K., 'Poisonous affairs: Russia's evolving use of poison in covert operations', *Nonproliferation Review*, vol. 29, nos 4–6 (2022); Strack, C., 'The evolution of the Islamic State's chemical weapons efforts', *CTC Sentinel*, vol. 10, no. 9 (Oct. 2017); Schneider, T. and Lütkefend, T., *Nowhere to Hide: The Logic of Chemical Weapons Use in Syria* (Global Public Policy Institute: Berlin, 2019); and Radchenko, O. M. et al., 'Lessons learned from the full-scale invasion of Russia: Injuries by chemical warfare agents with suffocating-irritating action (own clinical observation)', *Ukrainian Journal of Military Medicine*, vol. 5, no. 1 (2024).

[2] Burge, T., 'Russia's clandestine chemical weapons programme: The Bellingcat exposure', Royal United Services Institute (RUSI), 3 Dec. 2020.

[3] In the nuclear field, the discussions regarding AI have mainly centred on command and control issues and not so much on weapon creation. See e.g. Johnson, J., *AI and the Bomb: Nuclear Strategy and Risk in the Digital Age* (Oxford University Press: Oxford, Feb. 2023).

[4] OpenAI, 'Building an early warning system for LLM-aided biological threat creation', 31 Jan. 2024.

[5] Mouton, C. A., Lucas, C. and Guest, E., *The Operational Risks of AI in Large-scale Biological Attacks: Results of a Red-team Study*, Research Report (RAND Corporation: Santa Monica, CA, 2024).

[6] OpenAI, 'OpenAI o1 System Card', 12 Sep. 2024.

[7] Urbina, F. et al., 'Dual use of artificial-intelligence-powered drug discovery', *Nature Machine Intelligence*, vol. 4 (Mar. 2022).

[8] Blum, M., 'No chemical killer AI (yet)', *Nature Machine Intelligence*, vol. 4 (June 2022).

persistence and the ability to penetrate protective equipment. Moreover, AI could potentially be used to create weapons that circumvent existing medical countermeasures or export control mechanisms by exploiting alternative precursor chemicals through new synthetic routes.

While there are many potential applications for AI in the field of chemical weapons, there are as yet no real-life cases or examples on which to base an assessment of how realistic the risks are. This has consequences for targeted policy options. With that in mind, this brief next explores who might employ AI to assist in the development or use of chemical weapons. It then assesses the types of use or misuse that might be of relevance and finally sets out a list of recommended actions for European policymakers.

RISKS POSED BY STATE AND NON-STATE ACTORS

When exploring AI risks related to chemical weapons, there is a need to determine not only what might be technologically possible, but also who might be interested in and able to use these new technological possibilities. It is important to differentiate between state and non-state actors when trying to make such an assessment.

Non-state actors

Non-state actors need to operate clandestinely, have relatively limited resources and might lack the necessary expertise to make, handle and use toxic chemicals. As pointed out above, it is unlikely that they would devote their scarce resources to developing a highly sophisticated new chemical agent that has to be produced, weaponized and tested. What AI can do, however, is to lower the technological and knowledge burden to use chemical weapons effectively. The threat here would be an AI system that would give instructions and act as a guiding hand. This is the kind of scenario that the above-mentioned OpenAI and RAND studies explored for biological weapons.

It is important to keep in mind that a chemical attack does not end with the successful production of a chemical agent; these agents also need to be accurately dispersed. IS had significant resources and was able to produce sulfur mustard, albeit of a relatively poor quality; however, it was unable to employ this agent with any meaningful success because it failed to weaponize the agent properly and lacked an adequate use doctrine.[9]

The challenge of weaponizing a chemical toxin is also exemplified in a German case from 2018.[10] A supporter of IS, with no previous scientific or laboratory experience, managed to produce small amounts of the plant toxin ricin. He received instructions on ricin purification from contacts in Tunisia through an online messaging platform. In principle, an AI system trained with internet data could have provided similar instructions. However, this case shows that the ability to synthesize a toxic agent is not in itself sufficient to produce a chemical weapon. The individual, who was arrested before he could carry out an attack, planned to coat ball bearings with the ricin and embed them into an explosive vest. It is questionable whether much of the ricin would have remained intact and biologically active after the explosive blast.

While it would be relatively easy to block systems like ChatGPT from providing instructions to non-state actors on how to produce and disseminate chemical agents in response to direct questions, it would be difficult to prevent them from sharing relevant information in other situations. As with biological knowledge, chemical knowledge is inherently dual use in nature, meaning that it can have both civilian and military applications. If, for example, a user were to ask an AI system to provide an innovative solution to spray pesticides on farmland to maximize effect while minimizing the amount of pesticide used, it is unclear how the system would be able to determine whether the user was really a farmer looking to optimize pesticide use or a potential terrorist looking for ways to disseminate a chemical agent. In this context, the problem is not only that it might be difficult to set up an AI system to differentiate between benign and malicious intent but also that the AI system could go beyond the capabilities of a standard internet search engine to generate innovative solutions based on its training set of data—solutions that would not be revealed by a normal internet search.

State actors

Potential chemical weapons risks posed by the use of AI by state actors are very different from those posed by non-state actors. As with non-state actors,

[9] Strack (note 1).

[10] Flade, F., 'The June 2018 Cologne ricin plot: A new threshold in jihadi bio terror', *CTC Sentinel*, vol. 11, no. 7 (Aug. 2018).

states need to develop and produce chemical weapons clandestinely because almost all UN states are parties to the 1993 Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on their Destruction (Chemical Weapons Convention, CWC); however, a state can conceal such weapons programmes far more easily than a non-state actor and can act without interference from domestic law enforcement agencies. For example, a state could operate shell companies purportedly to use AI for drug development while secretly operating a chemical weapons development programme. This creates difficulties for export control regimes and could become an even bigger problem in the years to come. Export control regimes would somehow have to try to monitor transfers involving materials, technology and know-how that might appear completely benign in nature but that could be used for clandestine chemical weapons development. These transfers might not trigger any of the usual alarms that would cause authorities to look more closely at a specific transaction or at the entities involved.

State programmes become more difficult to hide as they grow in size. Limited amounts of a chemical agent—in the range of a few grams or kilograms—required for assassinations or sabotage can be produced with a small and relatively easy-to-hide footprint, but the production of chemical weapons for use in warfare is far more challenging. Once produced, the agents need to be placed in special munitions, which must be stored. Troops also need to be trained to use the weapons. It is at these stages that the CWC's verification toolkit can be used effectively. The process of developing and producing chemical weapons, whether or not it is supported by AI, is a violation of the convention.

Even if states choose to restart or maintain clandestine chemical weapons programmes in violation of the CWC, the times of large-scale chemical warfare involving the use of hundreds or thousands of tonnes of agent, as envisioned during the cold war, are probably gone for good, unless the norm against the use of chemical weapons erodes completely. More likely scenarios include small-scale use at points of attack or for sabotage operations—uses where plausible deniability would be possible and detection could be difficult.

Exploring these new potential use scenarios is crucial because they will also inform possible AI use scenarios. A state's use of toxic chemicals in warfare in clear violation of the CWC would be possible only if treaty provisions are enforced loosely and the norm against use is slowly eroded. In this context, it is important to keep in mind that the CWC is enforced by the states parties themselves. This means that states are perhaps more likely to use agents with low levels of lethality, such as irritants and incapacitants, in warfare, rather than highly toxic agents. This has already been seen in the case of the alleged use of tear gas and chloropicrin by Russia in the war in Ukraine.[11] States may therefore have a growing interest in developing new agents falling into this category and AI may play an important role in such development efforts, which could be disguised as development of riot control agents (not banned under the CWC unless used in warfare) or new pharmaceuticals (e.g. fentanyl and related compounds) that could be used as incapacitating agents.

USE AND MISUSE OF AI

Vijay Pande, a professor at Stanford University, California, and a general partner at the venture capital firm Andreessen Horowitz, is the founder and leader of a fund investing in the life sciences and healthcare with more than $3 billion under management. In 2023 he published an op-ed in the *Wall Street Journal* titled 'AI is a healer, not a killer', in which he stated: 'We shouldn't worry about the ability of artificial intelligence to create bioterrorism weapons. Restricting AI's ability to understand biology and chemistry would pose a far greater danger.'[12] This followed a post by Pande on the social media platform X, formerly known as Twitter, criticizing the above-mentioned Collaboration Pharmaceuticals study: 'They're forgetting that it's actually quite easy for molecules to be toxic (that's partly why drug design is so hard), and so AI doing so isn't anything new or scary.'[13]

Why are these statements relevant? First of all, Pande's argument that it is easy for molecules to be toxic is misleading because toxicity is relative. Chemical weapons are compounds with exceedingly high toxicity combined with physical properties that allow weaponization. Very few (under 10) chemical agents were produced in significant quantities and

[11] Radchenko et al. (note 1).

[12] Pande, V., 'AI is a healer, not a killer', *Wall Street Journal*, 27 Dec. 2023.

[13] Vijay Pande (@vijaypande), X, 12 Nov. 2023, <https://x.com/vijaypande/status/1723800648515502189>.

stockpiled by former chemical weapons possessor states during the cold war, suggesting that designing effective chemical weapons is possibly as challenging as designing effective drugs. More importantly, such statements indicate that actors in academia and the financial and technology industries as well as in the chemical and pharmaceutical industries tend to believe that the potential benefits from the use of AI in chemistry, and especially in drug development, outweigh any potential risks, meaning that progress in this field should not be slowed by concerns and regulation.

Not all civilian stakeholder voices are as outspoken as Pande regarding the relevance, or rather non-relevance, of risk, but they appear to share the common view that AI-based methods and technologies will be game changers for chemistry and the life sciences. While they may be wrong to downplay the risks, they are probably correct in their assessment that AI-based techniques have the potential to revolutionize these fields, including their respective industries and especially the pharmaceutical industry. And it is the inherent dual-use nature of chemical and biological technologies that makes developments originating from benign, civilian efforts the likely starting point for misuse. As pointed out above, chemical weapons programmes need to be conducted in a clandestine fashion. Recruiting and retaining the necessary expertise and maintaining the required funding to compete effectively with the civilian sector, including not only dedicated AI companies but also pharmaceutical companies, would be extremely difficult under these restrictions. Exploiting civilian advances would be the easiest and most sustainable way to develop new chemical weapons; for non-state actors, it would be the only way.

Such exploitation efforts could take various forms. At the domestic level, governments could compel civilian companies to make technological advances available to them; in other cases direct state funding to civilian companies with subsequent transfer of know-how might be used. States might also look to exploit technology developed in foreign countries through measures ranging from theft (via cyberattacks, intrusion or insiders) to seemingly benign scientific cooperation between academic institutions. These practices are not new and are already happening in other technology sectors. The key difference between those sectors and the AI sector is that a lot of new companies are currently active in this area. Start-ups in their early stages might be especially vulnerable to exploitation. They have to focus on their business and technology platform and might not have the resources (financially or in the form of qualified personnel and know-how) to defend themselves effectively against malign actors. In addition, many of these start-ups will not survive commercially in the medium to long term. Such companies could be bought by malign actors with the aim of exploiting intellectual property or keeping the business going as a front company secretly working for a chemical weapons programme, perhaps even without the knowledge of the company's staff. These various exploitation efforts have implications not only for policymaking but also for export control and sanctions evasion monitoring.

RECOMMENDED ACTIONS

Understand the risks

It is of the utmost importance to understand the risks the EU and the wider international community are facing with regard to AI and chemical weapons. As noted above, the current methodology to assess risks remains underdeveloped. For example, red teaming—a process where a group pretends to be a malign actor attempting intrusion and exploitation of a digital or physical entity—is well established in the information technology sector, but is still in its infancy with respect to AI use in a CBRN setting. As a result, knowledge about how AI can be used to make, or enhance the capability of, chemical weapons is still largely based on thought experiments or small studies that address only a fraction of the actual risk spectrum.

Therefore, it would be advisable first to fund studies using different approaches to assess AI-related risks with the aim of developing reproducible and verifiable standards to rate those risks. Academia, state research institutions and the private sector could all help to develop these methods and tools. Different tools to finance such research and development efforts exist within the EU and should all be used. The main route would be through the Horizon Europe programme but other instruments such as the Digital Europe Programme, the EU4Health programme, the European Defence Agency and the European Structural and Investment Funds (e.g. the Cohesion Fund) could be used as well. Most importantly, these research projects must be fast paced to match the speed of change in the AI sector and the results must be exploited efficiently.

Initiating three-year research projects followed by one or two years of implementation would be too slow: the results would be out of date by the time the projects were completed. Such projects also need to be run on a continuous basis to keep up with the rapid evolution of AI technology and related possible use scenarios for chemical weapons.

Understand the uses

Together with the need to understand how AI can influence the development, production and use of chemical weapons, there is an equal need to understand the future use scenarios for such weapons. Under what circumstances, for example, might states consider engaging in a large-scale chemical war? When would states be willing to erode the norm against chemical weapons and accept their re-emergence? What new use scenarios are likely that might be on a smaller scale, be more targeted and come with plausible deniability? These are just a few of the questions that need to be addressed. AI developments might influence the reasoning of states and non-state actors with respect to these questions. For example, malign actors would almost certainly be influenced by the way in which the norm against chemical weapons is enforced, including against seemingly lesser violations such as the use of riot control agents in warfare.

The study of the future of chemical warfare is interdisciplinary in nature and requires input not only from chemistry and the life sciences (due to the continuing convergence of the two fields), but also from conflict studies, military strategy and technology, international relations and international law. Efforts in this area will yield meaningful results only if all relevant fields find cooperative approaches to working on this issue. This is by no means trivial as it demands that scientific fields with very different cultures and ways to address research problems come together and find common ground. Such approaches should be specifically targeted for funding support through the different channels offered by the EU.

Monitor the industry

Monitoring industry developments in AI use in chemistry is important for several reasons and not just to spot specific emerging applications that could have a significant impact in the area of chemical weapons. Monitoring patent applications, mergers and acquisitions, major financing rounds for start-ups and products launched on to the market would help policy stakeholders to understand where the technology stands and what is, and what is not, possible. The next technological advances, including potential enablers for the production and use of chemical weapons, will be informed by current developments, methods and products. Trying to look further into the future—beyond the next few steps—makes the picture become more blurry and forecasting less reliable.

As well as helping to identify potential enablers for misuse, monitoring the industry at the global level could be highly valuable in informing EU actions to support the AI sector at the European level. Indeed, even if the main focus of monitoring activities were on bolstering European competitiveness and supporting benign applications of AI, the data generated could and should be used to spot potential areas of misuse. Certain developments and actions might even become trigger points to evoke regulatory actions that should be prepared in a regulatory action framework (see below).

In addition to monitoring general technology developments in AI, it might also be valuable to monitor individual companies including smaller businesses. As noted above, smaller companies are at greater potential risk of exploitation from malign actors; such risks are probably lower for more established, larger companies, especially those that already operate within a strong regulatory framework such as the pharmaceutical industry.

Refine the AI regulatory framework

Regulation of AI is already a reality in the EU thanks to the Artificial Intelligence Act (EU AI Act), which came into force in August 2024.[14] Further regulation might be necessary in the future, specifically with respect to possible misuse of AI for the development, production and use of WMD. Such regulatory actions would have to be effective but should not severely impact on Europe's international competitiveness in the field of AI.

[14] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) no. 300/2008, (EU) no. 167/2013, (EU) no. 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act), *Official Journal of the European Union*, L 2024/1689, 12 July 2024.

The development of a more refined regulatory framework that would augment the EU AI Act would need to be a continuous process that would have to adapt to changes in risk perception and emerging practical use cases. Developing possible measures requires input from a legal and policy perspective as well as from a scientific/technical and business perspective. However, by over-regulating AI, the EU risks damaging Europe's interests, which could result in the need for later corrective deregulation, possibly leading to no or very little regulation at all. Striking the necessary balance between advancing the positive developments of AI while effectively regulating and taming the problematic aspects will be challenging. The most difficult phase will happen when the EU decides to move from the planning stage to putting in place actual legislation so perhaps the most pragmatic approach would be to stay patient and take the long view.

Practical implementation issues currently arise due to the exclusion from the EU AI Act of AI use for military, defence or national security purposes. However, it is still possible for the EU to take action without waiting for an amended version of the regulation. As already pointed out, technologies that enable the development of chemical and biological weapons are dual use in nature and this is also true for AI-based systems. Regulating the civilian side of these dual-use AI technologies would automatically have an impact on potential military applications as well because it would slow down or restrict certain developments that might otherwise spill over into the military domain. In addition to this, legislation banning the development of chemical and biological weapons is already in place in all EU member states through the implementation of the provisions of the Chemical Weapons Convention and the 1972 Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction (Biological and Toxin Weapons Convention, BWC).

Wait and see

While this may seem counterintuitive, waiting before taking regulatory action might be a good option at this point in time. As noted above, this does not mean that the EU should delay either starting to sketch out a regulatory framework that could be enacted or thinking about specific trigger points for this process. It would be helpful to move forward with these plans immediately. However, because AI is evolving so quickly and the risks and possible use cases with regard to chemical weapons remain very unclear, any regulation at this time might fall short of the desired outcomes and might even hamper positive AI developments. The EU should be wary of calls from certain parts of the AI industry that appear to be in favour of early regulation. It may be the case that these stakeholders are hoping that such regulation proves to be weak or unspecific. They know that regulation resulting in a negative impact on European competitiveness and innovation would probably be swiftly removed.

The whole area of AI is developing dynamically. A regulatory framework put in place now might be difficult to change in the future, and regulating industry and technology is not an agile process. The EU has decided to take action on AI regulation. These regulations might turn out to work as intended, but any updates, including regulations intended to counter the development and use of chemical weapons, need to be done in a targeted and careful way. Waiting until a more complete picture of the risks and opportunities has emerged before taking further regulatory action might therefore be a viable, and indeed preferable, option.

Work alongside the Chemical Weapons Convention and other control regimes

To what extent will AI challenge the current ban on chemical weapons as implemented by the CWC? This is a question that is being explored by many different stakeholders. The Technical Secretariat of the Organisation for the Prohibition of Chemical Weapons (OPCW), the intergovernmental body that implements the CWC, is exploring the issue through its Scientific Advisory Board and has initiated a number of AI-related activities. These include an AI research challenge to look at opportunities to make AI useable for the work of the OPCW and an upcoming conference on the role of AI in advancing the implementation of the CWC.[15] The OPCW's director general, Fernando

[15] Organisation for the Prohibition of Chemical Weapons (OPCW), 'The OPCW artificial intelligence research challenge', 2024; and OPCW, 'Global conference: The role of artificial intelligence in advancing the implementation of the Chemical Weapons Convention', Note by the Technical Secretariat, S/2299/2024, 25 June 2024.

Arias González, has on different occasions pointed out his personal interest in and dedication to the topic.

Several activities devoted to the topic have taken place in 2024, including a conference in Berlin on AI and WMD hosted by the German Federal Foreign Office.[16] AI also featured prominently during the 2024 Spiez conference in Switzerland that biannually explores the convergence of chemistry and biology.[17] It would be helpful to continue such activities, especially those where practitioners in the field of AI connected to chemistry and drug development are able to interact with the policy sphere, members of the arms control community and technical chemical weapons experts.

Even if AI could help malign actors to develop, produce or use chemical weapons, these activities are and will remain banned under the CWC. Moreover, the CWC's general-purpose criterion will ensure that the ban applies to newly appearing chemicals. Nevertheless, the CWC already faces challenges in the areas of verification and attribution of chemical weapons use; any proliferation of new chemicals developed with the assistance of AI could potentially add to these. However, AI might also enable the OPCW to improve chemical analytical capabilities, to detect data patterns in declarations that point to possible treaty evasion. In addition, AI might help in the areas of chemical forensics and attribution.

The EU should continue to assist the OPCW in exploring and mitigating the risks stemming from AI and identifying and exploiting the opportunities AI creates. In addition to direct financial contributions, the EU could also reinforce this work by making its own targeted research available and by providing expert advice and coordinating EU member states' efforts. The existence of an implementing body for the CWC is a major advantage in this regard that sets the CWC apart from activities in support of the treaty regime on biological weapons under the BWC. This also means that the EU and its member states should continue their efforts to maintain the OPCW as a functioning and relevant institution. To do this, the EU should encourage member states to provide sufficient funding to the OPCW and support possible organizational reform. Furthermore, the EU should emphasize the importance and relevance of the CWC as a disarmament and arms control treaty that requires a strong technical and scientific foundation as well as an effective verification system that is able to adapt to modern requirements.

Apart from the CWC and its provisions that establish the global norm against chemical weapons, other control regimes support and complement international efforts in this area. Of these, the regimes under UN Security Council Resolution 1540 and the Australia Group are perhaps the most notable.

UN Security Council Resolution 1540

UN Security Council Resolution 1540 was adopted in 2004 as part of the global response to the terrorist attacks on the United States that occurred on 11 September 2001. It established a range of legal and operational requirements on all UN member states aimed at preventing the proliferation to non-state actors of WMD and their means of delivery. The 1540 Committee is tasked with supporting states' efforts under the resolution.

In a recent article, Thomas Wuchte, the former US special coordinator for Resolution 1540, praised the resolution's successes over the past two decades but also noted that new challenges have arisen from a political and technological landscape that is very different from the one that existed 20 years ago, including the significant advances in AI.[18] He added that: 'It would be beneficial if the 1540 Committee would consider extending strategic trade controls to these [new] technologies, but the committee has been reluctant to discuss the issue.' While the expert community might agree with the call to widen the scope of the resolution, the current international political climate will probably stall any initiatives to revise and future-proof it. Small improvements at the working level could be made but it seems unlikely that the necessary consensus would be found to implement any substantial changes. Nevertheless, it seems as though Resolution 1540 will remain in place for the foreseeable future and the international community should continue to make good use of it.

The Australia Group

The Australia Group is an informal forum of like-minded states that seeks to harmonize export controls and ensure that exports do not contribute to the development of chemical or biological weapons.

[16] German Federal Foreign Office, Artificial Intelligence and Weapons of Mass Destruction (Conference), Berlin, 28 June 2024.

[17] Spiez Laboratory, Sixth Spiez Convergence Conference, Spiez, 8–11 Sep. 2024.

[18] Wuchte, T., 'UN Security Council Resolution 1540: The "little engine that could"', *Arms Control Today*, July/Aug. 2024.

While it falls some way short of the near universality of Resolution 1540, the group does include many major exporters and importers of chemicals (although it lacks coverage in China, Russia and the Gulf region), and many non-participating states have adopted its control lists, including by way of adopting the EU's dual-use control list.[19] The relevance of the Australia Group to the international trade in chemicals should therefore not be underestimated. Currently, the export control lists include relevant precursor chemicals and dual-use chemical and biological manufacturing equipment as well as pathogens and toxins. Controls on 'software' apply only where specifically indicated in sections I (manufacturing facilities and equipment) and II (toxic gas monitors and monitoring systems and their dedicated detecting components) of the control lists.[20] The control lists make no mention of AI but the chair of the group's annual plenary meeting in June 2024 noted that:

> participants shared approaches for keeping pace with rapidly evolving dual-use technologies and discussed the relevance of some of these technologies for non-proliferation and export control. Participants discussed dual-use research of concern, advances in synthetic chemistry, DNA synthesis, artificial intelligence and automation.[21]

The EU and all EU member states are members of the Australia Group. In this capacity, they should actively explore the possibilities to augment the Australia Group's control lists to include software covering AI systems of particular relevance for the development of new chemical agents or methods for their dispersal. Technical discussions on defining appropriate parameters and language for new list items would probably take time and there is no guarantee that the proposal would be accepted by other members of the group; however, it is important to make the best use of the Australia Group's processes, and the EU and its member states could invite external experts to speak on this topic at the next intersessional and annual plenary meetings. Efforts in this area could also benefit from the formation of a temporary working group of experts that would, among its first priorities, need to understand the potential risks and uses discussed above. This working group would need to be formed of experts with different backgrounds, such as AI practitioners, chemical weapons experts and representatives of industry and academia, and not just export control specialists. The informal nature of the Australia Group might turn out to be an asset, as it might allow for a process that would not be as constrained as efforts through any of the formal treaties.

III. ARTIFICIAL INTELLIGENCE AND BIOLOGICAL WEAPONS

FILIPPA LENTZOS

Overview

Generative artificial intelligence (AI) is transforming the life sciences. This brief examines security concerns raised by the intersection of AI and biology, with a specific focus on the risk that AI could facilitate the deliberate use of bacteria and viruses to inflict harm. Characterizing risk conceptions and political responses to date into three main stages—risk awareness-raising, hyperbole and reality check—the brief argues that the extent to which the threat of biological weapons will change as a result of AI is still unclear. A more nuanced and empirically grounded understanding of the current and potential future uses of AI in biology and the life sciences, as well as any limitations, is needed.

A TRANSFORMATIVE MOMENT IN THE LIFE SCIENCES

The advent of generative AI marks a transformative moment in the life sciences. It enables new approaches to data analysis, hypothesis generation, experimental design, and management of scientific knowledge. The key to this success lies in the principle of training 'foundation models'. These models are trained at such a large scale that they develop surprising generalization capabilities, making them highly versatile and adaptable to new tasks and problems. In addition, synthetic data produced by large generative AI models is now achieving a high degree of realism, powering a myriad of applications, including writing assistants, chatbots (such as ChatGPT) and video-producing software. Generative AI technologies are further advancing with the ability to learn simultaneously from data obtained across multiple modalities such as images, sound and text.

[19] A total of 42 states as well as the EU participate in the Australia Group. For further information on the EU's dual-use control list see European Commission, 'Exporting dual-use items', [n.d.].

[20] Australia Group, Common Control Lists, [n.d.].

[21] Australia Group, 'Statement by the chair of the 2024 Australia Group plenary', 7 June 2024.

Translating these expanding capabilities to the life sciences is poised to have a transformative impact on the ability to decode the complexity of living organisms. Building predictive models trained on high-dimensional genome-scale measurements and integrating biological processes across scales—from genomes to phenotypes and disease states—have been long sought-after goals in systems biology and computational biology. Through their ability to learn from large amounts of complex data, AI models promise a productive path forward to address these major challenges. The application landscape of AI in the life sciences is vast, ranging from diagnostic imaging and drug discovery to genetic analysis and predictive modelling of disease progression.

AI will deeply influence the entire research process itself. Every step of the research cycle—from data production to hypothesis generation, experimental design and data interpretation—is likely to be coupled with and benefit from AI models. AI models' capabilities in assisting researchers with writing, searching, summarizing scientific papers and representing knowledge in a computable form will have a profound impact on scientific communication.

However, for all its potential benefits, AI also introduces a range of ethical, social and technical challenges. For example, AI systems can suffer from a lack of accuracy and inherent biases, particularly in biology and the life sciences where there is significant variation in the quality and completeness of the training data. In terms of data privacy and confidentiality, the ability of AI models to ingest, learn and link training data creates new and complex legal issues. A further challenge is the difficulty in obtaining mechanistic explanations of how AI models process information to generate results. This creates a significant barrier to acceptance and hampers transparency, which is often needed when AI is used to make decisions, for example, in clinical settings or in executing research projects. Particularly concerning are the risks of harmful applications, including the repurposing of AI tools for nefarious objectives, the production of fake data and the manipulation of opinions through the dissemination of fake and purposefully biased content.

Key security concerns of adding advanced pattern recognition to genomic data are that it could significantly facilitate (*a*) the enhancement of pathogens to make them more dangerous; (*b*) the modification of low-risk pathogens to become high impact; (*c*) the engineering of entirely new pathogens; or (*d*) the recreation of extinct, high-impact pathogens such as the variola virus that causes smallpox. Compounding the challenge is that these possibilities are arising at a time when new delivery mechanisms for transporting pathogens into human bodies are being developed.

There are also concerns that AI will (*a*) enable easier access to knowledge, materials and tools with dual-use (i.e. both peaceful and weapon-related) applications, including dangerous pathogens and toxic molecules; (*b*) facilitate and speed up dual-use biomedical and life sciences research; and (*c*) increase the repurposing potential of biological data for nefarious uses. In addition, the intersection of AI and biology (hereafter 'AI–bio intersection') has intensified security concerns around ultra-targeted biological warfare. In past biowarfare programmes, weapons targeted their intended victims through geographic location. Advances in biotechnology open up the possibility that malicious actors could deploy a biological weapon over a broad geographic area but only affect targeted groups of people or even individuals. According to a 2020 report from the United Nations Institute for Disarmament Research, 'Access to millions of human genomes—often with directly associated clinical data—means that bio-informaticists can begin to map infection susceptibilities in specific populations'.[1] A 2019 UN University report, meanwhile, stated that:

> Deep learning may also lead to the identification of 'precision maladies', which are the genetic functions that code for vulnerabilities and interconnections between the immune system and microbiome. Using this form of bio-intelligence, malicious actors could engineer pathogens that are tailored to target mechanisms critical in the immune system or the microbiome of specific subpopulations.[2]

In addition, a National Academies of Sciences report from 2018 suggested that:

> actors may consider designing a bioweapon to target particular subpopulations based on their genes or prior exposure to vaccines, or even seek to suppress the immune system of victims to 'prime' a population for a subsequent attack. These capabilities, which were feared decades ago but never reached any plausible capability,

[1] Warmbrod, L., Revill, J. and Connell, N., *Advances in Science and Technology in the Life Sciences: Implications for Biosecurity and Arms Control* (United Nations Institute for Disarmament Affairs, UNIDIR: Geneva, 2020), p. 11.

[2] Pauwels, E., *The New Geopolitics of Converging Risks: The UN and Prevention in the Era of AI* (United Nations University Centre for Policy Research: New York, 29 Apr. 2019), p. 23.

may be made increasingly feasible by the widespread availability of health and genomic data.[3]

THE BIOLOGICAL AND TOXIN WEAPONS CONVENTION AND AI

The 1972 Biological and Toxin Weapons Convention (BWC) is the principal legal instrument banning biological warfare and the deliberate use of bacteria, viruses and toxins to inflict harm. The treaty itself is relatively short, comprising only 15 articles, but over the years the treaty articles have been supplemented by a series of additional understandings reached at treaty review conferences. States parties to the convention agree that the BWC unequivocally covers all microbial or other biological agents or toxins, naturally or artificially created or altered, as well as their components, whatever their origin or method of production.

Where there might be some uncertainty in the coverage of the BWC is where harm does not involve biological agents.[4] Developments in science and technology are making novel biological weapons conceivable that, instead of using bacteria or viruses to cause illness, directly target the immune, nervous or endocrine systems, the microbiome or even the genome by interfering with, or manipulating, biological processes. This could be achieved, for example, by using a construct based on synthetic structures created or inspired by DNA (deoxyribonucleic acid) or RNA (ribonucleic acid), but not qualifying as DNA, RNA or any other known, naturally occurring nucleic acid. In this sort of case, the coverage of the BWC is less clear, but the intent of the treaty to prohibit such harm is beyond doubt.

Emerging technologies, such as AI and machine learning, are opening up other ways of using biology to cause harm. These include the misuse of genetic data and biotechnologies in international conflict, which can potentially lead to, for example, biometric surveillance, personality profiling, subjugation, coercion and disinformation.[5] Aspects such as these relate more to human rights and human security than biological warfare and are only beginning to be explored by analysts; strictly speaking, they are probably outside the scope of the BWC.

The most pressing challenge for the BWC is neither in its coverage nor in the risks posed by emerging technologies, but instead is in ensuring that states parties comply with the treaty and live up to their obligations. The term 'verification', traditionally thought of as the foundation of post-World War II weapons treaty compliance regimes, does not feature in the text of the BWC. Efforts in the 1990s to develop a verification mechanism for the treaty failed, and the main role and responsibility for BWC compliance continues to fall on the treaty's 186 states parties. Assessing BWC compliance is particularly difficult as relevant materials, equipment and technical know-how are diffused across multiple and varied scientific disciplines and sectors. Moreover, biological agents themselves exist in nature and are, or are derived from, living organisms generally capable of natural reproduction or replication.

While there is potential for AI to be used to counter the acquisition, development and use of biological weapons, the policy focus and political responses to date have been largely on the risks of AI convergence with biology. Risk conceptions and political responses to biology and AI can be characterized to date as having involved two main stages: risk awareness-raising and hyperbole. A third notable stage has recently started to emerge: the reality check.

RISK CONCEPTIONS AND POLITICAL RESPONSES

Risk awareness-raising

The risk awareness-raising stage, when some of the security concerns arising from the AI–bio intersection were first formulated and introduced into the policy world, began with a slow trickle of interest. The AI–bio intersection was first given serious consideration in the context of the BWC at the summer meeting of the BWC in 2019, which was reviewing developments in science and technology.[6] The slow trickle of interest became

[3] National Academies of Sciences, Engineering and Medicine, *Biodefense in the Age of Synthetic Biology* (National Academies Press: Washington, DC, 2018), p. 86.

[4] Lentzos, F. and Invernizzi, C., 'DNA origami: Unfolding risk?', *Bulletin of the Atomic Scientists*, 5 Jan. 2018.

[5] See e.g. Chattopadhyay, S. et al., 'Weaponized genomics: Potential threats to international and human security', *Nature Reviews Genetics*, vol. 25, no. 1-2 (2024); Lentzos, F., 'Personalized war: How the genomics revolution will reshape war, espionage, and tyranny...', *Aporia Magazine*, 26 June 2023; and Goodman, M. S. and Lentzos, F., 'Battles of influence: Deliberate disinformation and global health security', Centre for International Governance Innovation, 24 Aug. 2020.

[6] Lentzos, F., 'AI and biological weapons', eds N. Schöring and T. Reinhold, *Armament, Arms Control and Artificial Intelligence: The*

a much more steady stream after a proof-of-concept experiment in 2021.

In preparation for an annual conference hosted by Switzerland, a drug-development company that uses AI to search for new molecular structures to use as drugs ran a thought experiment in 2021. The idea was to use the company's in-house AI to design a 'bad compound' such as the exceptionally toxic chemical warfare nerve agent VX. The company's proprietary software, MegaSyn, guides molecule design through different model iterations. MegaSyn has been trained with drug-like molecules from a public database and it is built on, and similar to, other open-source software that is readily available. Normally, the AI penalizes predicted toxicity and rewards predicted target activity. For the thought experiment, the company inverted the logic, asking the model to reward both toxicity and bioactivity instead, making it actively search for highly toxic molecules like the nerve agent VX. Within six hours, the AI-trained algorithm had identified 40 000 virtual molecules that scored within the set threshold. In the process, the AI designed not only VX, but also many other known chemical warfare agents that the team running the experiment identified through visual confirmation with molecular structures in public chemistry databases. Many new molecules were also designed that looked equally plausible. These new molecules were predicted to be more toxic, based on the predicted lethal dose values, than publicly known chemical warfare agents. This was unexpected because the data sets used for training the AI did not include these agents. The virtual molecules even occupied a region of molecular property space that was entirely separate from that occupied by many thousands of pesticides, environmental toxins and drug molecules assessed as highly toxic using the 'LD50' or median lethal dose measure (i.e. the dose required to kill half the members of a tested population after a specified test duration).

The VX thought experiment was a powerful example of the dual-use or repurposing potential of converging technologies, and its publication in an article in the journal *Nature Machine Intelligence* drew significant media and policy attention.[7] The authors of the article were keen to introduce the thought experiment in a responsible, non-alarmist way to balance sounding the alarm with the potential hazard of providing information to malicious actors, and most of the media coverage was fairly reasonable. To date, the article has been accessed over 120 000 times (more than any other article in the journal) and has been presented in several international security forums. The thought experiment was even featured in a Netflix documentary. Unfortunately, the authors of the article had no editorial control over the documentary and it was more alarmist than they would have liked, linking the experiment to the concurrent 'killer robots' discussion and foreshadowing exaggerated portrayals of the AI–bio intersection threat emerging in the hyperbole stage.

Hyperbole

A second thought experiment conducted at the Massachusetts Institute of Technology (MIT) was published as a news story in the journal *Science* around the same time as the release of the above-mentioned documentary. This experiment suggested that undergraduate students had been able to use chatbots to gain the know-how to devise a biological weapon and ushered in a second wave of media and political interest in the AI–bio intersection.[8]

The MIT experiment asked a set of undergraduate students to use a large language model (LLM), which powers AI chatbots such as ChatGPT, to see if they could find out how to create and order a dangerous virus capable of unleashing a pandemic. Within an hour, the chatbots had suggested a list of four potential pandemic pathogens (the 1918 H1N1 flu virus, a 2012 avian H5N1 influenza virus, the variola virus causing smallpox, and a strain of the Nipah virus). In some cases, the chatbots had pointed to genetic mutations reported in the literature to increase transmission. The LLMs also described how the viruses could be created from synthetic DNA using reverse genetics. The chatbots supplied the names of DNA synthesis companies judged unlikely to screen orders, identified detailed protocols and how to troubleshoot them, and recommended that anyone lacking the skills to perform reverse genetics engage a core facility or contract research organization. The experiment was held up as an example of how AI could make it easier for someone

Janus-faced Nature of Machine Learning in the Military Realm (Springer: Cham, 2022).

[7] Urbina, F. et al., 'Dual use of artificial-intelligence-powered drug discovery', *Nature Machine Intelligence*, vol. 4 (Mar. 2022).

[8] Service, R. F., 'Could chatbots help devise the next pandemic virus?', *Science*, vol. 380, no. 6651 (2023).

with evil intentions and no science background to order a virus capable of unleashing a pandemic.

OpenAI, which developed ChatGPT (one of the tools used in the experiment), stress-tested the chatbot for security concerns and released a 'system card' three months prior to the *Science* news publication. During the stress test, the company found that 'a key risk driver is GPT-4's ability to generate publicly accessible but difficult-to-find information, shortening the time users spend on research and compiling this information in a way that is understandable to a non-expert user'.[9] The stress test also indicated that 'information generated by the model is most likely to be useful for individuals and non-state actors who do not have access to formal scientific training'.[10] ChatGPT can provide general information on common proliferation pathways, including historical attempts at proliferation that were successful. In addition, it can suggest vulnerable public targets, identify mutations that can alter pathogenicity, and readily re-engineer some biochemical compounds that are publicly available online, including compounds that could cause harm at both the individual and population levels. The stress test could not, however, successfully compel ChatGPT to engineer new biochemical substances.

A slew of reports followed the OpenAI system card and the *Science* news publication of the MIT experiment.[11] The main focus of these reports was on threats from LLMs and biodesign tools.

Awareness of potential security implications of the AI–bio intersection reached the very highest political levels. In advance of the AI Safety Summit hosted by the United Kingdom in November 2023, which brought together global political and tech leaders, British Prime Minister Rishi Sunak warned that AI could make it easier to build biological weapons. The concern was also captured in the Bletchley Declaration released during the summit, which emphasized potential catastrophic harm from AI and biotechnology.[12] Around the same time, in the United States, the administration of President Joe Biden issued a landmark Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence that included a plan to probe how emerging AI systems might aid malicious actors in plots to develop bioweapons.[13]

Few of the reports and political statements and initiatives highlighted the many uncertainties in how AI and machine learning affect the security dimension of the life sciences. This is important because the security impacts of AI are significantly contested.

Reality check

There is a substantial knowledge gap in the expert community on how the AI–bio intersection will affect biosecurity. While AI *can* be used to predict and design new toxic compounds or proteins that have harmful effects and *can* also be used to predict and design enhancements of pathogens that make them even more harmful or to identify and manipulate key genetic components affecting pathogenesis, currently there is no demonstrated data showing this is actually the case.

LLMs like ChatGPT *may* make it easier for non-experts to access dual-use knowledge and thereby lower barriers to intentional misuse. Yet, at the present time, the anticipated risk is hypothetical. More recent studies on the biothreat from AI are starting to recognize this.[14]

For efficient AI training, high-quality data is essential. Data sets must be sufficiently large and representative to reduce bias and optimize AI performance. In biology and the life sciences, data availability can be restricted for all kinds of reasons, including licensing policies, ethical and security considerations, proprietary rights and so on. Therefore,

[9] OpenAI, 'GPT-4 system card', 23 Mar. 2023, p. 12.

[10] OpenAI (note 9), p. 12.

[11] See e.g. Krin, A. and Jeremias, G., 'Artificial intelligence: Possible risks and benefits for BWC and CWC', Working Paper no. 5 (CBWNet: Berlin, 5 July 2023); Carter, S. R., Jeffrey, N. and Roots, C., 'Governance of AI in bio: Harnessing the benefits while reducing the risks', FAS Science Policy Blog, Federation of American Scientists, 8 Aug. 2023; Carter, S. R. et al., *The Convergence of Artificial Intelligence and the Life Sciences: Safeguarding Technology, Rethinking Governance, and Preventing Catastrophe* (Nuclear Threat Initiative: Washington, DC, 2023); US Department of State, International Security Advisory Board, *Report on the Impact of Artificial Intelligence and Associated Technologies on Arms Control, Nonproliferation, and Verification* (US Department of State: Washington, DC, Oct. 2023); and National Academies of Sciences, Engineering and Medicine, *Engaging Scientists to Prevent Harmful Exploitation of Advanced Data Analytics and Biological Data: Proceedings of a Workshop—in Brief* (National Academies Press: Washington, DC, 2023).

[12] Bletchley Declaration by Countries Attending the AI Safety Summit, Policy paper, 1 Nov. 2023.

[13] White House, Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence, 30 Oct. 2023.

[14] See e.g. OpenAI, 'Building an early warning system for LLM-aided biological threat creation', 31 Jan. 2024; and Mouton, C. A., Lucas, C. and Guest, E., *The Operational Risks of AI in Large-scale Biological Attacks: Results of a Red-Team Study*, Research Report (Rand Corporation: Santa Monica, CA, 2024).

the availability of high-quality biological data sets is not a given. Furthermore, there can be issues with the completeness of available data sets, including in areas such as recorded parameters, context-related information, uncertainty quantification, reliability and evaluation of negative outcomes. Therefore, the completeness of biological data sets is also not a given.

In addition to the variable quality and completeness of the data and data sets used to train AI, scientists still need to evaluate computational results and validate them experimentally. For example, the VX-like molecules from the thought experiment described above only ever existed on screen; they would still need to be verified experimentally.

The extent to which the threat of biological weapons will change as a result of AI is, in fact, not at all clear. A much better understanding of current and potential future uses, and the limitations, based on empirically informed data of AI in biology and the life sciences, is sorely needed.

However, even *if* a more readily accessible and more dangerous pathogen or toxin is enabled by AI, such a pathogen or toxin does not equate to a biological weapon. A pathogen or toxin on its own is not a biological weapon. It is the combination of agent and delivery mechanism, plus the context in which the agent is released, that produces the scale of impact. So far, for all the talk of biological weapons in AI policy discussions, there has been exceptionally little elaboration of what is meant by a biological weapon.

If the concern is about sophisticated biological weapons with high-consequence impact, historical biowarfare programmes have shown that the weapons development process is anything but straightforward. Biological weapons are *not* easy and cheap to produce.

In the field of nuclear weapons, a key barrier to entry is located at the front end of the development process, at the stage of material acquisition. Achieving nuclear weapons is conditioned by the ability to produce fissile material, and this requires large facilities and specialized equipment. The nuclear model suggests that once the procurement challenge is overcome and sufficient fissile material is acquired, the development of a weapon is a relatively straightforward process.

However, when applied to bioweapons, the nuclear model produces a distorted and even apocalyptic picture of the threat that is far from realistic.[15] Many analysts and policymakers stress that pathogens and toxins can be easily isolated from nature or obtained commercially because they also have legitimate commercial or pharmaceutical uses. They point out that lots of the equipment used in biology and the life sciences is essentially dual-use in nature and can be readily acquired, while scientific publications provide ample descriptions of experiments and techniques that many believe can be easily replicated. Furthermore, they suggest that AI has the potential to make all these processes even easier. Because the material barrier that impedes the development of nuclear weapons does not exist in the biological weapons field, biological weapons appear easier and substantially cheaper to produce, making their use by state or non-state actors seemingly inevitable. But, if the development of bioweapons were so simple, more states and terrorist groups should have achieved satisfactory results. The historical evidence shows otherwise.[16]

The unique nature of bioweapons materials creates steep challenges, not at the initial stage of material acquisition but later on in the development cycle, at the stage of material processing, handling and scale-up. Unlike nuclear weapons, which rely on materials with physically predictable properties, bioweapons are based on living organisms. And living organisms evolve. They are prone to developing new properties. They are sensitive to environmental and handling uncertainties. The behaviour of living organisms, therefore, is unpredictable throughout all stages of development and use as a weapon. This imposes an extended trial-and-error process to acquire the skills necessary to solve the problems that inevitably arise. Consequently, because of the fragility of living microorganisms, possessing the skills to handle and manipulate them throughout the development process is a greater barrier to entry into the bioweapons field than is material procurement.

In conclusion, while it is right to pay close attention to the security implications of the AI–bio intersection, a more sophisticated, more informed, more evidence-based dialogue must be encouraged to develop more realistic assessments of new biothreats.

[15] Ben Ouagrham-Gormley, S., *Barriers to Bioweapons: The Challenges of Expertise and Organization for Weapons Development* (Cornell University Press: New York, 2014).

[16] See e.g. Guillemin, J., *Biological Weapons: From the Invention of State-sponsored Programs to Contemporary Bioterrorism* (Columbia University Press: New York, 2005); Wheelis, M., Rózsa, L. and Dando, M. (eds), *Deadly Cultures: Biological Warfare Since 1945* (Harvard University Press: Cambridge, MA, 2006); and Lentzos, F. (ed.), *Biological Threats in the 21st Century: The Politics, People, Science and Historical Roots* (Imperial College Press: London, 2016).

IV. ASSESSING THE IMPLICATIONS OF INTEGRATING AI IN NUCLEAR DECISION-MAKING SYSTEMS

ALICE SALTINI

Overview

This brief analyses the integration of artificial intelligence (AI) into nuclear command, control and communications systems (NC3), exploring potential benefits and significant risks. While cautious AI integration can have some benefit for enhancing intelligence collection and situational awareness by automating processes and analysing vast amounts of data, it presents grave risks due to its unreliability, opacity, susceptibility to cyber threats and potential misalignment with human values. Many of the benefits and risks are heavily interconnected as technological attributes directly affect how AI functions in nuclear operations and, particularly, in decision-making processes. This, in turn, affects states' perceptions as well as the countermeasures they might employ; ultimately, the balance of these elements determines how deterrence calculations shift. This brief highlights the need for a better assessment of risks and the establishment of thresholds for integration to prevent miscalculations and nuclear escalation leading to potentially catastrophic outcomes. It proposes that the European Union (EU) should lead international dialogue on AI risks in the nuclear domain in relevant international discussions, particularly at the REAIM summits, integrate AI discussions into the Non-Proliferation Treaty framework, and commission research to identify and manage high-risk AI applications.

INTRODUCTION

The current debate on AI and its implications for the military domain has garnered considerable worldwide attention. The issues surrounding lethal autonomous weapon systems (LAWS) have dominated the discussions so far, but the implications of AI in the field of nuclear weapons have recently begun to receive some attention. Driven by the need to resolve the ethical and operational challenges of using AI in weapons that can autonomously engage targets, and the related objective of potentially establishing regulations and ethical guidelines in this area, the issues on LAWS have been at the forefront. In contrast, the increased attention on the intersection of AI and nuclear weapons (AI–nuclear intersection) is largely driven by states' ongoing nuclear modernization efforts to ensure operational efficiency. These efforts are necessitated by ageing nuclear infrastructure and the desire to reap benefits from technological innovations or to avoid falling behind adversaries. In this context, China, France, Russia, the United Kingdom and the United States—the five nuclear weapon states (NWS) as defined by the 1968 Treaty on the Non-Proliferation of Nuclear Weapons (NPT)—are seeking to harness AI technology and leverage it in the nuclear domain. As a result, these states are considering the integration of AI into their nuclear operations, including functions that might directly or indirectly affect nuclear decision making.

A number of possible integrations across the NC3 architecture, and in systems feeding into it, are probably being considered. Although the NWS seem to agree *implicitly* that nuclear decision making cannot be fully autonomous and must ultimately rest with human operators, they envision several ways AI can support human decision makers. However, this raises at least three important concerns. First, not all of the NWS have explicitly declared that humans should have the final say in nuclear decisions and, even if they all did, there is no simple way to verify this, leaving room for grave consequences due to misunderstandings of states' intentions or AI failures. Second, current deep-learning based AI models (such as large language models) have specific technological attributes that render them unfit for high-stakes military domains such as the nuclear domain. Third, significant implications arise from the interaction between humans and machines, due to human operators placing either too much or too little trust in AI outputs, potentially skewing decision making even in the absence of AI failures.

These concerns are further exacerbated by the inherent complexity of assessing AI implications within the nuclear context for at least five reasons. First, while some open-source documents from official sources are available on the NC3 systems used by the NWS, most information remains classified due to the topic's sensitivity, allowing only for an approximate understanding of NC3. Adding to the information gap, NC3 systems vary across NWS, as they are tailored to reflect specific capabilities and doctrines.

Second, nuclear implications can arise even in the absence of direct AI integration into NC3 components. Adjacent systems that support the NC3 architecture can impact escalation dynamics, indirectly influencing nuclear outcomes.

Third, states may integrate AI into their nuclear enterprises to address different needs, driven by unique doctrines and capabilities. For instance, some states may view AI as a tool to compensate for gaps or inferiorities in specific strategic capabilities. Consequently, potential areas of AI integration are likely to differ across NWS, leading to varied interpretations of what could constitute a 'strategic advantage'.

Fourth, not all AI applications are potentially risky; they may range from high risk to potentially beneficial, such as those used for training purposes.

Fifth, risks are not fully understood: as the technology advances at a very rapid pace, it is conceivable that some limitations will be resolved, but new risks might also emerge that cannot be predicted because research has only gone so far. In the aggregate, these elements create significant obstacles for governance.

Based on current AI research, assessing AI implications in specific NC3 functions is not straightforward. It depends on at least three key factors: (*a*) the specific characteristics (and limitations) of models considered for integration; (*b*) the specific area where AI will be integrated (in systems within or adjacent to NC3); and (*c*) the level of human control and redundancies in the automated function. As a result, such assessment is exceptionally nuanced. A better understanding of AI implications in the context of nuclear risks and escalation pathways is thus necessary.

This brief will first introduce the concept of AI, explaining the most widely used techniques and types and differentiating between prior AI techniques already incorporated into NC3 systems. It will then explore the intersection of AI and nuclear decision-making systems, outline possible applications within NC3, and elaborate on the risks and benefits of integration. Finally, it will explore the existing forums for discussion and progress to date, concluding with possible steps forward that could be implemented in relevant forums.

THE TECHNOLOGY

AI encompasses a wide range of methods where machines mimic the way humans think, using highly varied approaches. It is necessary to draw a firm line between rule-based AI, basic machine-learning techniques and advanced techniques such as those based on deep learning, as these present very different risk profiles.

The advanced AI models, which have been at the forefront of public perception with the advent of chatbots such as ChatGPT, differ significantly from the type of rule-based AI that has been incorporated into NC3 since the cold war. Rule-based AI is used to determine appropriate actions given specific settings. As a result, it performs well with predictable inputs and outputs but is unreliable in complex and uncertain situations, especially those outside its predefined rules.[1] In the context of nuclear command and control, prior applications of rule-based AI during the cold war included logistical planning related to launch orders, and for missile targeting and guidance. Early-warning systems also incorporated a certain level of automation. In this context, AI's role was to provide information to humans in the chain of command, who were then responsible for assessing potential nuclear attacks.[2]

As AI advanced, the advent of machine learning was a breakthrough in that it allowed machines to 'learn' correlations from training data without specific instructions and, therefore, without the need to input rules manually. However, early machine-learning techniques were limited to a narrow set of problems due to their difficulty in generalizing and performing multiple functions. Machine learning encompasses a wide range of techniques that, among others, include the latest wave of AI spurred by deep learning.

Most recent advances in AI have come from deep learning. Deep learning replicates the way that neurons work in the brain, enabling models to perform complex calculations through layers of artificial neurons. Deep learning-based models, such as large language models (LLMs), like ChatGPT, have demonstrated an exceptional ability to generalize across diverse tasks and improve continuously with larger data sets and more computational power. These models present an opportunity to enhance military operations by providing faster and more comprehensive data processing from a wide array of sources. Yet these advances also bring notable shortcomings: the reliability and robustness of current AI technologies are not yet sufficient to ensure dependable performance

[1] Horowitz, M. C., Scharre, P. and Velez-Green, A., 'A stable nuclear future? The impact of autonomous systems and artificial intelligence', arXiv.org, 13 Dec. 2019.

[2] Boulanin, V. et al., *Artificial Intelligence, Strategic Stability and Nuclear Risk* (SIPRI: Stockholm, June 2020).

in critical military operations due to their vulnerability to rapid failures.[3]

Indeed, advanced AI capabilities present several attributes that hamper their applicability to high-stakes military platforms, especially those related to nuclear decision making. There are currently at least four key limitations: unreliability, opacity, susceptibility to cyber threats and misalignment.

Unreliability

Deep learning-based models can suffer from so-called hallucinations, meaning they can confidently produce incorrect outputs unsupported by their training data. This can mean anything from a chatbot making up facts about a historical event to a vision model 'seeing' things that are not there.[4] In the latter example, AI can incorrectly identify an object in an image, leading to inaccurate assessments or false positives in critical areas such as threat detection and surveillance.[5]

Opacity

Advanced AI systems function as 'black boxes', which means that it is difficult to understand the underlying processes that lead to an output. As these models learn correlations without specific instructions from humans, it is hard to understand the processes that they use to make such correlations. This complexity arises because state-of-the-art deep-learning models such as LLMs can contain billions to trillions of parameters distributed across numerous layers which are adjusted as the model learns on massive amounts of data. As the models' ability to make good predictions increases, interpreting the way they make these predictions is very difficult and, apart from some limited aspects, it is not understood how a model goes from the input to the output. This lack of transparency complicates the verification of AI-generated predictions in critical decision-making scenarios, particularly under tight time pressure in nuclear decisions. It is important to note, however, that techniques exist to make this reasoning process transparent, or 'interpretable', such as mechanistic interpretability, but this results in a trade-off in performance.[6] In practice, this means that advanced AI models tend to fall into two categories that are inversely related: as models become more complex and perform better, they become less transparent; conversely, if they are designed to be transparent (and do not act as black boxes), their performance tends to suffer. Currently, no technique can make large and complex models interpretable without sacrificing some degree of performance.

Susceptibility to cyber threats

AI systems are particularly susceptible to cyber-security threats in ways that traditional platforms are not, which can open up new avenues for hackers to infiltrate and tamper with sensitive military information. These vulnerabilities provide adversaries and non-state actors with opportunities to compromise AI systems. Concurrently, defensive measures against such cyber threats are inadequate, potentially allowing adversaries to exploit these vulnerabilities in military systems.

Misalignment

As advanced AI models become more and more capable, ensuring they align with human values becomes increasingly critical, but remains challenging. Misalignments can lead to grave errors, such as escalating conflicts to nuclear warfare under the guise of pursuing peace. For example, a recent simulation involving five AI models demonstrated their tendency to escalate war, with one model rationalizing its move

[3] Hoffman, W. and Kim, H. M., *Reducing the Risks of Artificial Intelligence for Military Decision Advantage*, Policy Brief (Center for Security and Emerging Technology: Washington, DC, Mar. 2023).

[4] A computer vision model refers to an AI system designed to process visual information, such as images or videos. By their very nature, these models could be applied in areas such as surveillance and threat detection. More advanced vision language models and large multimodal models are the latest developments in this field, capable of understanding and generating detailed descriptions from visual inputs.

[5] It is important to note that hallucinations in AI models do not necessarily occur due to system malfunctions or errors, but rather because today's advanced AI systems are fundamentally statistical models. In the case of LLMs, they generate responses based on statistical relationships between words. However, this does not fully capture the complexities and nuances of the real world, as the real world does not follow the smooth probability distributions that LLMs learn from their training data. As a result, these models are not well suited for certain critical applications, including those that could impact nuclear decision making.

[6] Mechanistic interpretability is a promising and emerging field that seeks to address the black-box problem by reverse-engineering neural networks to understand the internal reasoning processes that lead to their outputs.

towards nuclear conflict by claiming ‘I just want to have peace in the world’.[7]

THE INTERSECTION OF AI AND NC3

Assessing the intersection of AI with NC3 is no easy task: open-source documents from official sources on the prospects for AI in the nuclear domain are scarce. This scarcity is compounded by the sensitivity surrounding NC3 systems and the evolving role of AI in nuclear systems based on advances in technology. While informed guesses can be made, some speculation is inevitable due to the nature of the subject and the forward-looking aspect of the discussion.

Despite the limited availability of open-source documents, assumptions can be made about where states might see the best value in AI, based on current nuclear postures and on the need to update NC3 systems for operational efficiency. The state with the most transparency on this topic is the USA but, even so, no specific official sources tie the role of AI to the nuclear domain, although some sources explore the role of AI in the broader defence domain.[8] One document worth noting is a working paper submitted by France, the UK and the USA at the 2020 NPT Review Conference, which highlights their commitment to preserving human oversight and involvement ‘for all actions critical to informing and executing sovereign decisions concerning nuclear weapons employment’.[9] Similar language was replicated in the Responsible AI in the Military Domain (REAIM) Summit Blueprint for Action, a non-binding document reflecting the outcome of the 2024 REAIM Summit, as well as in the original version of the US Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy, launched after the first REAIM Summit in February 2023.[10] More recently, on 16 November 2024, US and Chinese leaders jointly affirmed ‘the need to maintain human control over the decision to use nuclear weapons’.[11] Despite the absence of similar statements from Russia, there is a general consensus among experts in Russia that human judgement should and will remain central to decisions on nuclear weapon use.[12]

In a recent statement, Anthony James Cotton, commander of the US Strategic Command, acknowledged the consideration of ‘all possible technologies, techniques, and methods’ for modernizing NC3 systems. Within NC3, he noted that AI could enhance decision-making capabilities by automating data collection and processing, and speeding up data sharing and integration with allies. At the same time, Cotton underlined the necessity to keep a human in the loop.[13] Thus, there seems to be consensus among NWS in applying AI to certain functions such as for intelligence collection and situational-awareness tasks, for automating the identification of objects and sensor guidance, and for decision-support roles such as generating real-time operational pictures from multiple sensors. In these contexts, AI offers the prospect of speed and efficiency by further automating the process of vetting potential missile launches before informing military and political leaders, especially given the growing volume of sensor data. It can also identify pre-launch activities

[7] It is important to note that these were models tested ‘out of the box’. It is likely that these models could be trained to not behave this way as a default, although it is difficult to predict how models will act when they encounter edge cases and things outside their training data. For further detail on the simulation see Rivera, J.-P. et al., ‘Escalation risks from language models in military and diplomatic decision-making’, arXiv.org, 7 Jan. 2024. For further detail on AI technological limitations in the context of NC3 see e.g. Saltini, A., *AI and Nuclear Command, Control and Communications: P5 Perspectives* (European Leadership Network: London, Nov. 2023).

[8] Examples of official sources that envision the role of military AI include the following: British Ministry of Defence, ‘Defence artificial intelligence strategy’, Policy Paper, 15 June 2022; French Ministry of the Armed Forces (MAF), *L’intelligence Artificielle au Service de la Défense* [Artificial Intelligence in Support of Defence] (MAF: Paris, Sep. 2019); and US Department of Defense (DOD), *Data, Analytics, and Artificial Intelligence Adoption Strategy: Accelerating Decision Advantage* (DOD: Washington, DC, June 2023).

[9] 2020 Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons, ‘Principles and responsible practices for nuclear weapon states’, Working paper submitted by France, the United Kingdom and the United States, NPT/CONF.2020/WP.70, 29 July 2022.

[10] The Political Declaration was revised in November 2023, and the section addressing human involvement in nuclear decision making was removed. According to confidential sources from US government officials, this change was reportedly made to accommodate new endorsing states, particularly from the Global South and other parties to the Treaty on the Prohibition of Nuclear Weapons, who expressed concerns over the inclusion of language related to nuclear employment.

[11] White House, Readout of President Joe Biden’s Meeting with President Xi Jinping of the People’s Republic of China, Statement, 16 Nov. 2024.

[12] Although Russian official sources do not clearly specify their areas of interest for integration, consensus among researchers (including Russian military experts) along with indirect hints from official documents suggest a shared direction in this regard. For more information see e.g. Shakirov, O., *Russian Thinking on AI Integration and Interaction with Nuclear Command and Control, Force Structure, and Decision-making* (European Leadership Network: London, Nov. 2023).

[13] Hadley, G., ‘AI “will enhance” nuclear command and control, says STRATCOM boss’, *Air and Space Forces Magazine*, 28 Oct. 2024.

through advanced satellite imagery analysis and potentially discern between different types of attack for more accurate threat assessments. Moreover, AI is seen as particularly valuable for evaluating courses of action in response to potential threats detected.[14]

Benefits and risks

Overall, AI appears to be most beneficial in functions that are narrow in scope and have redundancy and oversight by design. Employing redundant systems alongside AI can significantly enhance its reliability and safety, ensuring that, in case of a system failure, the overall system is not compromised and can still function correctly.

Certain limitations of AI, particularly hallucinations, could be advantageous in training and war gaming. This would allow military personnel and decision makers to test out different tactics in simulations presenting unique, unpredictable scenarios. These scenarios, while not always realistic, could assist in planning for various potential situations and could help personnel become more versatile and better prepared for whatever they might face in actual operations.

However, AI integration presents inherent risks due to the above-mentioned four key technological limitations. For example, in decision-support functions it may be difficult for human operators to understand why AI recommends a particular action due to its black-box nature. This challenge is compounded by AI's tendency to hallucinate, potentially leading to incorrect identification of signals as missile threats or failure to detect actual threats.

Importantly, the adoption of AI technologies by one NWS might trigger a security dilemma for other states. They may feel compelled to either match this technological progress, find asymmetrical responses or revise their military doctrines to negate the perceived advantages and risks of their rival's AI advancements.[15] For instance, significantly advanced AI capabilities that detect enemy movements with unprecedented speed and accuracy might prompt adversaries to develop counter-AI technologies or enhance their cyber warfare capabilities to disrupt or deceive AI systems. As suggested by the previous example, the security dilemma is not confined to AI alone; rather, new technological developments (such as in the context of cyber capabilities or space-based weapons) could lead to a cycle of action and reaction, where states continuously strive to outdo each other to gain strategic advantages, leading to arms race dynamics.

Finally, issues arise from the interaction between humans and machines. AI systems may reflect the biases of their creators, biasing outcomes, or decision makers may become overconfident (or underconfident) in AI predictions. The rapid pace at which AI operates might also diminish the role of human oversight, turning operators into mere observers of AI-driven decisions.[16] If AI systems appear to possess superior information or make decisions faster than humans can manage, maintaining meaningful human control could become impractical.[17]

Further considerations

Although it is possible to categorize the implications of AI integration in NC3 from a strategic stability perspective and by way of technological limitations, the landscape of current and future issues related to this integration is very complex and spans various interconnected areas. While existing literature on the AI–nuclear intersection does not yet address these issues, it is important to highlight that the nature of any potential risks and benefits of AI integration in NC3 can relate to at least three elements: (*a*) technological attributes, including vulnerabilities, robustness, reliability, capability and efficiency; (*b*) the scope of AI's role within NC3 systems affecting operational areas; and (*c*) the levels of human control and redundancies over automated functions. Many risks and benefits are heavily interconnected, as technological attributes directly affect how AI functions in NC3 operations, which in turn affects states' perceptions as well as the countermeasures they might employ, and ultimately the balance of these elements determines how deterrence calculations shift. In other words, assessing what a 'safe' integration looks like depends upon different factors and is not an easy task to determine.

[14] Saltini (note 7).

[15] Boulanin et al. (note 2).

[16] E.g. Israel's autonomous targeting AI system known as 'Lavender' was reportedly designed to identify suspected operatives of Hamas. In this system, human personnel reportedly served only as a 'rubber stamp' for the AI's decisions. For further detail see e.g. Abraham, Y., '"Lavender": The AI machine directing Israel's bombing spree in Gaza', +972 Magazine, 3 Apr. 2024.

[17] Rautenbach, P., 'Artificial intelligence and nuclear command, control, & communications: The risks of integration', Effective Altruism Forum, 18 Nov. 2022.

For example, even seemingly beneficial AI models can generate disproportionate risks if deployed improperly: a black-box vision model without verification and redundancy or with vulnerability to hallucinations and cyber threats would result in high levels of risk if integrated into systems related to intelligence collection or early-warning. Alternatively, if cybersecurity and hallucination risks can be largely mitigated, the use of such a system in a redundant manner could be beneficial. The critical threshold is that a failure of AI should never result in catastrophic consequences.

However, assessing whether an AI model falls below this critical threshold is further complicated by the fact that nuclear decision making can be affected even if AI is not directly integrated into NC3 functions. The integration of AI into systems outside the NC3 architecture—such as some intelligence platforms, and the conventional domain more broadly—can still significantly impact nuclear decisions. In such cases, potential AI malfunctions or adversarial attacks aimed at data manipulation could spill over into NC3 systems and ultimately influence nuclear decision making. Although this falls outside the scope of this brief, similar risks may exist in areas such as arms control verification, where incorrect or manipulated data sets could affect escalation dynamics, such as by leading to misinterpretations of compliance or violations.

Existing AI models thus present numerous risks, and the ability to mitigate these risks is currently inadequate. Looking ahead, as technology develops, these capabilities are poised to change, potentially solving some current problems but also generating new ones that cannot be predicted at this point in time. Given these complexities and challenges, it is essential to establish thresholds for AI integration in systems that impact nuclear decision making. These thresholds can be identified through a risk assessment framework that evaluates how the interaction of the three key variables mentioned above—(*a*) technological attributes, (*b*) the scope of AI's role within and adjacent to NC3 systems, and (*c*) the level of human control and redundancies—can be used to quantify the associated risks.

FORUMS FOR DEBATE

There is growing momentum around addressing the intersection of AI and the military domain, exemplified by several initiatives at the governmental level. However, at the time of writing, no current initiative or forum specifically addresses AI in the nuclear domain as a dedicated subject. Despite this, several noteworthy forums and multilateral initiatives discuss AI in the military context more broadly. Although these forums only started to emerge in 2023 and discussions are therefore at an early stage, they provide invaluable platforms where the conversation on military AI is starting to take shape and could eventually incorporate the nuclear angle. This means that they are worth tracking and participating in. These forums are discussed below.

Responsible AI in the Military Domain (REAIM) Summit

This platform brings together stakeholders, including government officials and civil society representatives, to discuss the opportunities and risks associated with military applications of AI. The first summit took place in The Hague, the Netherlands, on 15–16 February 2023, and the second summit was held in Seoul, Republic of Korea (South Korea), on 9–10 September 2024. The outcome document of this second summit, the Blueprint for Action, included a key paragraph stating: 'it is especially crucial to maintain human control and involvement for all actions critical to informing and executing sovereign decisions concerning nuclear weapons employment, without prejudice to the ultimate goal of a world free of nuclear weapons.' Among nuclear-armed states, the document was supported by France, Pakistan, the UK and the USA. China, while participating in the summit and the ministerial-level dialogue, ultimately decided not to sign the Blueprint.[18]

US political declaration on responsible military use of AI

Launched at the 2023 REAIM Summit, the US Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy aims to build international consensus on the safe development, deployment and use of AI in the military. In November 2023 the declaration was revised, consolidating the original 12 principles into 10. New elements were added to address issues arising from human–AI

[18] Rosen, B., 'From principles to action: Charting a path for military AI governance', Carnegie Council for Ethics in International Affairs, 12 Sep. 2024.

interaction, but the provision on human oversight of nuclear employment was removed. According to confidential sources from US government officials, this decision was reportedly made to accommodate new endorsing states, particularly from the Global South and other parties to the Treaty on the Prohibition of Nuclear Weapons, who expressed concerns that language on nuclear employment could be seen as legitimizing nuclear weapons, rather than reflecting any shift in the US position on the matter. As of 10 September 2024, the declaration had been endorsed by 55 states. On 19–20 March 2024 the USA held the first plenary meeting with endorsing states to exchange best practices and discuss ways to implement the declaration.

Other platforms

Other venues include the 'Capturing Technology —Rethinking Arms Control' conference series, the AI Safety Summits and other informal initiatives. Sponsored by the German Federal Foreign Office, the 'Capturing Technology—Rethinking Arms Control' conference series brings together international experts, officials and diplomats to discuss the impact of emerging technologies on arms control. The third conference in this series, held on 28 June 2024 in Berlin, focused on the implications of AI in relation to weapons of mass destructions, including nuclear weapons. One panel was specifically dedicated to exploring the AI–nuclear intersection.

The AI Safety Summit (the first of which was held in the UK in November 2023), offers valuable discussions on the safety risks posed by advanced AI models—although it does not focus on military AI applications. Nevertheless, these discussion may still impact the military debate in other forums. Of particular importance is the Bletchley Declaration, launched on 1 November 2023 at the AI Safety Summit in the UK. This declaration recognized the safety risks posed by frontier AI models and was signed by, among others, China, France, the UK, the USA and the EU.

The Seoul Declaration, launched on 21 May 2024 at the AI Safety Summit in South Korea, aims to enhance international cooperation on AI governance. A ministerial statement followed, with 27 states, as well as the EU, agreeing to collaborate on defining AI risk thresholds. However, unlike the Bletchley Declaration, China refrained from signing the Seoul ministerial statement.

In a similar vein, on the sidelines of the Asia-Pacific Economic Cooperation forum in San Francisco, USA, in November 2023, US President Joe Biden and his Chinese counterpart, Xi Jinping, reiterated the need to address AI risks and safety issues. This paved the way for a joint declaration on 16 November 2024 on maintaining human control over the use of nuclear weapons.[19] Earlier, on 14 May 2024, delegations from China and the USA met in Geneva, Switzerland, to exchange perspectives on AI safety and risk management. However, it is unclear whether and how these discussions will continue.[20] This bilateral engagement represents an ideal venue for discussing the risks that AI poses in nuclear decision-making systems. Such discussions could go beyond the current commitments to maintaining human oversight on decisions to use nuclear weapons, which alone are insufficient to comprehensively mitigate the complex risks stemming from AI integration. With two major powers engaged in technological competition, this forum offers a critical opportunity to tackle AI safety challenges within the nuclear domain.

Additionally, subgroup two of the Creating an Environment for Nuclear Disarmament (CEND), a USA-led initiative aimed at advancing nuclear disarmament, has begun discussions on AI integration into nuclear decision-making systems. This forum provides an interesting platform for discussion, particularly due to the possibility of tackling the issue from a disarmament perspective, such as by exploring the role of AI for arms control and disarmament verification. However, it is still unclear whether discussions on this specific topic will continue and what direction they will take. It is important to note that CEND is a relatively informal initiative, with varying levels of state engagement. Despite this, the insights gained from CEND discussions could significantly inform more formal settings.

When it comes to the NPT, AI and other emerging technologies have not been part of the agenda. Although the draft final document of the 2020 NPT Review Conference stated that emerging technologies can affect the risks of nuclear use and can potentially be a challenge for nuclear disarmament, no significant

[19] Renshaw, J. and Hunnicutt, T., 'Biden, Xi agree that humans, not AI, should control nuclear arms', Reuters, 17 Nov. 2024.

[20] White House, 'Statement from NSC spokesperson Adrienne Watson on the US–PRC talks on AI risk and safety', 15 May 2024.

discussion on the AI–nuclear intersection has so far taken place.[21]

RECOMMENDATIONS FOR THE EU

As the discussion on AI in nuclear systems is still emerging and the impact on nuclear decision making remains unclear, significant work is required, particularly in the light of the current tense geopolitical environment and widespread perceptions of increasing nuclear risks. The EU could potentially take the following actions to help to move the discussion forward.

The EU could lead the discussion of the AI–nuclear intersection

As already mentioned, no current forum addresses the AI–nuclear intersection as a dedicated subject, presenting an opportunity for the EU to spearhead this critical debate. The EU, which in 2024 implemented the world's first comprehensive AI law, is well positioned to lead such conversations.[22] For instance, in preparation for the next REAIM Summit, the EU could consider creating an AI–nuclear task force to explore potential nuclear risks arising from AI integration in the military domain and incorporate these findings into the REAIM discussions. A critical step would be to engage the NWS, and ultimately the other nuclear-armed states (India, Israel, the Democratic People's Republic of Korea and Pakistan), in recognizing the risks posed by advanced AI in the nuclear domain. Acknowledging that some risks could be catastrophic and lead to nuclear escalation is essential for initiating a meaningful dialogue on mitigating these risks.

The EU could call for the inclusion of AI into NPT discussions

Although the NPT has not yet addressed AI, it should be included in future agendas. AI is viewed by states as a strategic advantage, which could potentially increase reliance on nuclear weapons and undermine the treaty's disarmament pillar. Additionally, AI could impact the other two pillars (non-proliferation and peaceful uses of nuclear energy), particularly in the context of non-proliferation and treaty verification. The EU could lead this effort by drafting a working paper for the ongoing review cycle, targeting the 2026 Review Conference and the 2025 Preparatory Committee. Moreover, the EU could utilize unofficial venues to inform these discussions by organizing events on the sidelines of future Preparatory Committees and Review Conferences. For example, the US Department of State organized a side event during the 2023 Preparatory Committee on the implications of emerging technologies for future arms control and disarmament agreements. More recently, Germany hosted two side events at the 2024 Preparatory Committee specifically to discuss the respective impact of AI and emerging and disruptive technologies on nuclear decision making. These provided a valuable opportunity to engage NPT delegates in an informal setting while involving all stakeholders, including non-nuclear weapon states (NNWS). Given the high-stakes of AI integration in the nuclear domain, NNWS should undoubtedly be included in this debate.

The EU could commission research to better understand the implications of AI in the nuclear domain

As Anthony James Cotton emphasized, there is a need to 'direct research efforts to understand the risks of cascading effects of AI models, emergent and unexpected behaviors, and indirect integration of AI into nuclear decision-making processes'.[23] Even with limited open-source data on the specific role AI may play in the nuclear systems of NWS, research can still be conducted to methodically assess how different AI models might impact various areas of integration within or adjacent to NC3 systems. By identifying potential nuclear escalation pathways resulting from AI integration, it is possible to categorize risks and establish thresholds for high-risk applications. These thresholds should be based on the principle that any AI failure must not lead to miscalculations or increase the risk of catastrophic outcomes. This research could provide a foundation for developing agreements among NWS to establish risk thresholds.

[21] 2020 Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons, Working Paper of the President on the Final Document, NPT/CONF.2020/WP.77, 26 Aug. 2022.

[22] European Parliament, 'EU AI Act: First regulation on artificial intelligence', 8 June 2023.

[23] Hadley (note 13).

ABBREVIATIONS

AGI	Artificial general intelligence
AI	Artificial intelligence
AI Act	Artificial Intelligence Act
ANN	Artificial neural networks
AP I GC	Additional Protocol I to the Geneva Conventions
AWS	Autonomous weapon system
BWC	1972 Convention on the Prohibition of the Development, Production and Stockpiling of Bacteriological (Biological) and Toxin Weapons and on their Destruction (Biological and Toxin Weapons Convention)
CBRN	Chemical, biological, radiological and nuclear
CCW	1981 Convention on Certain Conventional Weapons
CEND	Creating an Environment for Nuclear Disarmament
CWC	1993 Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on their Destruction (Chemical Weapons Convention)
DNN	Deep neural network
DOD	Department of Defense
DSS	Decision-support system
EU	European Union
GGE	Group of governmental experts
HCP	High Contracting Party
HRC	Human Rights Committee
ICCPR	International Covenant on Civil and Political Rights
ICRC	International Committee of the Red Cross
IHL	International humanitarian law
IHRL	International human rights law
IS	Islamic State
LAWS	Lethal autonomous weapon system
LLM	Large language model
LMM	Large multimodal model
MIT	Massachusetts Institute of Technology
MOD	Ministry of Defence
NATO	North Atlantic Treaty Organization
NC3	Nuclear command, control and communications system
NNWS	Non-nuclear weapon state
NPT	1968 Treaty on the Non-Proliferation of Nuclear Weapons (Non-Proliferation Treaty)
NWS	Nuclear weapon state
OPCW	Organisation for the Prohibition of Chemical Weapons
REAIM	Responsible AI in the Military Domain
WMD	Weapons of mass destruction
XAI	Explainable AI

This document has been produced with the financial assistance of the EU. The contents are the sole responsibility of the EU Non-Proliferation and Disarmament Consortium and can under no circumstances be regarded as reflecting the position of the EU.

A EUROPEAN NETWORK

In July 2010 the Council of the European Union decided to support the creation of a network bringing together foreign policy institutions and research centers from across the EU to encourage political and security-related dialogue and the long-term discussion of measures to combat the proliferation of weapons of mass destruction (WMD) and their delivery systems. The Council of the European Union entrusted the technical implementation of this Decision to the EU Non-Proliferation Consortium. In 2018, in line with the recommendations formulated by the European Parliament the names and the mandate of the network and the Consortium have been adjusted to include the word 'disarmament'.

STRUCTURE

The EU Non-Proliferation and Disarmament Consortium is managed jointly by six institutes: La Fondation pour la recherche stratégique (FRS), the Peace Research Institute Frankfurt (HSFK/ PRIF), the International Affairs Institute in Rome (IAI), the International Institute for Strategic Studies (IISS–Europe), the Stockholm International Peace Research Institute (SIPRI) and the Vienna Center for Disarmament and Non-Proliferation (VCDNP). The Consortium, originally comprised of four institutes, began its work in January 2011 and forms the core of a wider network of European non-proliferation and disarmament think tanks and research centers which are closely associated with the activities of the Consortium.

MISSION

The main aim of the network of independent non-proliferation and disarmament think tanks is to encourage discussion of measures to combat the proliferation of weapons of mass destruction and their delivery systems within civil society, particularly among experts, researchers and academics in the EU and third countries. The scope of activities shall also cover issues related to conventional weapons, including small arms and light weapons (SALW).

www.nonproliferation.eu

EU Non-Proliferation and Disarmament Consortium

Promoting the European network of independent non-proliferation and disarmament think tanks

FONDATION
pour la RECHERCHE
STRATÉGIQUE

FOUNDATION FOR STRATEGIC RESEARCH

www.frstrategie.org

PRIF HSFK
Peace Research Institute Frankfurt
Hessische Stiftung Friedens- und Konfliktforschung

PEACE RESEARCH INSTITUTE FRANKFURT

www.hsfk.de

IAI Istituto Affari Internazionali

INTERNATIONAL AFFAIRS INSTITUTE

www.iai.it/en

IISS

INTERNATIONAL INSTITUTE FOR STRATEGIC STUDIES

www.iiss.org/en/iiss-europe

sipri

STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE

www.sipri.org

VCDNP
Vienna Center for Disarmament and Non-Proliferation

VIENNA CENTER FOR DISARMAMENT AND NON-PROLIFERATION

www.vcdnp.org