UNIVERSITÄT OSNABRÜCK

Geostrategy and Geopolitics
Department 1
49069 Osnabrueck
apl. Prof. Dr. Dr. Klaus Saalbach

Artificial Intelligence and Nuclear Weapons

20 May 2024

Summary
Artificial Intelligence (AI) has a great potential in military applications and is seen by nuclear powers as indispensable and essential. AI and big data could be used for control of non-proliferation of nuclear weapons, could save time and costs in the research, design optimization, manufacturing, testing and certification, maintenance and surveillance of nuclear warhead systems and manage resources more efficiently. Military AI could enhance early warning and Intelligence, Surveillance and Reconnaissance (ISR) capabilities, communication reliability and accelerate decision-making. The integration of AI into the nuclear command, control, and communications (NC3) systems could also synchronize information across nuclear and non-nuclear command and control. For autonomous nuclear-weapon systems, AI can support obstacle detection and maneuverability, automated target identification, and long-range and loitering capability.
The debate on the use of AI for nuclear weapons covers three areas, the autonomy, the stability of military AI systems and the strategic stability. A potential autonomy of nuclear weapons is part of the broader debate on lethal autonomous weapons systems (LAWS). Automated and semi-autonomous nuclear decision-makings were already considered during cold war (SAGE; Perimeter). A specific military AI problem is the mission stability, as the AI systems lack context knowledge and may decide too quickly. The opacity of the systems leads to the explainability or interpretability issue; further problems may result from data poisoning, manipulated images, automation bias and artificial escalation. Wargame simulations with current generative AI Large language models (LLMs) from OpenAI, Anthropic and Meta showed that the systems tend to escalation up to nuclear strikes. As software system, AI is vulnerable for cyber attacks, generative AI also for prompt injections. A main concern of all involved parties is the strategic stability, i.e., to avoid everything that gives a reason for a nuclear first strike. Such a reason could be uncertainty about the capabilities of the adversary, because if it is not known what the other side can do in a certain situation, the only chance in a nuclear conflict is to strike first. AI can undermine strategic stability also by compression of decision-making time, which may result in escalation or inadvertent use of nuclear weapons, amplify misunderstandings and misperceptions during a crisis, and by encouraging a premature deployment of insufficiently tested AI. Moreover, AI systems facilitate the use of ‘dead hand’-systems and autonomous nuclear weapons. The rise of hypersonic weapons and the increasing speed of warfare both undermine strategic stability as well. Currently, the AI is still perceived as too immature to be used in high-risk strategic situations; there is a substantial risk of technical failures and of biased, incomplete, or inaccurate data. The nuclear powers agree that for command and control at the strategic level, the role of AI should remain supportive. United States, China and Russia have started dialogues on AI risks. This paper briefly presents the current state of the debate and the background.

Content

1 Introduction

1.1 Overview

Artificial Intelligence (AI) is commonly understood as the ability of machines to perform tasks that normally require human intelligence and is a key area of advanced computing. On 02 May 2024, the *US State Department* asked China and Russia to declare that they will not give control of nuclear weapons to AI systems, as the US already declared together with the United Kingdom and France[1].

The debate on the use of AI for nuclear weapons covers three areas, the autonomy, the stability of military AI systems and the strategic stability, i.e. to avoid everything that creates incentives for a nuclear first strike[2].

This paper briefly presents the current state of the debate and the background. After an introduction on the legal framework and key documents which are relevant for the debate, the discussions about autonomy, stability of AI systems and the strategic stability will be presented and summarized.

1.2 Brief History

Even for human intelligence, there is no standard definition, but the core of human intelligence definitions includes the mental capacity to recognize, analyze and solve problems, and a human being is then more intelligent if this can be done faster and/or for more complex problems. Many definitions of Artificial Intelligence focus on activities that require human intelligence, but already the simple pocket calculators of the 1970ies did something that requires human intelligence.

The time pressure and the amount of information during a potential nuclear attack already led in the early 1950ies to an automatization of data analysis and decision support and is meanwhile seen as first step to the use of AI in nuclear defense[3].

The United States used the fastest *Whirlwind II IBM* mainframe computers in command centers to receive, sort, and process data from radars and sensors to identify Soviet bombers and the SAGE supercomputer then coordinated US and Canadian aircraft and missiles for defense[4]. But with the development of nuclear *intercontinental ballistic missiles (ICBMs)* it became clear that SAGE computer could be destroyed and after a congressional hearing in 1966, the SAGE command centers were shut down[5].

This motivated the development of the first precursor of the internet, where decentralized computers in command centers were linked with each other.

Some autonomy was already embedded in nuclear missiles in the 1970ies when *multiple independent targeted reentry vehicles (MIRVS)* were invented. While the military can control the missile, the final distribution of the nuclear bombs is done automatically[6].

The Soviet Union developed the semi-autonomous *Perimeter* system for nuclear warfare. This is a **dead hand-system** which guarantees a second-strike capability even if the normal command and control chain is destroyed by a nuclear first strike[7]. If the *Perimeter* communication network gets silent, i.e., if the contact to the primary command center is lost and some other factors appear at the same time, like light radioactivity, seismic activity, and

[1] Norman 2024
[2] Shakirov 2023
[3] Lowther/McGriffin 2019
[4] Sankaran 2019
[5] Sankaran 2019
[6] Puwal 2024
[7] Lowther/McGriffin 2019

atmospheric pressure changes, the authority to launch nuclear weapons was transferred to the local commanders[8]. The system is still operational[9]. Originally, the Soviets considered a fully automated system that would transfer decision-making powers to machines during situations like a nuclear attack, but this plan was discarded[10].

In 2018, the Russian *Ministry of Defense (MoD)* held its first conference on Artificial Intelligence and in 2021, the *46th Central Research Institute (46th TsNII)* was designated for the development and integration of military-oriented AI technologies. In 2022, a new ministerial department was put in charge of AI development[11].

Human beings are emotional and make mistakes, but human experience and understanding of context can compensate technical errors as demonstrated by the *Petrov incident*: In 1983, an early warning system in Soviet Union indicated a launch of nuclear missiles from US, but lieutenant colonel Stanislav Petrov correctly interpreted this as malfunction due to the low numbers of missiles[12]. The system misinterpreted a natural phenomenon[13].

In 1962 during the Cuba crisis, the Soviet officer Vasili Arkhipov resisted in a submarine B-59 the pressure of co-commanders to release a nuclear torpedo when the submarine was surrounded by 11 US destroyers[14].

A key issue is the unexpected rapid progress of AI technologies. The so-called 'weak' AI can reproduce an observed behavior and can carry out tasks after while 'strong' AI would be an intelligent system with real consciousness and the ability to think, i.e., to think and say "I" and to ask "why". Strong AI is meanwhile discussed under the terms *Artificial General Intelligence AGI*[15] (reaching human level of cognition) and *Artificial Super-Intelligence ASI* which goes beyond human intelligence[16]. This led to an intense debate on AI for military applications including nuclear weapons.

A growing AI application is the *Generative AI* where the AI can create content like new images, texts, sounds, and videos based on short instructions, the prompts. While such programs were already developed since years, the public breakthrough came with the release of the *Large Language Model (LLM) ChatGPT (Generative Pretrained Transformer)* version 3.5 in November 2022 from the company *Open Artificial Intelligence (OpenAI)* that could generate complex responses with correct logic and grammar or expand initial entries[17]. There are other competitors such as *Copilot, Bard, Gemini, Llama, Claude* etc. which are permanently updated as well; and a lot of further applications e.g., for songs, images and books were released. Only 18 months later in May 2024, *ChatGPT4o* (o = omni = Latin for everything) was released that can utilize voice, text and images and has a lot of new functionalities, such as detection of user emotions, simulation of own emotions, real-time verbal chat including real-time translation into 50 languages, reading of hand-writing, solving math problems etc.[18]

The problem is that military systems are typically complex, large-scale technologies and before an AI application would be embedded everywhere and combat-ready, it would be already

[8] Saltini 2023a/Depp/Scharre 2024
[9] Lowther/McGriffin 2019, Saltini 2023a, Depp/Scharre 2024
[10] Saltini 2023a. This matter was the topic of the Stanley Kubrick's movie *Dr. Strangelove or: How I Learned to Stop Worrying and Love the Bomb* from 1964.
[11] Shakriov 2023
[12] Depp/Scharre 2024
[13] Kaur 2024
[14] Kaur 2024
[15] Kölling 2023
[16] Zia 2023
[17] Iqbal 2023
[18] OpenAI 2024

outdated, i.e., due to the rapid and unpredictable expansion of AI the military AI sector is very dynamic and unstable.

On 30 October 2023, US President Joe Biden signed an *Executive Order on safe, secure, and trustworthy Artificial Intelligence* which in Section 4.4 mandates the *Department of Homeland Security DHS* to investigate how AI may enhance *CBRN (chemical, biological, radiation and nuclear)* threats and to establish the *AI Safety and Security Board*[19].

On 02 May 2024, an official from the *US State Department* asked China and Russia to declare that they will not give control of nuclear weapons to AI systems, as the US already committed together with the United Kingdom and France in 2022[20].

United States, China and Russia have started dialogues on AI risks. In 2023, the *Nuclear Threat Initiative NTI* released a report on '*Reducing Cyber Risks to Nuclear Weapons*' produced by a Track II process between Russian and US experts[21]. In January 2024, the US National Security Advisor Jack Sullivan and Chinas foreign minister Wang Yi met in Bangkok and made plans to establish a dialogue on AI risks by the new generation of AI platforms, including the need to look on military applications[22].

2 Legal Framework and Key Documents

2.1 Legal Framework

The key international document is the *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy* agreed in February 2023 at the *Responsible AI in the Military Domain Summit (REAIM 2023)* in The Hague[23]. Initiated by the United States, this is a non-binding guidance which aims to build international consensus around responsible behavior and guide states' development, deployment, and use of military AI and is intended as discussion platform between states for further steps. In late November 2023, approximately 50 states signed this document. The aim is not a ban as it includes the right develop and use AI in the military domain, but with the aim to embed this into strong and transparent norms.

The *Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy* provides definitions which are in line with the discussions in the literature[24]:

Artificial Intelligence (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. […] **Autonomy** may be understood as a spectrum and to involve systems operating without further human intervention after activation. […] and explains further that "**Military AI capabilities** include not only weapons but also decision support systems that help defense leaders at all levels make better and more timely decisions, from the battlefield to the boardroom...."

For military practice, the *DoD Directive 3000.09* "*Autonomy in Weapon Systems*" from November 2012 was revised in 2023 to establish a policy and assigns responsibilities for developing and using autonomous and semiautonomous functions in weapon systems, to minimize the probability and consequences of failures in autonomous and semi-autonomous

[19] WhiteHouse 2023, Heslop/Keep 2024
[20] Joint Statement 2022, Norman 2024
[21] Shakirov 2023
[22] Heslop/Keep 2024
[23] USA 2023
[24] USA 2023

weapon systems that could lead to unintended engagements and, as new unit in 2023, to establish the *Autonomous Weapon Systems Working Group*[25].

A widely agreed classification of human involvement[26] is

- "Human in the loop": weapon systems that use autonomy to engage individual targets or specific groups of targets that a human can and must decide to engage.
- "Human on the loop": weapon systems that use autonomy to select and engage targets, but human controllers can halt their operation, if necessary
- "Human out of the loop": weapon systems that use autonomy to select and engage specific targets without any possible intervention by human operators.

On 30 October 2023, US President Joe Biden signed an *Executive Order on safe, secure, and trustworthy Artificial Intelligence*[27]. The order requires that developers of the most powerful AI systems share their safety test results and other critical information with the U.S. government and to develop standards, tools, and tests to help ensure that AI systems are safe, secure, and trustworthy[28].

In 2024, the European Union plans to release a regulatory and legal AI framework, the *Artificial Intelligence Act (AI Act),* to classify and regulate AI applications based on their potential risks.

2.2 Further Key Documents

In 2022, United States, the United Kingdom and France released the *Principles and responsible practices for Nuclear Weapon States*. They reaffirmed the *Joint Leaders' Statement on Preventing Nuclear War and Avoiding Arms Races* from 03 January 2022, in particular the understanding that nuclear war cannot be won and must never be fought and in *Point vii*, that "*Consistent with long-standing policy, we will maintain human control and involvement for all actions critical to informing and executing sovereign decisions concerning nuclear weapons employment*"[29].

The *US Department of Defense's (DoD) 2022 Nuclear Posture Review (NPR)* which is part of the *2022 National Defense Strategy*[30] also states that current policy is to "maintain a human 'in the loop' for all actions critical to informing and executing decisions by the President to initiate and terminate nuclear weapon employment" in all cases[31].

A bipartisan initiative of US Senators in April 2023 proposed the *Block Nuclear Launch by Autonomous Artificial Intelligence Act of 2023*[32]. The act aims to prevent AI from launching a nuclear weapon and clearly prohibits to use an "*autonomous weapons system that is not subject to meaningful human control or to launch a nuclear weapon; or to select or engage targets for the purposes of launching a nuclear weapon"* [33]. Moreover, the act would ensure that no federal funds can be used for any launch of any nuclear weapon by an automated system without meaningful human control[34].

The main competitor of United States in the AI sector is China. According to the 2017 *New Generation AI Development Plan*, China is aiming to become the global AI leader and develop

[25] DoD 2023a
[26] CoE 2022, Sayler 2023, DoD 2023a, where US intends to restrict autonomous weapons to humans in or on the loop.
[27] WhiteHouse 2023, Heslop/Keep 2024
[28] WhiteHouse 2023
[29] Joint Statement 2022
[30] DoD 2022
[31] Markey et al. 2023
[32] Markey et al. 2023, Congress 2023
[33] Markey et al. 2023, Congress 2023
[34] Markey et al. 2023, Congress 2023

a domestic AI market worth USD 150 billion by 2030[35]. The Chinese government views AI as an opportunity to "leapfrog" the United States by focusing on AI for enhanced battlefield decision-making, cyber capabilities, cruise missiles, and autonomous vehicles in all military domains[36].

To accelerate the transfer of AI technology from commercial companies and research institutions to the military as *Civil-Military Integration (CMI)*, the Chinese government created a *Military-Civil Fusion Development Commission* in 2017[37]. Aim is the development of warfare from mechanization to '*informationisation*' and now with A.I. to '*intelligentisation*'. Thus, for the Chinese army PLA, AI is essential for "*intelligentised warfare*"[38], this means to equip military technology with AI.

China has released two key position papers in relation to AI and its associated risk, the *2021 Position Paper on Regulating Military Applications of Artificial Intelligence* and the *2022 Position Paper on Strengthening Ethical Governance of Artificial Intelligence*[39]. China stated that nuclear-weapons states should refrain from using AI-enabled systems to strike each other, including nuclear capabilities[40].

In June 2022, the United Kingdom's *Ministry of Defence (MOD)* released the '*Defence Artificial Intelligence Strategy*' and a policy paper on the *'Ambitious, safe and responsible'* use of AI[41].

3 The Debate on AI and Nuclear Weapons

3.1 The Military Potential of AI

Artificial Intelligence has potential in military applications and is seen as indispensable and essential[42]. The leading nuclear powers consider the integration of generative AI models in military systems before their adversaries do the same, i.e., there is a kind of AI arms race[43].

AI and big data could be used for control of non-proliferation of nuclear weapons, could save time and costs in the research, design optimization, manufacturing, testing and certification, maintenance and surveillance of nuclear warheads and manage resources more efficiently[44].

AI is not only relevant for the use of nuclear weapons, but also for the development and testing. Since France ratified the *Comprehensive Nuclear-Test-Ban Treaty (CTBT)* in 1998, the nuclear program runs entirely on mathematical modelling by a program called '*Simulation*' of the *Atomic Energy Commission* that probably uses algorithms and AI-enabled tools, to certify the efficiency of the French nuclear weapons[45].

In 2023, the US DoD released the *DoD Data, Analytics, and AI Adoption Strategy* to combine and replace the *2018 AI Strategy* and the *2020 Data Strategy* to make rapid, well-informed decisions by expertly leveraging high-quality data, advanced analytics, and AI[46]. The primary goal is **decision advantage** based on battlespace awareness and understanding, adaptive force

[35] Hoadley/Sayler 2019, p.1, NATO 2019, p.10
[36] NATO 2019, p.10
[37] Hoadley/Sayler 2019, p.20-22
[38] Bommakanti 2020, p.3-4
[39] Su/Yuan 2023
[40] Saltini 2023a
[41] Saltini 2023b
[42] Fayet 2023, Shakirov 2023, House of Lords 2023
[43] Saltini 2023a
[44] ISAB 2023, NNSA 2023, House of Lords 2023
[45] Fayet 2023
[46] DoD 2023b

planning and application, fast, precise, and resilient kill chains, and a resilient sustainment support of operations and efficient operations.

AI may enhance early warning and *Intelligence, Surveillance and Reconnaissance (ISR)* capabilities[47], i.e., the coordinated sensing, acquisition, fusion, processing, and dissemination of accurate, relevant, and timely information and intelligence to support military decision-making processes[48], command and control with better protection against cyber-attacks and an improved force and stockpile management, with precision strikes and attacks with improved or autonomous navigation and support accompanying non-nuclear operations as well[49]. Also, AI could contribute to communication reliability[50] and accelerate decision-making[51].

A key strategic issue is that the current *nuclear command, control, and communications (NC3)* systems of the United States needs to be updated as it is more than 30 years old and made up of more than 200 individual systems[52]. The integration of AI into the NC3 systems could also synchronize information across nuclear and non-nuclear command and control[53]. For autonomous nuclear-weapon systems, AI can support obstacle detection and maneuverability, automated target identification, and long-range and loitering capability[54].

China's approach for future nuclear command and control systems is the combined 'new trinity' of '*AI-Cyber-Nuclear*'[55]. Autonomous weapon systems can make nuclear command and control systems more resilient against cyber interference and attacks, but autonomous weapon systems combined with cyber offensive weapons could also be used to attack enemies' nuclear warheads and related systems[56].

Various authors also argue that the concerns about AI are exaggerated. They see AI as a useful tool while an *Artificial General Intelligence AGI* does not exist[57]. While the proposed *Block Nuclear Launch by Autonomous Artificial Intelligence Act of 2023* plans to prohibit to select or engage targets for the purposes of launching a nuclear weapon, military AI could be useful for a *look-shoot-look*-Strategy, i.e., find, hit, and assess damage[58]. In 2022, a commanding US general advocated in the US Congress for *human-on-the-loop* instead of *human-in-the-loop* for detection of approaching nuclear missiles to utilize the higher speed of machines[59].

The AI risks may be controlled by employing AI with discrete capabilities that do not evolve into a sentient system[60].

3.2 Autonomy and Dead Hand Systems

Some authors argue that lethal autonomous weapon systems are not so much different from the current situation, as for example in fighter jets decision-making is already highly dependent on automated software interfaces that characterize, sort, interpret, and prioritize the output of a huge range of sensors more precisely and more efficiently than any human could do[61].

[47] Jalil 2023
[48] Chaudry/Klein 2023
[49] Jalil 2023
[50] Hruby/Miller 2021
[51] Kaur 2024
[52] Hruby/Miller 2021
[53] Hruby/Miller 2021
[54] Hruby/Miller 2021
[55] Su/Yuan 2023
[56] Su/Yuan 2023
[57] Puwal 2024
[58] Puwal 2024
[59] Hruby/Miller 2021
[60] Lowther/McGriffin 2024
[61] Ford 2020

Nevertheless, some authors argue that autonomous weapons are still not reliable enough and that there is a creeping automation of military systems[62]. Currently, the US use the *dual phenomenology approach*, i.e., missiles are only fired if confirmation of an enemy attack came from two independent sensor systems, e.g. ground-based radar and satellite data[63].

Among the approximately 800 AI-related projects[64] and unmanned device (UxS) programs of the *US Department of Defense (DoD)*, the US *Defense Advanced Research Projects Agency (DARPA)* is supporting the development of the *ShELL (Shared Experience Lifelong-Learning)* project. Normally, an AI is developed by training of an algorithm on a data set. Once the training is complete, the AI application is released. The application could be regularly updated and upgraded, but a certain version is somewhat 'static' while for military purposes it is essential to have actual data.

The ShELL concept is based on EDGE computing, i.e., intermediate computers between the internet and the central computer that can sort, analyze, and update data while the central computer can then combine and share the data and utilize them by AI[65]. This would be a permanently learning 'dynamic AI'. However, a first internal paper of the *OpenAI Superalignment Team* which should accompany and safeguard the development of future AIs showed how a smaller AI model may safeguard a larger one (Chat-GPT 2 versus Chat-GPT 4), but the paper could not show how a dynamically growing AI could be safeguarded[66].

The need for a *dead hand system* in the United States is disputed. Russia developed the semi-autonomous *Perimeter* system for nuclear warfare which guarantees a second-strike capability even if the normal command and control chain is destroyed by a nuclear first strike[67], but a malfunction could lead into a nuclear disaster[68]. On the other hand, the United States are confronted with an increasing attack-time compression from modernized nuclear systems and hypersonic missiles and needs a modernized *nuclear command, control, and communications (NC3)* system[69]. Threats for the US are e.g., the Russian *Kaliber-M* and Kh-102 cruise missiles, the *Poseidon Ocean Multipurpose System Status-6* unmanned underwater vehicle also known as *Kanyon*, and the *Avangard Objekt 4202* hypersonic weapon[70].

A borderline case is the *US Air Force* B-21 bomber which can fly without a crew and is nuclear capable[71]. Currently, the rule is that a crew should be always present if nuclear weapons are on board[72].

3.3 The Stability of Military AI

A specific military AI problem is the **mission stability**[73]. Autonomous military systems can improve reconnaissance and intelligence and can speed up decision making and may also allow rapid reaction, but also may destabilize military missions. Examples:

- An autonomous system may decide to attack a relevant target, but by this disclose a military presence and jeopardize Special Forces or Intelligence Operations.

[62] Depp/Scharre 2024
[63] Depp/Scharre 2024
[64] Raasch 2023 For example, autonomous supersonic aircraft capabilities are being developed for the US DoD by EpiSci.
[65] NTA 2021
[66] Burns et al. 2023
[67] Lowther/McGriffin 2019
[68] Depp/Scharre 2024
[69] Lowther/McGriffin 2019 and 2024
[70] Lowther/McGriffin 2019 and 2024, Hruby/Miller 2021
[71] Air Force 2024
[72] Depp/Scharre 2024
[73] Masuhr 2019, Johnson 2020

- In the *DARPA Cyber Challenge* of 2016, the best computer was a machine that defended itself on the expense of the defense systems[74].
- A computer may decide that a combat at a certain location may be a waste of resources and withdraw e.g., a drone swarm, but may not understand that sometimes a certain location has a symbolic and psychological value, or is maybe foreseen as anchor point of a new front line or that the fight is only done to distract adversaries from more important areas. The question is: will an advanced military AI really be able to think strategically or only tactical? Context is still not fully understood by the systems[75].
- Mission authority problem: In civil airplanes, pilots already had to fight against defect autopilots which could not be overridden in critical situations[76].
- An AI may decide to fight too quickly, leaving the conventional forces unprepared or closing the door to a peaceful solution. Five large language models (LLMs) including three different versions of *ChatGPT (OpenAI), Claude (Anthropic)*, and *Llama 2 (Meta)* were used in simulated wargames and diplomatic scenarios where they could make foreign policy decisions without human oversight[77]. All models tended to choose an aggressive approach, including using nuclear weapons[78]. Findings were statistically significant and a finding was that the systems sometimes made sudden changes[79].
- An intruded AI system can be turned against its controller or used as double agent (i.e., it sends observations of both sides to both sides).

AI models which combine learning algorithms with up to hundreds of hidden 'neural' layers and up to billions of parameters, which makes them to opaque black-box systems, this is known as **explainability** or **interpretability** issue[80].

Data poisoning: machines can be systematically misled by mislabeled data[81]. As AI heavily relies on data sets and data bases, the manipulation of data and the data poisoning by mislabeled data can mislead AI-driven technologies with corrupting or destroying data bases[82]. An escalation and misperception by AI-generated disinformation could also result from **deep fakes**[83]. Deliberate disinformation provided to the AI system, early warning or unmanned systems or emitters could trick an AI into believing that a nuclear strike is incoming[84].

---

[74] DARPA 2016

[75] Wright 2020, p.7, Depp/Scharre 2024. However, a machine algorithm may interpret the context of history, politics, ethics etc. only as confounding factors which bias straight-forward logical thinking. The most logical timepoint for a nuclear attack may then be not during conflicts where information analysis and target selection must happen under high time-pressure, but during stable peace-time where the machine has plenty of time to detect, analyze and select targets of the adversary which together with the surprise effect could minimize the damage from a potential retaliatory strike.

[76] Voke 2019 wrote in his analysis on page 33: *„Moreover, if AI is showing improper intentions or acting poorly, humans must be able to override its behavior. Although the system did not perform as required, the human must be able to exercise control once recognition of a hazardous situation occurs. Transparency is a requirement for control, and control is a requirement for trust."*

[77] GPT-3.5 (gpt-3.5-turbo-16k-0613), Claude-2.0 (claude-2.0), Llama-2-Chat (Llama-2-70b-chat-hf), GPT-4-Base (gpt-4-base). Eight fictious nations interacted with 27 discrete actions in three scenarios: a neutral scenario without any initial events, an invasion scenario where one nation invaded another before the start of the simulation. and a cyberattack scenario where one nation conducted a cyber attack on another before the start, Rivera et al. 2024

[78] Duboust 2024

[79] Rivera et al. 2024

[80] Arrieta et al. 2020, p.83, Chaudry/Klein 2023

[81] Wolff 2020

[82] Pauwels 2019, 2021

[83] Chaudry/Klein 2023

[84] Jalil 2023

Manipulated images can confuse autonomous systems. Already smallest changes in digital images can cause systematic misinterpretation by AI, a process known as **adversarial machine learning**[85].

Automation bias: humans may over-trust the technology and may also take too long to realize that the machine misinterpreted data or malfunctioned which is critical in nuclear conflicts where high information loads are combined with extreme time pressure[86].

Artificial escalation: inadvertent escalation in which AI systems make calculations originating from other AI systems creating a positive feedback loop that continuously escalates conflict[87]. Kaur defines artificial escalation in a different way "*as the risk of inadvertent escalation due to the potential for AI systems to misinterpret or misattribute signals, leading to miscalculations or unintended consequences*"[88].

For security reasons, it was suggested that weapon systems that can potentially use lethal autonomy should have a data recording function to document whether engagement decisions were made autonomously[89].

Cyber attacks: As any other software, AI is a complex software that is susceptible for cyber attacks. Generative AI can be attacked with manipulative instructions, the prompt injections. Typical attacks are prompt injections with direct commands, imagination, and reverse psychology. Prompt injections can also be used to conduct further attacks, facilitate the creation of malware, polymorphic viruses, ransomware, and other malevolent applications. Further problems of generative AI systems are hallucinations, contamination of search engines and the efflux of sensitive data. On the other hand, generative AI is also very useful for cyber defense for advanced data analysis, advanced pattern recognition, creation and analysis of threat repositories and code analysis[90].

Overall, the use AI is seen positive for basic functions like communication, design, testing etc. while concerns dominate for decision-making processes and autonomous launches[91]. The opacity, unpredictability, and susceptibility to cyberattacks are arguments not to involve AI in decision-making processes[92].

3.4 The Strategic Stability Debate

A main concern of all involved parties is the strategic stability, i.e., to avoid everything that gives a reason for a nuclear first strike[93]. Such a reason could be uncertainty about the capabilities of the adversary, because if it is not known what the other side can do in a certain situation, the only chance in a nuclear conflict is to strike first[94]. In a worst-case scenario, AI could make nuclear war 'winnable'[95] and has the potential to undermine nuclear deterrence by posing a threat to the second-strike capabilities of nuclear states.[96]

AI can undermine strategic stability also by compression of decision-making time, which may result in escalation or inadvertent use of nuclear weapons, amplify misunderstandings and misperceptions during a crisis, and by encouraging a premature deployment of insufficiently

[85] Wolff 2020
[86] Chaudry/Klein 2023
[87] Chaudry/Klein 2023
[88] Kaur 2024
[89] CNA 2023
[90] Saalbach 2023
[91] Saltini 2023a
[92] Saltini 2023a
[93] Shakirov 2023
[94] See also House of Lords 2023
[95] Boulanin 2019
[96] Rooth 2023

tested AI[97]. AI-enhanced conventional capabilities could also exacerbate the risk of inadvertent escalation caused by combination of nuclear and non-nuclear weapon systems[98].

Due to the rapid advances in AI technology, an extensive testing may take too long and the military AI is already outdated and behind the adversaries' AI systems, i.e., the military AI sector is both dynamic and unstable. As AI systems facilitate the use of dead hand systems and autonomous nuclear weapons such as the Russian *Perimeter* and *Poseidon*, they contribute to further instability. The rise of hypersonic weapons which cannot be timely detected are also critical[99], as the increasing speed of warfare will both undermine strategic stability and increase the risk of nuclear confrontation[100].

Currently, the AI is still perceived as too immature to be used in high-risk strategic situations[101]; there is a substantial risk of technical failures and of biased, incomplete, or inaccurate data[102]. The leading nuclear powers agree that for command and control at the strategic level, the role of AI should remain supportive[103].

4 Summary

This paper briefly presented the current state of the debate and the background. Artificial Intelligence has an enormous potential in military applications and is seen by nuclear powers as indispensable and essential. AI and big data could be used for control of non-proliferation of nuclear weapons, could save time and costs in the research, design optimization, manufacturing, testing and certification, maintenance and surveillance of nuclear warheads and manage resources more efficiently. Military AI could enhance early warning and Intelligence, Surveillance and Reconnaissance (ISR) capabilities, communication reliability and accelerate decision-making. The integration of AI into the nuclear command, control, and communications (NC3) systems could also synchronize information across nuclear and non-nuclear command and control. For autonomous nuclear-weapon systems, AI can support obstacle detection and maneuverability, automated target identification, and long-range and loitering capability. The debate on the use of AI for nuclear weapons covers three areas, the autonomy, the stability of military AI systems and the strategic stability.

A potential autonomy of nuclear weapons is part of the broader debate on lethal autonomous weapons systems (LAWS). Automated and semi-autonomous nuclear decision-makings were already considered during cold war. A specific military AI problem is the mission stability, as the AI systems lack context knowledge and may decide too quickly. The opacity of the systems leads to the explainability or interpretability issue; further problems may result from data poisoning, manipulated images, automation bias and artificial escalation. Wargame simulations with current generative AI *Large language models (LLMs)* from *OpenAI, Anthropic* and *Meta* showed that the systems tend to escalation up to nuclear strikes. As software system, AI is vulnerable for cyber attacks, generative AI also for prompt injections.

A main concern of all involved parties is the strategic stability, i.e., to avoid everything that gives a reason for a nuclear first strike. Such a reason could be uncertainty about the capabilities of the adversary, because if it is known what the other side can do in a certain situation, the only chance in a nuclear conflict is to strike first. AI can undermine strategic stability also by compression of decision-making time, which may result in escalation or inadvertent use of

[97] Fayet 2023
[98] Johnson 2020
[99] Hruby/Miller 2021
[100] Johnson 2020
[101] Saltini 2023a
[102] Jalil 2023
[103] Su/Yuan 2023

nuclear weapons, amplify misunderstandings and misperceptions during a crisis, and by encouraging a premature deployment of insufficiently tested AI. Moreover, AI systems facilitate the use of dead hand systems and autonomous nuclear weapons. The rise of hypersonic weapons and the increasing speed of warfare both undermine strategic stability as well.

Currently, the AI is still perceived as too immature to be used in high-risk strategic situations; there is a substantial risk of technical failures and of biased, incomplete, or inaccurate data. The nuclear powers agree that for command and control at the strategic level, the role of AI should remain supportive.

5 References

Air Force (2024): B-21 Raider Fact Sheet. Article 2682973 www.af.mil

Arrieta, A.B. et al. (2020): Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities, and challenges toward responsible AI. Information Fusion 58 (2020), p. 82–111

Bajak, F. (2023): Pentagon's AI Initiative accelerate hard decisions on lethal autonomous weapons. AP News 25 Nov 2023

Bommakanti, K. (2020): A.I. in the Chinese Military: Current Initiatives and the Implications for India Observer Research Foundation (ORF) Occasional Paper 234 February 2020

Boulanin, V. (2019): The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk Volume I Euro-Atlantic Perspectives Stockholm International Peace Research Institute (SIPRI) May 2019

Burns et al., (2023): Weak-To-Strong Generalization: Eliciting Strong Capabilities With Weak Supervision. Joint project paper of the OpenAI Superalignment Generalization team.

Chaudhry, H., Klein, L. (2023): Artificial Intelligence and Nuclear Weapons: Problem Analysis and US Policy Recommendations. policy@futureoflife.org 14th November 2023 Future of Life Institute (FLI)

CNA (2023): Arms Control and Lethal Autonomy CNA Corporation Analysis Paper

CoE (2022): Emergence of lethal autonomous weapons systems (LAWS) and their necessary apprehension through European human rights law Draft resolution unanimously adopted by the Committee on Legal Affairs and Human Rights of the Council of Europe on 14 November 2022 AS/Jur (2022)

Congress (2023): Block Nuclear Launch by Autonomous Artificial Intelligence Act of 2023 Bill Text (PDF) BUR23348 GC1

DARPA (2016): Cyber Grand Challenge https://www.cybergrandchallenge.com 05 Aug2016

Depp, M. and Scharre, P. (2024): Artificial Intelligence and Nuclear Stability. War on the Rocks 16 January 2024, warontherocks.com

Duboust, O. (2024): AI models chose violence and escalated to nuclear strikes in simulated wargames. https://www.euronews.com/next/2024/02/22/ai-models-chose-violence-and-escalated-to-nuclear-strikes-in-simulated-wargames. Updated 23/02/2024

DoD (2022): 2022 National Defense Strategy of The United States of America including the Nuclear Posture Review NPR 2022 and the Missile Defense Review MDR 2022, https://media.defense.gov/2022/Oct/27/2003103845/-1/-

DoD (2023a): DOD DIRECTIVE 3000.09. Autonomy In Weapon Systems. Originating Component: Office of the Under Secretary of Defense for Policy Effective: January 25, 2023 Releasability: Cleared for public release

DoD (2023b): DoD Data, Analytics, and AI Adoption Strategy. Cleared for open publication June 27, 2023, Department of Defense/Office of Prepublication and Security Review

Dresp-Langley, B. (2023): The weaponization of artificial intelligence: What the public needs to be aware of. Front. Artif. Intell. 6:1154184. doi: 10.3389/frai.2023.1154184

Fayet, H. (2023): French thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making. Report 13 November 2023 European Leadership Network (ELN)

Ford, C.A. (2020): Arms Control and International Security Papers Volume I, Number 2 I April 20, 2020 Office of the Under Secretary of State for Arms Control and International Security

Heslop, D, Keep, J. (2024): The 2024 China-US AI Dialogue Should start with an Eye on Chem-Bio Weapons. The Diplomat 09 March 2024

Hoadley, D.S., Sayler, K.M. (2019): Artificial Intelligence and National Security Congressional Research Service R45178 Version 6 Updated November 21, 2019

House of Lords (2023): Proceed with Caution: Artificial Intelligence in Weapon Systems. AI in Weapon Systems Committee. Report of Session 2023–24 HL Paper 16

Hruby, J., and Miller, M.N. (2021): Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapon Systems Nuclear Threat Initiative (NTI)

ISAB (2023): Report on the Impact of Artificial Intelligence and Associated Technologies on Arms Control, Nonproliferation, and Verification October 2023. A report of the International Security Advisory Board (ISAB), a Federal Advisory Committee for the Department of State

Jalil, G.Y. (2023): Artificial Intelligence and Nuclear Weapons: Way to the Future or Path to Disaster? Arms Control & Disarmament Centre, Institute of Strategic Studies Islamabad ISSI – Issue Brief. December 11, 2023

Johnson, J.S. (2020): Artificial Intelligence: A Threat to Strategic Stability. Strategic Studies Quarterly Spring 2020, p.16-39

Joint Statement (2022): Principles and responsible practices for Nuclear Weapon States 2022. Working paper submitted by France, the United Kingdom of Great Britain and Northern Ireland and the United States of America

Kasperowicz, P. (2023): Pentagon moving to ensure human control so AI doesn't 'make the decision for us'' Fox News 21 April 2023

Kaur, S. (2024): Artificial Intelligence and the Evolving Landscape of Nuclear Strategy Union of Concerned Scientists (UCS) https://blog.ucsusa.org/science-blogger/artificial-intelligence-and-the-evolving-landscape-of-nuclear-strategy/ March 4, 2024

Kölling, M. (2023): Künstliche Superintelligenz ist in Sicht. Neue Zürcher Zeitung, 06 Oct 2023, p.17

Longpre et al. (2022): Longpre, S., Storm, M. and Shah, R. MIT Science Policy Review August 29, 2022, vol. 3, pg. 47-55 This article is licensed under a Creative Commons Attribution 4.0 International License http://creativecommons.org/licenses/ by/4.0/

Lowther, A., McGriffin, C. (2019): America needs a dead hand. https://warontherocks.com/2019/08/america-needs-a-dead-hand/16 Aug 2019

Lowther, A., McGriffin, C. (2024): America needs a dead hand more than ever. https://warontherocks.com/2024/03/america-needs-a-dead-hand-more-than-ever/

Markey E.J. et al. (2023): Markey, Lieu, Beyer, and Buck Introduce Bipartisan Legislation to Prevent AI From launching a Nuclear Weapon. April 26, 2023 https://www.markey.senate.gov/news/press-releases/markey-lieu-beyer-and-buck-introduce-bipartisan-legislation-to-prevent-ai-from-launching-a-nuclear-weapon

Masuhr, N. (2019): AI in Military Enabling Applications. CSS Analyses in Security Policy No. 251, October 2019

NATO (2019): Artificial Intelligence: Implications for NATO's Armed Forces. Science and Technology Committee (STC) - Sub-Committee on Technology Trends and Security (STCTTS) Rapporteur: Matej Tonin (Slovenia) 149 STCTTS 19 E rev. 1 fin Original: English 13 October 2019

NNSA (2023): Artificial intelligence for nuclear deterrence strategy. National Nuclear Security Administration NNSA Department of Energy DOE/NNSA-0145

Norman, G. (2024): State Department wants China, Russia to declare that AI won't control nuclear weapons, only humans. Fox News, 02 May 2024

NTA (2021): DARPA issues AI exploration opportunity for Shell Project – Proposals due July, 27. nta.org

OpenAI (2024): https://openai.com/index/gpt-4o-and-more-tools-tochatgpt-free Update ChatGPT4o on 14 May 2024

Pauwels, E. (2019): The New Geopolitics of Converging Risks: The UN and Prevention in the Era of AI, United Nations University Centre for Policy Research, 29 April 2019.

Pauwels, E. (2021): Cyber-biosecurity: How to protect biotechnology from adversarial AI attacks. The European Centre of Excellence for Countering Hybrid Threats (Hybrid CoE). Hybrid CoE Strategic Analysis / 26 May 2021

Porter, T. (2023): The Pentagon is moving toward letting AI weapons autonomously decide to kill humans. Business Insider 22 Nov 2023

Puwal, S. (2024): Should AI be banned from nuclear weapon systems? NATO Review 12 April 2024

Raasch, J.M. (2023): Cheap drones can take out expensive military systems, warns former Air Force Pilot pushing AI-enabled force. Fox News online, 08 Dec 2023

Rivera, J.P. et al. (2024): Escalation Risks from Language Models in Military and Diplomatic Decision-Making. arXiv:2401.03408v1 [cs.AI] 7 Jan 2024

Rooth, C. (2023): The impact of Artificial intelligence on nuclear deterrence July 2023 Info Flash FINABEL - European Army Interoperability Centre.

Saalbach, K. (2023): Artificial Intelligence and Cyber Attacks. Working Paper. https://doi.org/10.48693/413

Saltini, A. (2023a): AI and nuclear command, control and communications: P5 perspectives. November 2023 European Leadership Network (ELN) Published under the Creative Commons Attribution-ShareAlike 4.

Saltini, A. (2023b): UK thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making Report 13 November 2023 European Leadership Network (ELN)

Sankaran, J. (2019): A different use for Artificial Intelligence in Nuclear Weapons Command and Control https://warontherocks.com/2019/04/a-different-use-for-artificial-intelligence-in-nuclear-weapons-command-and-control/

Sayler, K.M. (2023): Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems Congressional Research Service CRS Paper IF 11150 Updated May 15, 2023

Shakirov, O. (2023): Russian thinking on AI integration and interaction with nuclear command and control, force structure, and decision-making Report 13 November 2023 European Leadership Network (ELN)

Su, F., Yuan, J. (2023): Chinese thinking on AI integration and interaction with nuclear C2, force structure, and decision making. Report 13 November 2023 European Leadership Network (ELN)

US (2023): Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/ Bureau of Arms Control, Deterrence, and Stability November 09, 2023

Voke, M.R. (2019): Artificial Intelligence for Command and Control of Air Power. Wright Flyer Paper No. 72 Air University Press

White House (2023): FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/

Wolff, J. (2020): How to Improve Cybersecurity for Artificial Intelligence. Brookings Report 08 June 2020

Wright, N.D. (2019): Artificial Intelligence, China, Russia, and the Global Order Technological, Political, Global, and Creative Perspectives. Air University Press in October 2019

Zia, H. (2023): Information Revolution and Cyber Warfare: Role of Artificial Intelligence in Combatting Terrorist Propaganda Pakistan Journal of Terrorism Research, Vol-03, Issue-2, 133