Chapter 17
Nuclear Weapons and the Militarization of AI

Guglielmo Tamburrini

Abstract This contribution provides an overview of nuclear risks emerging from the militarization of AI technologies and systems. These include AI enhancements of cyber threats to nuclear command, control and communication infrastructures, proposed uses of AI systems affected by inherent vulnerabilities in nuclear early warning, AI-powered unmanned vessels trailing submarines armed with nuclear ballistic missiles. Taken together, nuclear risks emerging from the militarization of AI add new significant motives for nuclear non-proliferation and disarmament.

17.1 Introduction

The major powers are busy incorporating Artificial Intelligence (AI) into existing and emerging military systems [3, 15, 24]. Summarizing the drive towards pervasive military uses of AI technologies and systems, the US National Security Commission on Artificial Intelligence (NSCAI from now on) stated that AI-enabled technologies are going to be integrated into "every facet of warfighting" ([23], p. 79). In a similar vein, China's "New Generation Artificial Intelligence Development Plan" underscored the need to "promote all kinds of AI technology to become quickly embedded in the field of national defense innovation" [11]. And Russian President Vladimir Putin confidently claimed that whoever becomes the leader in AI will rule the world [26].

Inspired by similar objectives and aspirations, recent actions towards the militarization of AI include both the launch of NATO's innovation fund—which prioritizes investments into artificial intelligence, big-data processing, autonomous systems, and other dual-use emerging technologies [22]—and the use that the British Army made for the first time in 2022 of an AI system to process terrain and other contextual

G. Tamburrini (✉)
DIETI - Department of Electrical Engineering and Information Technology, Università di Napoli Federico II, Naples, Italy
e-mail: guglielmo.tamburrini@unina.it

P. Cotta-Ramusino et al. (eds.), *Nuclear Risks and Arms Control - Problems and Progresses in the Time of Pandemics and War*, Springer Proceedings in Physics 291,
https://doi.org/10.1007/978-3-031-29708-3_17

information during a military exercise in Estonia (https://www.gov.uk/government/news/artificial-intelligence-used-on-army-operation-for-the-first-time).

Nuclear defense systems are not exempted from ongoing military plans and actions focusing on AI technologies and systems. The US NSCAI recommended that "AI should assist in some aspects of nuclear command and control: early warning, early launch detection, and multi-sensor fusion." ([23], p. 104). Moreover, the emerging use of AI in cyber operations may increase the quantity and quality of cyber threats to nuclear command, control, and communication (NC3) infrastructures, thereby impinging on nuclear defense systems and nuclear escalation risk in conflicts.

This contribution provides an overview of major nuclear risks emerging on account of the militarization of AI. Section 17.2 reviews cyber threats to the military nuclear infrastructure and their impending enhancement by means of AI-powered cyber attacks. Section 17.3 examines proposals to use AI technologies to modernize nuclear early warning systems, and related risks arising from AI inherent vulnerabilities, brittleness, and information processing opacity. Section 17.4 briefly overviews wider destabilizing implications of AI on nuclear deterrence policies, arising from AI-powered, autonomously navigating vessels trailing submarines armed with ballistic missiles, AI deepfakes eroding the credibility and consistency of political leaders of nuclear powers, and autonomous weapons systems having the potential to tilt conventional military equilibria and to provide adversaries with new incentives to threat the use of nuclear weapons to avoid defeat. Section 17.5 concludes.

17.2 AI-Powered Cyberthreats and Nuclear Escalation Risks

Cyber attacks are, roughly speaking, cyber operations targeting computers and computer information systems, networks, or infrastructures, with the aim of stealing, exposing, altering, disabling, and destroying data or disrupting their normal computational processing. Cyber attacks are carried out using a variety of tools and methods, including malware, phishing, and denial of service.

The question has often been raised whether one might respond to a cyberattack not in kind, but rather using conventional weapons deployed in other traditional warfare domains or even weapons of mass destruction. An instance of this general question is whether (and, if so, in which circumstances) one might consider employing nuclear weapons to respond to a cyber attack. The latter question is meaningfully related to a general claim made in the 2018 US Nuclear Posture Review, according to which one may consider the use of nuclear weapons "in extreme circumstances to defend the vital interests of the US, its allies, and partners." It is further specified there that extreme circumstances "could include significant non-nuclear strategic attacks."

([31], p. 21). And to exemplify, two broad scenarios are mentioned there which may qualify as significant non-nuclear strategic attacks:

(i) attacks on the US, allied, or partner civilian population or infrastructure;
(ii) attacks on US or allied nuclear forces, their command and control, or warning and attack assessment capabilities ([31], p. 21).

Both scenarios are described in terms of their objectives and effects, without introducing any restriction on the types of weapons employed to achieve those objectives and effects. This approach agrees with a US Air Force doctrine document stating that a strategic attack "is not defined by the use of a particular weapon or the focus on a specific target" ([2], p. 3). Its strategic character depends instead by the goal of achieving "war-winning effects by the most direct, effective, and efficient means possible. [A strategic attack] disrupts critical leadership functions, infrastructure, and strategy, achieving results by affecting the psychological, cognitive, and behavioral aspects of warfare." ([2], pp. 3–4).[1]

If no restrictions are placed on the kinds of weapons to achieve strategic effects, one may legitimately ask whether a cyberattack might count as a significant non-nuclear strategic attack. If a positive answer is given to this question, one may further ask—in the light of both the 2018 Nuclear Posture Review and similar positions expressed in the more recent 2022 Nuclear Posture Review—whether there are conceivable circumstances licensing the use of nuclear weapons as a response to a significant non-nuclear strategic cyberattacks.

Ch. Ford addressed these questions at a time when he was assistant secretary of state for international security and non-proliferation: "Lest there be any confusion about whether a cyber attack could potentially constitute a 'significant non-nuclear strategic attack', I can say with confidence that it most certainly could if it caused kinetic effects comparable to a significant attack through traditional means." [14]. Commenting on this claim, Herbert Lin remarked that "the proposal for possible first use of nuclear weapons in response to devastating cyber attack is likely less of a departure from previous policy than it might seem" ([21], p. 28).

To begin with, let us consider Ford's claim in the framework of scenario (i), that is, in connection with "attacks on the US, allied, or partner civilian population or infrastructure". Cyberattacks against civilian infrastructures are legion, their list is rapidly expanding, and their disruptive effects are increasing. Cyber attacks which made the headlines for their relatively significant effects were carried out against the Colonial Oil Pipeline in 2021 and against the Ukrainian power facility PrykarpattyaOblenergo in 2015. The malware attack against the Colonial Oil Pipeline—which supplied almost half of diesel, gas, and jet fuel needed in the US East Coast—resulted in the shutdown of this facility for a few days. The cyberattack to PrykarpattyaOblenergo disrupted power supply and affected more than 200,000 consumers for up to 6 h. Clearly, neither one of these events qualifies as a significant non-nuclear strategic attack on civilian population and infrastructure.

[1] See Air Force Doctrine Publication 3–70, *Strategic Attack,* 21 November 2021, pp. 3–4.

Can a cyber attack conceivably achieve the effects of a significant non-nuclear strategic attack on civilian population and infrastructure? Undoubtedly, the artisanal character of activities that human operators must perform through all stages of the cyber kill chain is a bottleneck hindering their speed, volume, and destructiveness. Time-consuming and labor intensive operations include identifying software and hardware vulnerabilities, developing suitable tools for attack delivery, exploring target environments, and taking on command and control of penetrated systems. But what about *future* developments of cyber warfare in the light of ongoing attempts to remove or mitigate this bottleneck? Are cyber attacks bound stay below the threshold of significant non-nuclear strategic attacks, if these labor intensive activities get automated?

Initiatives leveraging on AI systems based on machine learning methods are well under way to automate and increase the speed, volume, and destructiveness of cyber attacks. Even though there are presently "no publicly known reported cases of machine learning being used to directly attack a system or an application" (Abaimov Martellini, p. 122), the 2021 NSCAI final report warns that "machine learning has current and potential applications across all the phases of cyber attack campaigns and will change the nature of cyber warfare and cyber crime. The expanding application of existing AI cyber capabilities will make cyber attacks more precise and tailored, further accelerate and automate cyber warfare, enable stealthier and more persistent cyberweapons, and make cyber campaigns more effective on a larger scale." (NSCAI, pp. 50–51). Accordingly, current and prospective applications of machine learning methods to expand the cyber capabilities of AI systems must be continually reviewed to assess the potential destructiveness of cyber attacks and their impact on future warfare. Even though known cyber attacks to civilian infrastructures have not caused kinetic effects comparable to a significant strategic non-nuclear attack, one cannot exclude that this situation will radically change in the future.

Let us now turn to consider the evoked nuclear response to cyber attacks in the framework of scenario (ii), that is, "attacks on US or allied nuclear forces, their command and control, or warning and attack assessment capabilities". The computational infrastructure of the US nuclear defense complex offers an extended cyber attack surface, that is, software or hardware elements that one may explore and penetrate for malicious purposes. These elements are notably present in off-the-shelf hardware and software used by military contractors and available to hackers for examination, in computing modules embedded into nuclear weapons delivery systems, and in NC3 software supporting nuclear planning and situational awareness ([21], pp. 38–90). This extended cyber attack surface raises doubts about the reliability and integrity of nuclear weapons systems, especially concerning the ability to launch a weapon when authorized or to prevent its inadvertent launch, to maintain uninterrupted command and control over nuclear weapons, or to preserve the functionality of nuclear communications ([28], p. 3).

Moreover, the modernization of nuclear defense systems may further extend the cyber attack surface, due to the implementation of new and more advanced computational functionalities. In general, Lin emphasized the tension between modernization of the military nuclear complex and its cybersecurity needs, suggesting that a sensible trade-off must be reached, moderating appetites for added computational functionalities in the light of attending cybersecurity risks ([21], pp. 123–4). The prospect of AI-powered cyberattacks can only exacerbate these concerns about the cybersecurity of a more extensively computerized military nuclear complex.

Let us take stock. Past cyber attacks to civilian infrastructures have not caused kinetic effects expected from a significant strategic non-nuclear attack. Accordingly, these events are not meaningfully related to scenario (i) envisaged in the 2018 Nuclear Posture Review. However, this situation must be continually reassessed, in the light of machine learning methods being used to expand the cyber capabilities of AI systems, to automate cyber warfare, and to increase the speed, volume, and destructiveness of cyber attacks. Additional concerns about scenario (ii) arise from modernization plans involving the extension of computing infrastructures for the military nuclear complex and AI-powered cyber attacks on this infrastructure compromising the reliability and integrity of nuclear weapons systems.

These cyber risks and their potential exacerbation flowing from AI applications in the cyber kill chain call for the establishment of permanent venues to discuss AI's impact on nuclear crises and stability. The AI research community has a central role to play in this context—to foster dialogue and exchange scientific information, to advance specific trust and confidence building measures, and to raise the awareness of political decision makers and public opinion on AI-related cyber risks affecting nuclear weapons systems and infrastructures.

17.3 Nuclear Early Warning and AI Misclassification Risks

The NSCAI final report emphasized that the decision to authorize the employment of nuclear weapons should firmly remain in human hands and should never be delegated to an AI-enabled system. It is further recommended there that that the US "should include a statement to this effect in the next Nuclear Posture Review and should seek an analogous commitment by Russia, China, and other nuclear powers." ([23], p. 98). At the same time, however, NSCAI recognized some role for AI to play in the modernization of NC3: "AI should assist in some aspects of the nuclear command and control apparatus: early warning, early launch detection, and multi-sensor fusion, to validate single sensor detections and potentially eliminate false detections" ([23], p. 104, n. 22). Let us critically examine this claim, pointing to risks distinctively arising from this suggested use of AI technologies *within* nuclear early warning systems.

Early warning systems play a central role in nuclear deterrence based on second strike retaliation capabilities. By increasing automation of early warning systems, one expects to reduce information processing time and to buy more extended temporal windows for human decision-making, thereby alleviating the enormous pressure involved in evaluating whether a nuclear attack is in progress and deciding what is to be done. As noted in Borrie ([8], p. 49), "[i]n the absence of declassified information about current nuclear early-warning and command-and-control systems, it is difficult to assess the pros and cons of AI-enabling aspects of these systems." Nevertheless, independently of the availability of this detailed information, one may still identify a variety of potential risks depending on distinctive features of automation in general, and AI-powered automation in particular. Automation bias is one of these risks, which applies to AI-powered automation, but is not specific to it. Indeed, the tendency to trust machine decision-making more than contrasting human judgments has been observed across a variety of automation technologies, leading to accidents in both civilian and military application domains. Other risks are specific to the use of AI in nuclear early-warning and in other critical domains. In what follows, we focus on distinctive fragilities affecting deep neural networks (DNN), motivated by the fact that these AI systems have contributed most to determine the pervasive impact of AI over the last decade and across a variety of application domains.

To train a learning AI system formed by a DNN to perform early launch event detection, one usually relies on the availability of relevant "big data", that is, vast amounts of sensor data about launch and non-launch situations. Accordingly, the scarcity of nuclear launch event data may hamper the training process, negatively affecting downstream the accuracy of the trained AI system. Let us suppose, for the sake of argument, that this preliminary bottleneck can be overcome, that enough training data can be collected or synthetically generated, and that an AI system trained on these data is found to achieve "good" classification accuracy on early launch detection. One should carefully note that the estimated achievement of good classification accuracy does not exclude the occurrence of mistakes, for the possibility of an error is intrinsic in the statistical nature of AI decision-making. Clearly, a mistake occurring in nuclear early warning, no matter how infrequent, may have existential implications: an AI classifier detecting a false positive launch of intercontinental ballistic missiles (ICBMs) may trigger an unjustified use of nuclear weapons.

The high risk associated to an infrequently occurring early warning mistake demands that the responses of AI systems must be carefully verified by human decision-makers. However, the additional time required by this verification process may offset the expected reduction in processing time allegedly flowing from AI-powered automation, thereby defeating the goal of buying more extended temporal windows for human decision-making, which is one of the pros one may adduce for using AI in nuclear early warning.

Additional risks involved in AI-powered nuclear early warning emerge by reflecting—in connection with the need to countervail automation biases—on the difficulty of interpreting AI information processing and explaining its outcomes. Humans in command-and-control positions are expected to act on the basis of a proper understanding of machine behaviors, rather than blindly trusting its responses. Hence, they must be able to obtain enough humanly understandable information about machine information processing, mapping the latter into perceptual and cognitive domains that humans can make sense of. However, AI learning systems based on DNN raise major stumbling blocks towards the fulfilment of this "interpretability" requirement. Indeed, classification outcomes of AI learning systems depend on features of input data that may significantly differ from features that humans use to perform the same classification task. To illustrate, to decide whether there is a cat in an image, humans usually focus on salient features of typical cats—such as whiskers, ears, nose, and tail—and their spatial arrangement. In contrast with this, AI image classification processes may rely on distributed sets of image parts and pixels that the human perceptual system does not meaningfully associate, as a rule, to distinctive features of cats.

The “semantic gap” between human and machine knowledge representation and processing ([18], p. 20) extends well beyond the identification of salient features in input data. AI learning systems process information *subsymbolically*, without operating on humanly understandable declarative statements and without applying stepwise logical or causal inference [25]. Because of these remarkable differences between machine and human information processing, AI systems are mostly opaque and hardly interpretable to human users and decision-makers.

Another risk arising from the use of AI systems in nuclear early warning flows from the unexpected and counterintuitive mistakes that AI systems make and that a human operator would unproblematically detect and avoid. These fragilities were discovered by means of adversarial machine learning [7] early on in the history of learning AI systems formed by DNNs. A variety of errors were identified that are most relevant to military uses of AI systems. Notably, visual perceptual systems based on DNN architectures were found to mistake images of school buses for ostriches [27] and 3-D renderings of turtles for rifles [6]. These mistakes were induced by small input perturbations crafted on the basis of adversarial machine learning methods. A human operator would not incur in such mistakes, for the small adversarial input perturbations inducing the machine to err are hardly noticeable by the human perceptual system. Clearly, these mistakes are potentially catastrophic in a wide variety of conventional warfare domains, for normal uses of school buses are protected by International Humanitarian Law, and someone carrying a harmless object in the hand may be mistakenly taken by an AI system to wield a weapon, thereby triggering an unjustified use of force [4]. By the same token, one cannot exclude that AI systems for nuclear early warning will make counterintuitive and potentially catastrophic errors of the same sort that adversarial machine learning has enabled one to highlight in other critical application domains.

To detect and correct machine errors that human operator would easily prevent, nuclear decision makers should be put in a position to understand the reasons *why* an AI-powered early warning system provided a certain classification of sensor data. To fulfil this “explainability” condition, AI systems should be endowed with the capability to provide elements of a good and humanly understandable explanation of their decisions and classification results. However, causal or logical forms of reasoning are often needed to provide these explanations. But logical and causal reasoning, as already noted above, is beyond the current capabilities of the more successful AI learning systems. Accordingly, the development of explanation-giving AI systems raises formidable research problems, which now characterize the goals of the eXplainable AI (or XAI in brief) research area (https://www.darpa.mil/program/explainable-artificial-intelligence). Pending significant breakthroughs in XAI, one cannot but acknowledge the difficulty of fulfilling interpretability and explainability conditions that are necessary for nuclear decision makers supported by early warning AI systems to achieve the required situational awareness.

Adversarial machine learning demonstrates the possibility of *accidental* misclassifications leading to surprising and potentially disastrous mistakes. More recent developments in this research area have additionally showed that *deliberate* adversarial attacks can be maliciously exploited to induce an AI system to make classification mistakes. Indeed, adversarial AI attacks have been systematically carried out against AI systems operating in the real world. By altering the illumination of a stop signal on the street—in ways that are hardly perceptible to human eyes—an AI system was induced to classify it as a 30-mph speed limit sign [16]. To carry out this optical attack, AI scientists used inexpensive and readily available equipment only: a low-cost projector, a camera, and a computer. These developments pave the way to intentional adversarial attacks which manipulate input data, inducing AI-powered early warning systems to make perceptual mistakes. Similar hostile motivations may prompt intentional attacks of a different kind, carried out by "poisoning" AI learning systems. Poisoning attacks aim at corrupting datasets for learning, degrading the learning procedure or even the resulting AI system. There are no patches available to avoid either input manipulation or poisoning attacks, insofar as these are based on inherent weaknesses of the deep learning methods and systems that are prevalent today [12].

Let us take stock. Alleged advantages one may expect to flow from automated AI systems supporting nuclear early warning include a reduction of information processing time, buying more extended temporal windows for humans to assess whether a nuclear attack is in progress. However, specific AI fragilities discussed in this section, leading to accidental or intentionally induced counterintuitive misclassifications, erode confidence in the reliability of this technology in this high-risk application area. To avoid disastrous consequences of false positive or false negative early warning classifications, human decision makers should do their best to control the correctness of responses produced by AI classifiers. But this process is hindered by the lack of transparency and explainability of AI information processing and its outcomes, so that its enactment may offset reductions of processing time allegedly flowing from AI-powered automation. Therefore, the NSCAI recommendation that "AI should assist in some aspects of the nuclear command and control apparatus," including early warning and early launch detection ([23], p. 104, n. 22), cannot be taken at face value, but stands in need of a thorough critical assessment taking in due account fragilities, opacities, and unintended consequences of AI classification mistakes.

17.4 Wider Implications of AI for Nuclear Stability

Potential impacts of AI technologies and systems on nuclear defense and stability extend well beyond AI enhancements of cyber threats to NC3 and the envisaged use of AI-powered systems for nuclear early warning. Unmanned vessels, whose autonomous navigation capabilities are powered by AI systems, are likely to have a significant impact on the prong of nuclear deterrence which is based on SLBMs (submarine launched ballistic missiles). Autonomous unmanned vessels are being developed as new elements of anti-submarine warfare, to identify submarines as they emerge from port or at maritime chokepoints and to trail them for extended periods of time henceforth. The surface vessel Sea Hunter is a significant case in point: originally prototyped in the framework of the DARPA anti-submarine warfare ACTUV (Autonomous Continuous Trail Unmanned Vessel) program, Sea Hunter is now undergoing further development by the US Office of Naval Research, to perform autonomous trailing missions lasting up to three months (https://www.darpa.mil/news-events/2018-01-30a). Another case in point is the autonomous extra-large unmanned undersea vehicle (XLUUV) Orca, manufactured by Boeing to meet a variety of undersea operations including anti-submarine trailing missions and warfare (https://www.naval-technology.com/projects/orca-xluuv/). According to a report of the National Security College of the Australian National University, "oceans are, in most circumstances, at least likely and, from some perspectives, very likely to become transparent by the 2050s." In particular, counterdetection technologies will be of little or no avail, so that submarines carrying ballistic missiles will be "detected in the world's oceans because of the evolution of science and technology." ([5], p. 1). A similar suggestion was advanced earlier on in a 2016 British Pugwash report in connection with SLBM-enabled Continuous At Sea Deterrence (CASD): "…adaptable long-endurance or rapidly-deployable unmanned underwater vehicles (UUV) and unmanned surface vehicles (USV), look likely to undermine the stealth of existing submarines." [10].

AI systems are used to generate synthetic data called *deepfakes*—a word blending the expression *deep learning* with the word *fakes*. Malicious uses of AI deepfake technology include the fabrication of videos imitating political leaders. These increasingly realistic and deceitful videos may induce misconceptions about the personality, behaviors, political positions, and actions of the represented political leaders. Deepfake videos of leaders of nuclear powers like Barack Obama, Donald Trump, and Vladimir Putin were widely circulated. Fueling doubts about their consistency and rationality, these videos may undermine the effectiveness of nuclear deterrence policies, which are crucially based on the credibility of second-strike threats to deter first uses of nuclear weapons.

The race to the militarization of AI was initially fueled by the rise of autonomous weapons systems (AWS). These are AI-enabled weapons systems that select and apply force to targets without human intervention [19, 30]. Instances include loitering munitions and autonomous drones. Loitering munitions overfly an assigned area in search of targets to dive-bomb and destroy without requiring any further human intervention after their activation. The loitering munition Kalashnikov ZALA Aero KUB-BLA was allegedly used by Russian forces in Ukraine [20]. And the Turkish unmanned aerial vehicle STM Kargu-2 was reportedly employed in autonomous attack mode during the Second Libyan Civil War against Haftar-affiliated forces. The repertoire of existing autonomous weapons is continually expanding, with an initial comprehensive survey provided in [9]. AWS raise serious concerns about the respect of IHL in conventional conflicts [4, 19]. Moreover, AWS have the potential to give large conventional military advantages to adopters. It was claimed in this connection that if AWS happen to tilt the conventional military balance, a nuclear-armed adversary may feel incentivized to threat the use of nuclear weapons to avoid military defeat ([17], p. 31).

17.5 Concluding Remarks

The decision that Lieutenant Colonel Stanislav Petrov made on September 26th, 1983, is an enduring lesson about nuclear risks arising from technological efforts to automate nuclear early warning systems and reduce the role of human judgment. The Soviet early warning system OKO wrongly signaled an incoming nuclear attack as it mistook sensor readings of sunlight reflecting on clouds for signatures of ICBMs engines. However, Petrov correctly concluded that OKO had signaled a false positive. Commenting some years later on the mental processes that led him to the conclusion that saved humanity from a nuclear war, Petrov remarked that "when people start a war, they don't start it with only five missiles." (https://www.armscontrol.org/act/2019-12/focus/nuclear-false-warnings-risk-catastrophe). This is an instance of human commonsense reasoning at its best. In contrast with this, AI still lacks commonsense reasoning, unable to respond properly and often revealing its brittleness to changing contextual situations that fall outside the scope of narrow sets of assumptions and boundary conditions [13].

The brittleness to changing context and the inherent vulnerabilities of AI information processing clearly support the US National Security Commission on Artificial Intelligence recommendation to "clearly and publicly affirm existing U.S. policy that only human beings can authorize employment of nuclear weapons", and to include "such an affirmation in the DoD's next Nuclear Posture Review", seeking "similar commitments from Russia and China" ([23], p. 98). There are, however, additional risks arising from AI inherent vulnerabilities and information processing weaknesses, especially in connection with plans and proposals to use AI to modernize nuclear early warning, including NSCAI own proposals in this respect. AI Computer scientists can and should do better to highlight risks to nuclear stability induced by limitations affecting current AI technologies and systems, by AI-powered cyber attacks, and by AI-induced erosion of nuclear deterrence. Taken together, nuclear threats emerging from the militarization of AI reveal additional limitations of nuclear deterrence policies and provide new significant motives to support nuclear non-proliferation and disarmament.

Acknowledgements Support by MUR, Italian Ministry for University and Research, grant PRIN 2020SSKZ7R, is gratefully acknowledged.

References

1. S. Abaimov, M. Martellini, *Machine Learning for Cyber Agents. Attack and Defense* (Springer, Cham, 2021). Book available at https://doi.org/10.1007/978-3-030-91585-8
2. AFDP Air Force Doctrine Publication 3-70, *Strategic Attack* (2021)
3. D. Amoroso, D. Garcia, G. Tamburrini, The weapon that mistook a school bus for an ostrich. Sci. Dipl. (2022). https://www.sciencediplomacy.org/article/2022/weapon-mistook-school-bus-for-ostrich
4. D. Amoroso, G. Tamburrini, Toward a normative model of meaningful human control over weapons systems. Ethics Int. Aff. **35**(2), 245–272 (2021). https://doi.org/10.1017/S0892679421000241
5. ANU-NSC, *Transparent Oceans? The Coming SSBN Counter-Detection Task May Be Insuperable.* Australian National University—National Security College Publication (2020). https://nsc.crawford.anu.edu.au/publication/16666/transparent-oceans-coming-ssbn-counter-detection-task-may-be-insuperable
6. A. Athalye, L. Engstrom, A. Ilyas, K. Kwok, Synthesizing robust adversarial examples. Proceed. Mach. Learn. Res. **80**, 284–293 (2018). https://proceedings.mlr.press/v80/athalye18b.html.
7. B. Biggio, F. Roli, Wild patterns: ten years after the rise of adversarial machine learning. Pattern Recogn. **84**, 317–331 (2018)
8. J. Borrie, Cold war lessons for automation in nuclear weapon systems, in *The impact of Artificial Intelligence on strategic stability and nuclear risk, vol. I: Euro-Atlantic Perspectives*, ed. by V. Boulanin (SIPRI, Stockholm, 2019), pp. 41–52. https://www.sipri.org/publications/2019/other-publications/impact-artificial-intelligence-strategic-stability-and-nuclear-risk-volume-i-euro-atlantic
9. V. Boulanin, M. Verbruggen, *Mapping the development of autonomy in weapon systems* (Stockholm International Peace Research Institute, Solna, 2017)
10. S. Brixey-Williams, *Will the Atlantic Become Transparent? British Pugwash Report*, 3rd edn, 2018. (2016). Pugwash_Transparent_Oceans_update_nov2016_v3b_April2018–1.pdf
11. China State Council, *New Generation Artificial Intelligence Development Plan* (translation) (2017). www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017
12. M. Comiter, *Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It* (Belfer Center Science and International Affairs, Harvard Kennedy School, 2019). www.belfercenter.org/publication/AttackingAI
13. M.L. Cummings, Rethinking the maturity of artificial intelligence in safety-critical settings. AI Mag. **42**(1), 6–15 (2021)
14. Ch. Ford, International security in cyberspace: new models for reducing risk, in *Arms Control and International Security Paper Series*, vol. I, no. 20 (2020). https://www.newparadigmsforum.com/p2818
15. D. Garcia, Stop the emerging AI cold war. Nature **593**(7858), 169 (2021)
16. A. Gnanasambandam, A.M. Sherman, S.H. Chan, Optical adversarial attacks, in *IEEE/CVF International Conference on Computer Vision Workshops* (ICCVW 2021) (2021), pp. 92–101
17. M.C. Horowitz, P. Scharre, A. Velez-Green, A stable nuclear future? the impact of autonomous systems and artificial intelligence (2019). arXiv:1912.05291v2
18. ICRC, Artificial intelligence and machine learning in armed conflict: a human-centred approach, in *International Committee of the Red Cross Report* (Geneva, 2019). https://www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach
19. ICRC, ICRC position on autonomous weapons systems, in *International Committee of the Red Cross Report* (2021). www.icrc.org/en/document/icrc-position-autonomous-weapon-systems
20. Z. Kallenborn, Russia may have used a killer robot in Ukraine. Now what? Bull. Atom. Sci. (2022). https://thebulletin.org/2022/03/russia-may-have-used-a-killer-robot-in-ukraine-now-what

21. H. Lin, *Cyber Threats and Nuclear Weapons* (Stanford UP, 2021)
22. NATO, NATO launches innovation fund, News release (2022). https://www.nato.int/cps/en/natohq/news_197494.htm
23. NSCAI, *Final Report of the US National Security Commission on Artificial Intelligence* (2021). www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf
24. K. Payne, *I, warbot: The dawn of artificially intelligent conflict* (C. Hurst & Co., London, 2021)
25. J. Pearl, D. Mackenzie, *The Book of Why. The New Science of Cause and Effect* (Penguin, London, 2019)
26. Russia Today, *Whoever Leads in AI Will Rule the World: Putin to Russian Children on Knowledge Day* (2017). www.rt.com/news/401731-ai-rule-world-putin
27. Ch. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, *Intriguing Properties of Neural Networks* (2014). arXiv.org. https://arxiv.org/abs/1312.6199v4
28. B. Unal, P. Lewis, Cybersecurity of nuclear weapons systems: threats, vulnerabilities and consequences. Chatham House Rep (2018)
29. United Nations, Final Report of the Panel of Experts on Libya Established Pursuant to Security Council Resolution 1973 (2011), March 8, 2021, UN Doc. S/2021/229, para. 63 (2021). https://digitallibrary.un.org/record/3905159?ln=en
30. US DoD, Department of Defense, Autonomy in Weapons Systems (Directive 3000.09) (2012). www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf. Accessed 21 Nov 2012
31. US DoD, Nuclear Posture Review 2018, US Department of Defense (2018). https://dod.defense.gov/News/SpecialReports/2018NuclearPostureReview.aspx. Accessed Feb 2018