Preemptive Discussions: the Potential Implications of Integrating Deep Learning into Early Warning Systems

The Disruptive Technology Working Group

Authors: Alisha Anand, Liza Arias, Belen Bianco, Fabian Hoffmann, Artur Honich, Natasha Karner, Niels Renssen, Elisabeth Suh, Lydia Wachs, Alexa W

Abstract

Early warning systems (EWS) are a critical part of the global nuclear command, control, and communications (NC3) enterprise. As nations begin to modernise these systems, discussion of further integration of artificial intelligence (AI) and machine learning (ML) approaches into various aspects of NC3 systems has publicly (and presumably privately) emerged via international expert communities. AI and ML are concepts that have become 'buzz words' but are often discussed without reference to the exact meaning and context in which they could be applied. This paper seeks to explore the areas in which EWS could be subject to the integration of novel machine learning (ML) techniques, particularly deep learning (DL) — an integration which is presumably already occurring in the intelligence, surveillance, and reconnaissance (ISR) spheres of the nuclear realm. In doing so, the authors hope to both raise awareness of

this ML technique given its rise in popularity across various sectors, as well as provide an example of the way in which novel technologies could be discussed in the nuclear context. In order to assess the consequences of such integration, it is necessary for stakeholders to both understand the technology and discuss its significance in open fora. Although knowledge is limited and security concerns remain, such discussions are vital to encouraging transparency and risk reduction as well as mitigating negative implications. The following analysis offers actionable processes for stakeholders regarding the potential opportunities and risks associated with the potential integration of DL into EWS in order to mitigate risks and increase crisis stability.

Defining Nuclear Early Warning Systems and Deep Learning

EWS provide states with both an awareness of incoming missiles and the ability to respond. These are essential building blocks for first strike survivability, second strike capability, and strategic stability to various degrees in all nuclear possessor states. Furthermore, EWS are also important for those states that benefit from (extended) nuclear deterrence, including those that possess or contribute components.

Modern EWS consist of a network of satellites and ground and air-based radars. In the pre-launch phase, these components allow states to engage in ISR missions, such as monitoring international nuclear force readiness. In the boost phase, infrared sensors on the satellites detect heat signatures released during missile launch, track missiles during early flight stages and communicate trajectories to their respective command centers. Ground and air-based radar systems track missiles in later flight stages during the midcourse and terminal phases of flight.[32] For states without advanced satellite constellations, radars may be the only viable detection mechanism within their EWS. Information from EWS systems can also be used for operations with preemptive or preventive purposes.

Broadly, EWS have implemented AI and ML processes, to various extents, for decades. An increase in the commercial application of DL approaches paired with aging legacy systems, advancements in delivery systems, increases in dual-capability and various global calls for nuclear modernisation, could incentivise states to consider integrating DL into EWS modernisation attempts.[33]

ML (in broad terms) consists of a computer finding ideal parameters of an unknown function. DL is an approach to ML composed of logical structures of algorithms, with at least one 'hidden layer', modeled after the biological brain, called artificial neural networks.

> *"Deep learning uses networks that contain layers of nodes that in some ways mimic the neurons in the brain. Each layer of neurons takes the data from the layer below it, performs a calculation, and provides its output to the layer above it."*[34]

---

32 Kingston Reif, "Missile Defense Systems At a Glance," Arms Control Association Fact Sheets & Briefs, accessed 29 March 2021, https://www.armscontrol.org/factsheets/missiledefenseataglance#q2.

33 See, for example: Page O. Stoutland, "Artificial Intelligence and the modernization of U.S. nuclear forces," in *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk* Volume 1 Euro-Atlantic Perspectives ed. By Vincent Boulain (SIPRI, May 2019), 63-67, https://sipri.org/publications/2020/other-publications/impact artificial-intelligence-strategic-stability-and-nuclear-risk-volume-iii-south-asian; Lora Saalman, "China's Artificial Intelligence-Enabled Offense: Hypersonic Glide Vehicles and Neural Networks," in *Artificial Intelligence, China, Russia, and the Global Order: Technological, Political, Global, and Creative Perspectives* ed. by Nicholas D. Wright (Alabama: Air University Press, 2019), 162-168, https://bit.ly/3sBkQb3.

34 Ben Buchanan and Taylor Miller, *Machine Learning for Policymakers: What It Is and Why It Matters* (Cambridge, MA: Belfer Center for Science and International Affairs, 2017), 14-15, https://www.belfercenter.org/sites/default/ files/ files/publication/MachineLearningforPolicymakers.pdf.

Ongoing DL research primarily aims to further improve its input-efficiency, precision, and speed as well as algorithms' ability to process unknown information or unpredictability in data.[35] For these reasons, among others, plausible integration of DL into the pre-launch and launch phase of EWS would most likely occur within extremely specific areas with specific tasking, such as components that address image and audio recognition and processing.

Opportunities for DL- Enhanced EWS

The net effects of technological developments and their application are not predetermined, nor static across time and space. Novel technologies can prove more or less disruptive depending upon the context in which they are introduced and the ways in which humans interact with them. Given the advantages of DL in efficiently processing and fusing large amounts of data and self-improving its own accuracy over time, careful integration into EWS can mitigate the risk of accidental or inadvertent escalation due to false positives (i.e. detection of non-existent pre-launch/boost activities), enhance strategic stability by reducing the likelihood of false negatives (i.e. non-detection of actual pre-launch/boost activities),[36] and afford decision-makers more time to consider necessary and proportionate courses of action in the case of true positives (i.e. detection of actual pre-launch/boost activities).[37]

The integration of DL into EWS creates significant opportunities during the pre-launch phase and the boost phase. In the pre-launch phase, the integration of DL into EWS could allow for an increase in accuracy for computer vision, particularly with regard to aspects of EWS that are tasked with image classification, pattern recognition and anomaly detection. It could enhance ISR capabilities, thereby facilitating awareness of the existence, nature, and imminence of threats before they materialise (i.e., before missiles are launched). During the boost phase, integration of DL could improve the overall performance of EWS by progressively enhancing accuracy, speed, and data processing capacity.[38] DL-enhanced EWS could provide human decision-makers with more time to both decide upon and execute a range of countermeasures, such as 1) pre-emption, 2) crisis diplomacy, and 3) target protection.

First, the integration of DL could improve so-called 'left-of-launch' capabilities such as "persistent overhead coverage and all-weather ISR; and rapid processing, exploitation, and dissemination of targeting information."[39]

Second, the implementation of DL into EWS could allow for an increased output specificity based on the specific environment in which it is applied.[40] In addition, DL approaches can provide increased efficiency and

---

35 Additionally, increasing advances in quantum computing "could enable a quantum approach to deep learning neural networks for greatly enhanced AI and data analytics"; see: NATO Science & Technology Organization, *Science & Technology Trends 2020-2040 Exploring the S&T Edge*, (Brussels: NATO, Report March 2020), 23.

36 Lora Saalman, "Fear of False Negatives: AI and China's Nuclear Posture," *Bulletin of the Atomic Scientists*, April 24, 2018, https://thebulletin.org/2018/04/fear-of-false-negatives-ai-and-chinas-nuclear-posture/. Thomas C. Schelling and Morton H. Halperin, *Strategy and Arms Control* (New York: Twentieth Century Fund, 1961), 50.

37 Michal Onderco and Madeline Zutt, "Will Humans and/or Machines Save Us From Nuclear Doomsday?," *Clingendael Spectator*, January 6, 2021, https://spectator.clingendael.org/en/publication/will-humans-andor machines-save-us-nuclear-doomsday.

38 Edward Geist and Andrew J. Lohn, *How Might Artificial Intelligence Affect the Risk of Nuclear War?* (Santa Monica, California: RAND Corporation, 2018), 10, https://www.rand.org/pubs/perspectives/PE296.html.

39 *2019 Missile Defense Review* (Washington DC: Office of the Secretary of Defense, January 2019), 33, https:// media.defense.gov/2019/Jan/17/2002080666/-1/-1/1/2019-MISSILE-DEFENSE-REVIEW.pdf.

40 The term 'output' is contingent on the environment in which the DL technique is applied. One could imagine outcomes for DL application within EWS ranging from decision-making options, to specific vulnerability identification to increased image classification options.

speed of complex data processing, which could in turn provide more outputs for human decision-makers as well as the potential for increased decision-making timelines. Further, DL approaches, when applied to various pre-launch and boost phase environments, could decrease the risk of potential miscalculations or technical misperceptions, and, in doing so, could aid crisis diplomacy and the peaceful resolutions of various situations.[41]

Third, DL-enhanced early warning during the pre-launch phase widens the timeframe within which measures can be adopted to protect counterforce or counter-value targets, such as civilian populations and infrastructure. Given the time-sensitive nature of such measures (e.g., evacuation or missile interception), their effectiveness and efficiency could be significantly enhanced through DL-enabled anticipation of attacks.[42] DL-enhanced early warning during the boost phase could widen the timeframes for measures of active defense, such as missile defense response and interception, or timely attribution and reciprocal retaliation.

Risks for Nuclear Stability

Despite these opportunities, incorporating DL into EWS could have destabilising effects and exacerbate nuclear risks, stemming from 1) technical shortcomings, and 2) human-machine interactions.

First, the effectiveness of DL integration largely depends on both the quantity and quality of available data, in order to accurately make connections and identify patterns.[43] As operational datasets for events like nuclear attack are non-existent, the system would need to draw from artificially constructed datasets and testing scenarios.[44] As such, challenges could arise when trying to match the training data to the deployed environment and ensuring representativeness of the data.[45]

Further, DL systems are vulnerable to data poisoning, wherein biased data is fed unintentionally into the training model, which could result in the machine producing biased outputs when 'triggered' by the attacker — possibly pushing human operators towards a certain interpretation or decision. In addition, attacks utilizing adversarial examples[46] — "an instance with small, intentional feature perturbations that cause a machine learning model to make a false prediction",[47] can exploit a DL system by coaxing it toward a certain output via intentionally injecting it with imperceptibly corrupted input.[48] This could be particularly concerning in a nuclear EWS context as one could imagine the possibility of a compromised DL system prompting the boost phase due to data poisoning or an adversarial example attack. Because adversarial examples do not

41 Phil Stewart, "Deep in the Pentagon, a Secret AI Program to Find Hidden Nuclear Missiles," *Reuters*, June 5, 2018, https://www.reuters.com/article/us-usa-pentagon-missiles-ai-insight-idUSKCN1J114J.

42 Vincent Boulanin, ed., *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk* (Stockholm: SIPRI, 2019), 23, 28, 34, 41-49, 53-54, 65-66, 69-70, 80, 83, 86, 91, 136-137, https://www.sipri.org/publications/ 2019/ other-publications/impact-artificial-intelligence-strategic-stability-and-nuclear-risk-volume-i-euro-atlantic.

43 Eda Kavlakoglu, AI vs. Machine Learning vs. Deep Learning vs. Neural Networks: What's the Difference? *IBM Blog*, May 27, 2020, https://www.ibm.com/cloud/blog/ai-vs-machine-learning-vs-deep-learning-vs-neural-networks

44 Boulanin, *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, 19.

45 iPRAW, Focus on Computational Methods in the Context of LAWS, 2017, https://www.ipraw.org/publications/ computational-methods/, 11-12.

46 Louise Matsakis, "Artificial Intelligence May Not 'Hallucinate' After All," *wired*, August 5, 2019, https:// www.wired. com/story/adversarial-examples-ai-may-not-hallucinate/

47 Christoph Molnar, *Interpretable machine learning. A Guide for Making Black Box Models Explainable*, Chapter 6.2, last updated 28 July 2021,, https://christophm.github.io/interpretable-ml-book/adversarial.html.

48 Erin D. Dumbacher and Page O. Stoutland, "U.S. Nuclear Weapons Modernization: Security and Policy Implications of Integrating Digital Technology," (Washington DC: NTI, November 2020), 25, https://media.nti.org/ documents/ NTI_Modernization2020_FNL-web.pdf.

necessarily need access to the DL system's training data or parameters and are highly imperceptible, these attacks are incredibly challenging to defend against.[49] Adversarial attacks writ large are not necessarily always intentional, particularly when large amounts of data are not available — they can be a result of lack of a robust learning model due to underspecification.[50] Accidental robustness issues would be equally, if not more, concerning than intentional data poisoning or intentional adversarial attacks, especially in the context of nuclear EWS.

Further, the complex nature of such systems engenders the possibility of accidental use. Hidden interactions (such as feedback loops) may be unforeseen by human operators, whilst tightly coupled (i.e. interdependent) connections mean there is no "buffer" time between different internal EWS interactions.[51] As such, accidents can be inevitable or "normal" in complex technological systems.[52] Given the function of EWS, a single error or system failure could quickly trigger a devastating sequence of events toward accidental nuclear use.

Second, increased risk of accidental or inadvertent escalation can also stem from human behaviour, including the interaction between the system and a human operator. Specifically, complications could arise when attempting to understand why the system reacted (or failed to react) to certain real-life stimuli.[53] 'Un-interpretability' of DL systems may compromise trust in the system's reasoning — a serious problem given the dire consequences of false positives (false alarms) and false negatives (no alarms) in the nuclear context.[54] In an effort to counteract possible un-interpretability, predictability and understandability are necessary safety requirements to ensure the system is functioning correctly and thus must be worked into the testing, evaluation, validation and verification (TEVV) regime.[55] A lack of transparency over how and why certain correlations or conclusions are reached would compromise both assurances in the system, and its ability to trigger a fail-safe. Moreover, human "over-trust or uncritical trust" of machine capabilities (automation bias) may be heightened in crisis situations with the often-quoted requirement to "fight at machine speed."[56] Taken together, opacity of a system and automation bias could undermine effective human oversight which is of critical importance in the nuclear realm.

---

49 Paul Scharre, "Killer Apps: The Real Dangers of an AI Arms Race," *Foreign Affairs*, May/June 2021, https:// www.foreignaffairs.com/articles/2019-04-16/killer-apps.

50 Alexander D'Amour et al. "Underspecification Presents Challenges for Credibility in Modern Machine Learning" *arXiv* (November, 2020) https://arxiv.org/abs/2011.03395

51 UNIDIR, *Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies*, (Geneva: UNIDIR, 2016), 6, https://unidir.org/files/publications/pdfs/safety-unintentional-risk-and accidents-en-668.pdf.

52 Leveson, Nancy, Nicolas Dulac, Karen Marais, and John Carroll, "Moving Beyond Normal Accidents and High Reliability Organizations: A Systems Approach to Safety in Complex Systems," *Organization Studies* 30, no. 2–3 (2009), https://doi.org/10.1177/0170840608101478.

53 Michael C. Horowitz and Paul Scharre, *AI and International Stability: Risks and Confidence Building Measures* (Washington DC: CNAS, 2021), 7, https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/AI-and International-Stability-Risks-and-Confidence-Building-Measures.pdf.

54 Arthur Holland Michel, *The Black Box, Unlocked: Predictability and Understandability in Military AI* (Geneva: UNIDIR, 2020), https://unidir.org/publication/black-box-unlocked; Michael C. Horowitz, Paul Scharre and Alexander Velez-Green, "A Stable Nuclear Future? The Impact of Autonomous Systems and Artificial Intelligence," *arXiv* (December 2019), 4, https://arxiv.org/abs/1912.05291.

55 Holland Michel, *The Black Box, Unlocked*.

56 Margarita Konaev, Tina Huang and Husdanjot Chahal, "Trusted Partners: Human-Machine Teaming and the Future of Military AI," *CSET Issue Brief* (February 2021), 17, https://cset.georgetown.edu/wp-content/uploads/ CSET-Trusted-Partners.pdf; Horowitz, Scharre and Velez-Green, *A Stable Nuclear Future?*, 1.

Furthermore, significant enhancements of ISR capabilities kindle target states' perceptions of insecurity and vulnerability. By enabling the precise location, tracking, and targeting of second-strike capabilities, in particular mobile ICBMs, DL-enhancements in EWS could decrease deliberate ambiguity and undermine target states' first strike survivability, second strike capabilities, and thereby effective strategic deterrents.[57] This could have immediate effects on escalation spirals in crises. As likely consequences, target states could be incentivised to expand their nuclear forces and quickly advance their own AI enabled capabilities, potentially implementing immature technologies.[58]

In sum, the possible integration of DL in EWS could carry significant destabilising implications for crisis stability. In both cases, the mere perception of an adversary's capabilities often matters more than the actual capabilities themselves. To make things worse, false alarms—due to technical errors, inadequate data, or adversarial perturbations—paired with opaque algorithms and the human proclivity for automation bias, could cause crisis instability by setting off a proactive reaction that leads to accidental or inadvertent escalation.[59] The potentially devastating consequences would be particularly acute with nuclear weapons on high alert status.

Conclusion

It is imperative to create a context which could facilitate the potential integration of DL into aspects of EWS in a manner which enables optimal harnessing of the opportunities DL presents as well as mitigation of the risks it may create. Thus, states ought to be selective about when and where to integrate which technologies, must invest in technological literacy for decision-makers, allow for independent oversight, exchange best practices, be transparent about their approaches, and negotiate norms and confidence-building measures. The following section proposes measures that could contribute to the creation of such a context by fostering improved understanding and careful implementation of DL-enhanced EWS.

Recommendations

The possible adoption of DL into EWS could have mixed implications for nuclear risks. To some extent, these will be determined by the manner in which these AI learning techniques are developed and tested, and integrated into existing and new equipment, as well as the way this integration is accompanied by adequate training and doctrinal choices. The discussed implications from the possible integration of DL into EWS substantiate the call for open and transparent dialogue to contribute to the understanding of how technology could reduce and/or aggravate nuclear risks. Such dialogue needs to be comprehensive in terms of content and participation, covering all technologies with potentially disruptive effects to stability and engaging both governmental and non-governmental stakeholders, including industry and technical experts.

The following set of recommendations for diplomats and decision-makers outlines different pathways of how to multilaterally initiate such dialogue, what milestones are duly needed, and which activities on the national level are conducive to mitigating risks and increasing crisis stability.

Awareness Raising Measures

1. This paper sought to contribute to an open and public discussion of potentially disruptive implications from technological enhancements in the military domain. The complete range of stakeholders, including

57 Andrea Gilli and Mauro Gilli, "Rethinking the Impact of Emerging Technologies on Strategic Stability," in *Europe's Evolving Deterrence Discourse* ed. by Amelia Morgan and Anna Péczeli (Center for Global Security Research, Lawrence Livermore National Laboratory, 2021), 113, https://cgsr.llnl.gov/content/assets/docs/ CGSR_euro_det_final.pdf.

58 Horowitz and Scharre, *AI and International Stability*, 8.

59 Boulanin, *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, 92.

private and public actors, should engage in the discussion concerning the implications and interaction effects of emerging technologies for nuclear risks and strategic stability.

2. High-level intergovernmental initiatives, such as the P5 Process, the "Creating an Environment for Nuclear Disarmament" and the "Stepping Stones for Advancing Nuclear Disarmament" initiatives, should discuss the implications of emerging technologies on nuclear risks and strategic stability and report back to nuclear-specific multilateral fora, such as the upcoming Review Conference of the Treaty on the Non-Proliferation of Nuclear Weapons (NPT). These state groupings include the majority of relevant state stakeholders and provide the diplomatic platforms to informally exchange views and explore potential paths forward. Below-mentioned transparency and confidence-building measures suggest milestones for these intergovernmental exchanges.

3. In addition to above-mentioned intergovernmental initiatives reporting back to the NPT Review Conference or crafting agreed language to be included in final texts, topics concerning potentially disruptive technologies should be discussed in the Main Committee I and Subsidiary Body 1 to the NPT Review Conference.

4. In the absence of dedicated multilateral fora to systematically review developments in technology that could have radical and novel implications for strategic stability, intergovernmental initiatives should lay the groundwork for the establishment of a majority in the UN General Assembly First Committee to mandate a Group of Governmental Experts (GGE), involving the views of technical, industry and other non-governmental stakeholders, to periodically and systematically examine technological developments in relation to the nuclear domain and develop recommendations for the consideration of Member States.[60]

5. In line with its discussions and decisions in 2018, the Conference on Disarmament should re-establish the Subsidiary Body 5 and encourage discussions in line with its mandate.[61]

Transparency and Confidence-Building Measures

1. As the source of innovation, research and development does not lie primarily with state actors, it is essential not only to include, but to encourage private actors to participate in transnational discussions. Such discussion should also take place outside diplomatic fora, allowing for technical experts to interact and engage in transnational capacity building through sharing of best or worst practices on competence development and training, safety and security provisions, which could in turn contribute to confidence-building.[62] States parties should provide appropriate funding for such events.

60 The establishment of a Group of Scientific Experts (GSE) under the auspices of the Conference on Disarmament presents another or additional pathway to facilitating technical discussions, as the example of the GSE and its contributions to the Comprehensive Nuclear-Test-Ban Treaty shows, see Ola Dahlman, Frode Ringdal, Jenifer Mackby and Svein Mykkeltveit, "The inside story of the Group of Scientific Experts and its key role in developing the CTBT verification regime," *The Nonproliferation Review*, 2020.

61 Renewed discussions within the Subsidiary Body 5 will contribute to increase understanding and trust among delegations on the implications of particular developments in science and technology, see Conference on Disarmament's decision CD/2119, decision CD/2126, and report CD/2141.

62 The Wiesbaden Process that focused on global export controls and non-proliferation measures in view of UNSC Resolution 1540 provides a useful example of how to organize open dialogue between private and public stakeholders, see "The Spirit of Wiesbaden: Preventing the proliferation of weapons of mass destruction," German Federal Foreign Office, November 24, 2017, accessed on 31 March 2021, https://www.auswaertiges-amt.de/en/aussenpolitik/themen/aussenwirtschaft/-/692082.

2. Crisis communication channels and designated points of contact between states, such as those that participate in above-mentioned intergovernmental initiatives, are vital for strengthening strategic stability and mitigating nuclear risks by reducing the likelihood and consequences of misperception, enabling de-escalation and peaceful resolution of disputes through crisis diplomacy, providing opportunities for information exchanges between military and scientific communities, and enhancing trust.

3. The goal of increasing transparency is already on the agenda of the P5 process and should not only be reinvigorated, but also expanded to include discussions of technological enhancements and modernisation of EWS. Such discussions need to aim to foster mutual understanding, reduce the risk of misperception, and eliminate incentives for arms racing, as well as to define the threshold for nuclear escalation.

4. In line with existing proposals within the NPT Review Conference context to de-alert nuclear forces, this discussion of disruptive technologies only adds to the need to reduce dependence on high-alert forces.[32]

5. Additionally, intergovernmental initiatives should discuss the viability of missile launch notifications and the possibility of agreeing to such agreements in writing. States could build upon existing bilateral arrangements in order to establish a multilateral notification framework. This would generate an additional source of information that could be employed to scrutinise the performance of EWS, thereby reducing the risk of misperception.

State-Level Recommendations

1. Given the understandability and predictability problem of DL, it is difficult to anticipate how a DL system may behave in contexts and environments for which they have not been specifically tested. This calls for States that are integrating DL in EWS to set up a common task framework in order to minimise above mentioned risks stemming from technical shortcomings and human machine interactions by addressing the broadest possible conditions the system is likely to face if deployed.

2. Operational and technical training for all human operators that are likely to interact with the system through the development and deployment phase will contribute to enhancing the understanding of the DL systems' likely behaviours, notwithstanding their inexplicability problem, thereby reducing the above-mentioned risks stemming from technical shortcomings and human-machine interactions. Therefore, states whose militaries are likely to interact with DL integrated EWS should set up detailed and comprehensive operational and technical training regimes for their military personnel.

3. Even though it might be tempting for decision-makers to rush the integration of DL into EWS, decision-makers should be conservative when deciding when and where DL can be safely implemented. The self-learning capabilities of machines powered with this technique could create inadvertent escalation with potential devastating consequences with nuclear weapons on high alert status.

32 Two working papers to the 2019 Preparatory Commission for the 2020 Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons address nuclear de-alerting specifically, see NPT/CONF.2020/PC.III/WP.23 and NPT/CONF.2020/PC.III/WP.31.