Machine Learning and Nuclear Command:

How the technical flaws of automated systems and a changing human-machine relationship could impact the risk of inadvertent nuclear use

Peter Rautenbach

2022-11-09

Completed as part of the Cambridge Existential Risks Initiative Summer Research Fellowship

A special thank-you to my CERI mentor Haydn Belfield for his non-stop support and to Uliana Certan for her impeccable editing talent. I'd also like to thank those who donated their time to chat about the research including Michael Aird, Matthew Gentzel, Kayla Matteucci, Darius Meissner, Abi Olvera, and Christian Ruhl. And of course, one final thank-you to the entire CERI team and the other research fellows for all their support and rigorous debate.

1. Summary:

1.1 The Problem

The increasing autonomy of nuclear command and control systems stemming from their integration with artificial intelligence (AI) stands to have a strategic level of impact that could either increase nuclear stability or escalate the risk of nuclear use. Inherent technical flaws within current and near-future machine learning (ML) systems, combined with an evolving human-machine psychological relationship, work to increase nuclear risk by enabling poor judgment and could result in the use of nuclear weapons inadvertently or erroneously. A key takeaway from this report is that this problem does not have a solution; rather, it represents a shift in the paradigm behind nuclear decision making and it demands a change in our reasoning, behavior, and systems to ensure that we can reap the benefits of automation and machine learning without advancing nuclear instability.

The figure below demonstrates the general path from AI/ML systems integration with nuclear command and control toward an increased risk of nuclear weapons use.

In one sense, states desire this integration for the strategic advantages that it promises.[1] ML systems can function without rest, look at enormous amounts of data, and, perhaps most importantly, find patterns and connections in a way that often outperforms human analysts by also being able to draw conclusions from seemingly unrelated data points.[2] Furthermore, current NC3 systems are aging and the last major update was during the 1980s.[3] The need to ensure the technical effectiveness of NC3 is clear.

Beyond working for deterrence as intended , there is also interest in ensuring that systems are progressively safer. The automation of ML stands to reduce the number of near-calls related to human error, cognitive bias, and fatigue.[4] Despite concerns over integrating AI, it's also blatant that human operators are far from perfect and prone to allowing biases or making mistakes, especially when completing repetitive tasks or assessing adversarial moves.

---

[1]James Johnson, 'Delegating Strategic Decision-Making to Machines: Dr. Strangelove Redux?', *Journal of Strategic Studies* 45, no. 3 (16 April 2022): 463, https://doi.org/10.1080/01402390.2020.1759038.
[2]Ibid., 453.
[3] Jon Harper, 'Nuclear Command, Control, Comms Under Scrutiny', *Center For Strategic Deterrence Studies: News and Analysis*, no. 1357 (2019): 7
[4] Johnson, 'Delegating Strategic Decision-Making to Machines', 452 ; James Johnson, 'Rethinking Nuclear Deterrence in the Age of Artificial Intelligence'. Modern War Institute, 28 January 2021. https://mwi.usma.edu/rethinking-nuclear-deterrence-in-the-age-of-artificial-intelligence/.

Finally, states have an interest in keeping up with, if not surpassing, their adversaries' military technology.[5] This is particularly concerning regarding the implementation of ML in nuclear command because in "an effort to gain a real or perceived nuclear strategic advantage against their adversaries, while engaging in an AI race, states may place less value on AI safety concerns and more on technological development."[6] While the development of AI technologies by major states may not currently be best characterized as a 'race', the pressure to keep up is unmistakable: "AI has the potential to drastically change the face of war and the world at large… This in turn not only drives general integration, but raises the risk that both our advancements, and those of adversarial nations, increase the speed at which we field AI-enabled systems, even if testing safety measures are lacking."[7] Not only will modernized NC3 incorporate ML but there is a real risk of rushed integration with higher risk tolerance than normally accepted.

3.2 Modernization

Given these pressures to integrate ML with nuclear command, what concrete evidence is there that integration is contemplated within the U.S. defense establishment? Additionally, how might this integration specifically take place within early warning decisions support systems?

Despite the inherently classified nature of these developments, public statements by US leadership indicate that NC3 modernization could include further automation and AI integration.[8] When asked about AI and NC3 modernization, former USSTRATCOM Commander General Hyten stated: "I think AI can play an important part."[9] Former director of the USSTRATCOM NC3 Enterprise Center publication stated the desire and need for AI experts for NC3 modernization.[10] This follows statements by former U.S. Secretary of Defense James Mattis who also expressed an interest in the use of militarized AI and its ability to fundamentally change warfare.[11]

---

[5]Michael Horowitz and Paul Scharre, 'AI and International Stability: Risks and Confidence-Building Measures', Technology & National Security (Center for a New American Security, 2021), 5, https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.

[6]Maas, Matteucci, and Cooke, 2022, 23

[7] Michèle A Flournoy, Avril Haines, and Gabrielle Chefitz, 'Building Trust through Testing: Adapting DOD's Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, Including Deep Learning Systems' (WestExec Advisors, 2020), 5.

[8] Stoutland, 2019, 63; Boulanin et al, 2020, 22

[9] Philip Reiner and Alexa Wehsener, 'The Real Value of Artificial Intelligence in Nuclear Command and Control', War on the Rocks, 4 November 2019, https://warontherocks.com/2019/11/the-real-value-of-artificial-intelligence-in-nuclear-command-and-control/.

[10]Ibid.

[11]Yuna Huh Wong et al., 'Deterrence in the Age of Thinking Machines' (RAND Corporation, 27 January 2020), 4, https://www.rand.org/pubs/research_reports/RR2797.html.

Building off this, "The U.S. budget for Fiscal Year 2020, for instance, singled out AI as a research and development priority and proposed $850 million of funding for the American AI Initiative."[12] China has "declared its intentions to lead the world in AI by 2030, estimated to exceed tens of billions of dollars."[13] Modernization and integration of AI with NC3 will likely take many different forms over the next decade with over $70 billion going towards command and control and early warning systems as part of the NC3 modernization.[14]

Given these statements and the potential benefits of AI, it is reasonable to assume that efforts to incorporate further AI and automation into the military include NC3. At the very least, it should be assumed that this prospect is being explored, and if integration doesn't happen now, it could easily happen in future years.

Considering these varied indications that integrating AI into military systems, including NC3, is apparently beneficial and actively explored by the US military, it's valuable to discuss the two important overarching ways in which integration could impact risk: firstly, through early warning systems, and secondly through a form of 'predictive forecasting' (similar to a 'cognitive maneuver').

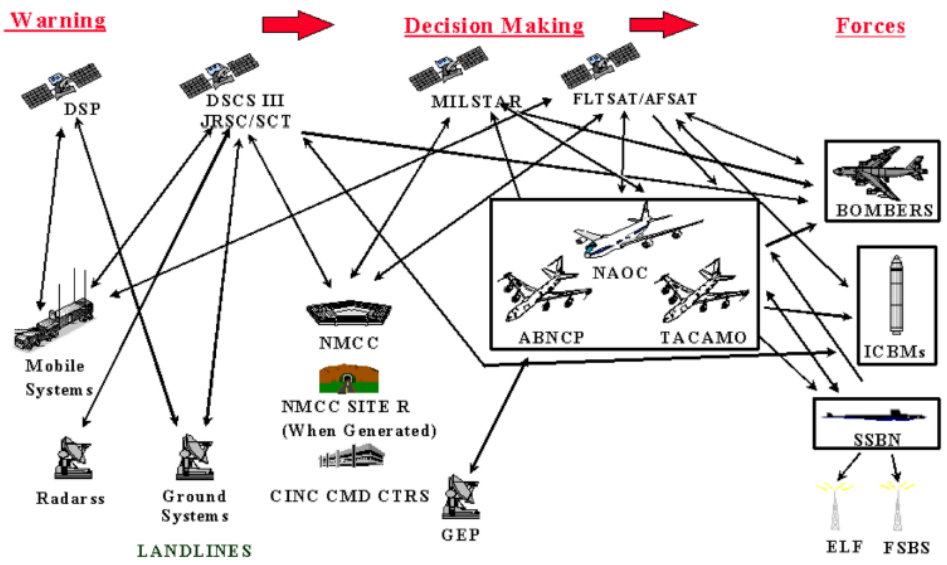
3.3 Early Warning:

The early warning system is a core part of NC3 where integration of AI systems will be particularly incentivised and impactful.The US early warning system uses a combination of pace-based infrared (IR) sensors and ground-based radars to detect potential incoming ballistic missiles.[15] This then triggers an alert at the North American Aerospace Defense (NORAD) Command in Colorado where analysts work to confirm and authenticate the warning before quickly submitting the information to leadership if it is deemed reliable.[16]

---

[12] Michael C. Horowitz et al., 'Policy Roundtable: Artificial Intelligence and International Security', Texas National Security Review, 2020, https://tnsr.org/roundtable/policy-roundtable-artificial-intelligence-and-international-security/.
[13] Ibid.
[14] Arms Control Association, 'U.S. Nuclear Modernization Programs, 'https://www.armscontrol.org/factsheets/USNuclearModernization
[15] Hruby and Miller, 2019, 15
[16] Ibid.

Figure 3: Notional Connectivity for Nuclear Command and Control in the United States



Source: "National Command & Control: The National Military Command System (NMCS)," October 2001.

This entire process takes place within a few minutes- an already short amount of time to analyze a high-stakes situation and avoid errors. And yet, detection and analysis of incoming attacks is getting increasingly complicated. Early warning systems must be capable of detecting multiple targets and also discriminating between: type of attack, launch and impact points, validity, and more.[17] AI could remove the worst aspects of humanity, while providing analysis for mass amounts of data, in a position that is traditionally both mentally taxing and boring for human analysts. This is arguably an important driver behind AI integration as both bias and exhaustion are human limitations that can easily impede good decision making.

ML systems could effectively replace at least some of the analytical work done by humans who assess early warning information to determine the credibility of a threat. "Recent publications have highlighted the potential for machine learning-based algorithms to provide better discrimination abilities in radar applications. If used in early-warning systems, this could in principle result in fewer false alarms."[18] Early warning information is getting progressively complicated with the timeframe to determine the validity of the danger getting smaller. New weapons technologies, such as hypersonic delivery systems, can complicate traditional detection by appearing later or confounding detection systems.[19]

Additionally, the advent of technologies such as hypersonic weapons or even the automation of attacks means that combat could soon reach speeds much too fast for human

[17] Hruby and Miller, 2019, 15
[18] Stoutland 2019, 65.
[19] Ibid., 16.

cognition. AI-augmented systems would be essential for any offensive or defensive operations occurring at 'machine-speeds' including cyber warfare or automated weapon systems. [20]

Given the importance of early warning systems within NC3, the possibilities for their modernization, and the significant reasons to do so, it is feasible and even probable that ML will be integrated with early warning systems.

3.4 Predictive Forecasting of the Imminent Use of Nuclear Weapons

Once the realm of science fiction, there are now discussions around the use of machine learning in command and control to detect nuclear threats *before* they occur. This could involve ML systems analyzing relevant factors such as troop movements, supply lines, communication, and other intelligence to calculate where nuclear threats, not only could, but likely will, come from.

A prime, if somewhat rudimentary, example of this would be the Soviet Union's response to the events leading up to Able Archer 83.[21] In response to fears that the U.S. would first-strike the Soviet Union, "some 300 operatives [were tasked] with examining 292 different indicators—everything from the location of nuclear warheads to efforts to move American 'founding documents' from display at the National Archives."[22] This information "was then fed into a primitive computer system, which attempted to calculate whether the Soviets should go to war to pre-empt a Western first strike."[23] An AI could theoretically be tasked with forecasting future attacks in a similar, though more advanced, manner.

A relatively modern example of this kind of technology is the Defense Advanced Research Projects Agency's (DARPA) Real-time Adversarial Intelligence and Decision Making (RAID) machine learning algorithm "designed to predict the goals, movements, and even the possible emotions of an adversary's forces five hours into the future."[24] Acting as support tools for decision makers, "future iterations of these systems may be able to identify risks (including risks unforeseen by humans), predict when and where a conflict will break-out, and offer strategic solutions and alternatives, and, ultimately, map out an entire campaign."[25]

---

[20]Johnson, 'Delegating Strategic Decision-Making to Machines', 441.
[21]Able Archer 83 was the annual NATO Able Archer exercise conducted in November 1983. The purpose was to simulate a period of escalation that ended in the US military reaching DEFCON 1 and simulated a coordinated nuclear attack. The realism of the exercise, combined with new lows in U.S-Soviet Relations, led Soviet leadership to believe that Able Archer was a ruse and a genuine preparation for a first-strike on the Soviet Union. This perceived aggression led to the Soviet Union preparing to use their nuclear forces and the U.S. had little to no idea that this was occuring. While this ended without incident, this is arguably one of the closet times we have come to nuclear use.
[22] Nate Jones and Peter J. Scoblic, 'The Week the World Almost Ended', *Slate*, 7 June 2017, https://slate.com/news-and-politics/2017/06/able-archer-almost-started-a-nuclear-war-with-russia-in-1983.html.
[23]Ibid.
[24] Johnson, 'Delegating Strategic Decision-Making to Machines', 453
[25]Ibid.

This is not an attempt to see the future but rather a kind of 'predictive analytics' already used to combat crime in cities across the globe.[26] This kind of technology works to determine where crime will occur before it actually does. An AI system tasked with this responsibility conducts analysis, finds correlations, and then draws conclusions from the data and makes complex statistical predictions about future behavior, providing decision support and suggestions to the experts in the field.[27] In a basic sense, this isn't too different from modern image classifiers or language models in which systems learn from patterns (guided by human-attributed labels and rewards). The predictive forecasting machine in NC3 would also seek patterns that indicate a potential incoming threat by continuously monitoring an immense amount of real world data and calculating their significance to nuclear threats..

The strategic value of an accurate predictive and preemptive method in nuclear security is invaluable. Rather than reacting and responding to a nuclear strike or a similarly threatening attack, a military power would anticipate their adversaries' moves and either thoroughly prepare for them, hinder or inhibit them, or take the offensive. AI systems could theoretically offer the safest way to achieve preemption (an already contentious concept) if they can outperform their human counterparts, remain unbiased, and provide a calculated warning or suggestion. Additionally, AI can assess data and bring together different pieces of intelligence in a manner that a human potentially never would.[28]

Of course, even if such systems are deployed to assist with strategic decision-making, the question remains whether humans would act solely, or at least primarily, based on their recommendations. At a series of workshops on artificial intelligence and nuclear risk held by the Stockholm International Peace Research Institute (SIPRI) in 2020, "workshop participants found it hard to believe that a nuclear-armed state would find such a system reliable enough to initiate a pre-emptive nuclear attack based only on the information that its algorithms produce."[29] Participants believed that states would likely wait for tangible evidence such as early warning system detections to confirm the AI conclusions .[30]

I personally question this confidence in the reliability of decision-makers and their ability to wait to confirm the results of this predictive forecasting. First, crisis situations are unpredictable, and emotions running high or the other systems failing could result in less patience than imagined. Second, the human-machine relationship could evolve to a point where trust in machine intelligence is far higher, almost implicit. Finally, waiting for systems such as those used in early warning to confirm the statistical prediction of AI would completely negate the reason for deploying it in the first place.

---

[26] Keith Dear, 'Artificial Intelligence and Decision-Making', *The RUSI Journal* 164, no. 5–6 (19 September 2019): 20, https://doi.org/10.1080/03071847.2019.1693801.
[27] Boulanin et al., 2020, 121
[28] Johnson, 'Delegating Strategic Decision-Making to Machines', 453.
[29] Boulanin et al., 2020, 121
[30] Ibid.

ML powered predictive analytics in nuclear command and control could potentially be, or already is being, pursued by militaries. Even if the SIPRI participants are correct that decision-makers will reluctantly follow the ML recommendations, there are nevertheless a myriad of risks that must still be addressed following this integration.

4. Inherent Dangers of Integration

This section outlines the risks of integrating ML with nuclear command. It offers an explanation of the contributing factors that result in these risks and also an explanation of the mechanisms - the how-and-what could go wrong and result in actual nuclear use or a significant increase in nuclear use risk. Each issue here presents a problem on its own, while also often working in conjunction with the other outlined problems to magnify the increased chance of inadvertent nuclear use.

In section 4.1 I outline the technical flaws within AI systems and how they interact with nuclear deterrence and decision making. In section 4.2 I explore how increasing automation could change the human-machine relationship, which stands to alter the decisions-making process of nuclear deterrence. As previously established, *any* potential changes to decision-making around nuclear weapons demand the utmost scrutiny.

4.1 Technical Flaws in AI Technology

The problems outlined here are technological hurdles that face current and near-term AI. Some of these seem solvable while others may simply be inherent and unsolvable. Regardless of these hurdles, the current and coming modernization means that integration may occur while these problems still persist. Additionally, there is the simple adage that ‘complexity breeds accidents’. Even with rigorous testing, the coding error rate is often between 0.1 to 0.5 errors per 1000 lines of code.[31] Given that luxury automobiles have around 100 million lines of code,[32] one can imagine that the amount of code required in AI systems used for command and control would be immense. Errors would be inevitable.

The nuclear weapon context also has specific implications for looking at technical flaws in AI. It is the paramount example of a ‘safety critical’ environment. The consequences of failure are at their highest which demands the utmost scrutiny into how decisions are made and what tools are used to aid the process. Additionally, the flaws outlined below would likely become a serious issue in a time of crisis. Periods of inflamed tensions and dangerous rhetoric are when mistakes are most likely. Finally, in critical circumstances, humans would likely have a very limited amount of time, if any, to scrutinize the data or suggestions provided by an AI.

[31] Scharre 2019, 157
[32] Ibid.

4.1.1 The Alignment Problem

AI systems alignment, or misalignment, is a key problem facing AI deployment. While typically considered in the context of artificial *general* intelligence (AGI) development, ML experts struggle to design even modern ML systems that act exactly as intended without behaving even slightly wrong in unexpected and surprising ways.

Recent work done by Anthropic built on this claim when they found that "large generative models have an unusual combination of high predictability - model capabilities scale in relation to resources expended on training - and high unpredictability — specific model capabilities, inputs, and outputs can't be predicted ahead of time."[33] It is important to recognize that generally speaking, the impact of technical flaws do not cause the machine to break or fail to work. Rather, the machine does exactly as instructed, but not what is wanted from its programmers. Brian Christian effectively illustrated the "alignment problem," as by describing those who employ AI systems in "the position of the 'sorcerer's apprentice': we conjure a force, autonomous but totally compliant, give it a set of instructions, then scramble like mad to stop it once we realize our instructions are imprecise or incomplete."[34]

The following problems fall under the greater issue of misalignment or contribute to it. . These specific technical flaws can result in the deployment of ultimately misaligned systems..

4.1.2 The 'Black Box' problem

This is perhaps the most central problem in the context of nuclear command integration as it creates its own issues while also magnifying the other technical problems. 'Black Box' refers to ML systems being "opaque in their functioning, which makes them potentially unpredictable and vulnerable."[35] These systems are somewhat unknowable as "neural nets essentially program themselves… they often learn enigmatic rules that no human can fully understand."[36] We can test their outputs, but we don't really know why or how they reach their conclusions. For all we know, the system is misaligned and either using data incorrectly or measuring the wrong data altogether. Since we don't understand the system's inner workings, its behavior can seem odd and completely unexpected, almost alien. Reliability is key in terms of nuclear decision making.

One example of this problem's consequences is discussed in a 2016 article from OpenAI describing their efforts to conduct new reinforcement learning (RL) experiments using the game CoastRunner. The team trained the AI to obtain the highest possible score in each level, assuming that this concrete goal would reflect their informal goal for the AI to finish the race. However, the RL agent determined that it could more effectively achieve a high score by simply

---

[33]Deep Ganguli et al., 'Predictability and Surprise in Large Generative Models' (arXiv, 15 February 2022), 2, http://arxiv.org/abs/2202.07785.
[34] Brian Christian, The Alignment Problem: Machine Learning and Human Values (New York: W.W. Norton, 2020), pp. 12-13.
[35] Boulanin et al., 2020, 128
[36]Ariel Bleicher, 'Demystifying the Black Box That Is AI', Scientific American, 2017, https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/.

going around in circles and continually knocking over the same three targets - as shown in the CoastRunners 7 video.[37] Despite catching on fire, crashing into a boat, and going the wrong direction, the AI scored 20 percent higher than human players. This perfectly illustrates the Black Box problem: we cannot fully predict its behavior, and we cannot foresee or control its interpretation of our goals. While this is harmless in a video game, this could be catastrophic in a nuclear context.

The 'Black Box' issue has been recognized by the U.S. Department of Defense (DoD) as the "dark secret heart of AI" which has acted as a significant hurdle to the military use of ML systems.[38] This innate opacity is particularly problematic in safety-critical circumstances with short timeframes such as a nuclear crisis. Crises by their very nature are unpredictable: "testing is vital to building confidence in how autonomous systems will behave in real world environments, but no amount of testing can entirely eliminate the potential for unanticipated behaviors."[39] It is important to note that "there is active research, often called 'explainable AI,' or 'interpretability' to better understand the underlying logic of ML systems."[40] However, at this time the 'Black Box Problem' remains and may be inherent to ML systems and thus not 'solvable'.

Most importantly, the 'black box' makes it significantly more difficult to address the other technical flaws of AI systems. The need to determine and address AI system 'flaws' seems antithetical to the unknowable nature of these issues. Technical issues can very well remain dormant until they are triggered when the system is already in operation.

4.1.3 Brittleness

In the face of complex operating environments, machine learning systems often encounter their own brittleness – the tendency for powerfully intelligent programs to be brought low by slight tweaks or deviations in their data input that they have not been trained to understand.[41] On the technical side, this is known as a 'distributional shift' where ML systems can "make bad decisions – particularly silent and unpredictable bad decisions – when their inputs are very different from the inputs used during training."[42]

In one case, despite a high success rate in the lab, graduate students in California found that the AI system they had trained to consistently beat Atari video games fell apart when they

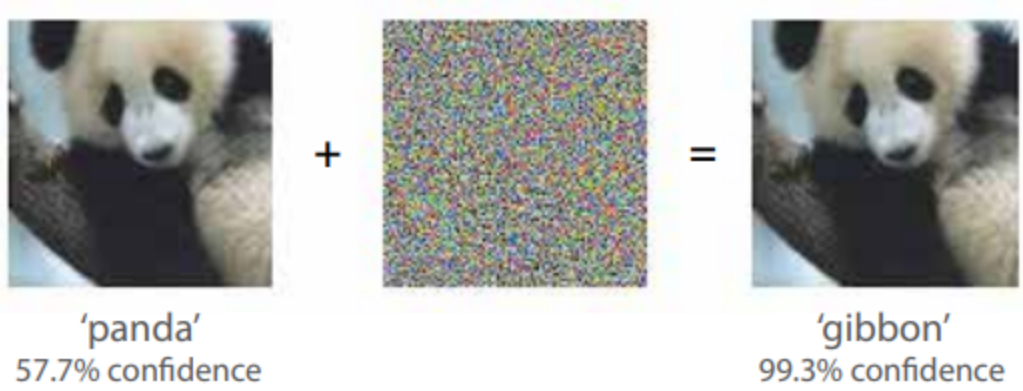
---

[37]Jack Dario, 'Faulty Reward Functions in the Wild', OpenAI, 22 December 2016, https://openai.com/blog/faulty-reward-functions/.
[38]Shin-Shin Hua, 'Machine Learning Weapons and International Humanitarian Law: Rethinking Meaningful Human Control', *Georgetown Journal of International Law* 51, no. 1 (2020 2019): 135.
[39] Scharre 2019, 32
[40] Hruby and Miller, 2019, 7
[41] Ibid.
[42]Dario Amodei et al., 'Concrete Problems in AI Safety' (arXiv, 25 July 2016), 3, https://doi.org/10.48550/arXiv.1606.06565.

added just one or two random pixels to the screen.[43] In another case, trainers were able to throw off some of the best AI image classifiers by simply rotating the objects in an image.[44] This is an unacceptable issue when considering AI employment in an ever-changing real-world environment. Unlike human operators, autonomous ML systems " lack the ability to step outside their instructions and employ 'common sense' to adapt to the situation at hand".[45]

The figure below illustrates how easily an adversary could confuse an image classifier by simply adding pixels(invisible to humans) to a picture,drastically changing the AI's interpretation while it signals evidentiary confidence.

Deception of an artificial intelligence by an adversary



Source: JCN 1/18, Human-Machine Teaming, Page 30

While this report does not focus on intentional attempts to deceive ML systems, the ease of this deception effectively demonstrates the concept and gravity of brittleness.

Indeed, ML systems struggling with brittleness are unprepared and unsuitable to function outside of the lab, let alone in a military setting. War is an atypical situation – while we can spend a prodigious amount of time training and preparing for it, the 'fog of war' will always lead to unexpected developments and surprises. An ML system would likely have a hard time adapting to new situations, and even if it was possible, taking the time to "adapt" could result in serious costs.[46] They would be held to extreme standards, as the need for accuracy and safety in the military context is unmatched. This is far more relevant in safety-critical nuclear security. We can imagine the impermissible consequences of a brittle AI system either failing to provide accurate information in a nuclear security context, or worse, providing inaccurate information with high confidence, leading to unwarranted decisions and actions by trusting human operators.

---

[43]Douglas Heaven, 'Why Deep-Learning AIs Are so Easy to Fool', *Nature* 574, no. 7777 (9 October 2019): 165, https://doi.org/10.1038/d41586-019-03013-5.
[44]Ibid.
[45]Scharre 2019, 146
[46] Horowitz et al., 'Policy Roundtable'.

This vulnerability is compounded by inadequate data sets, resulting in even more unreliable systems.[47] This is especially true in the nuclear security sphere, where so much of the training data is simulated[48] Although there are extensive records associated with the launch of older ballistic missiles, newer, less tested models require the use of simulation.[49] The lack of data in terms of real-world offensive nuclear use is undoubtedly fortunate for the world, but it nevertheless means that much of the data involved in training machine learning programs for NC3 systems will be artificially simulated. Even real-world data is often insufficient to train ML systems for real-world environments, so training on simulated data leaves systems woefully unprepared for deployment in military operating environments. Despite the best efforts to ensure that simulated data is accurate and robust it will always be an imperfect, short-sighted imitation of world events.Therefore, when a ML system encounters even slightly unprecedented, real-world data, it it may be enough to throw it off track, with or without anyone noticing. Watching AlphaGoZero play the game of Go has been described as ”watching an alien, a superior being, a creature from the future, or a god play.”[50] The program has the same goal as human players, but the way it achieves them - the actions it takes to get there - can be “almost impossible to comprehend.”[51] It may seem obvious that an AI system is not human, but its inherently inhuman nature matters more than it may first appear, and should never be forgotten, underestimated, or overlooked. This is linked to the ‘Black Box’ problem outlined earlier. We don’t truly know how or why an AI does what it does. We can see its output, but its reasons or justifications remain a mystery and should never be entirely trusted.

This issue of trusting the output of ML systems is further by the fact that these systems can be confident in their output even if they are wrong. Give a dog classifier a picture of a cat and it will tell you it's a Corgi with 99% confidence. Not only can systems be wrong but they can be wrong while telling you they are right with a high degree of confidence. This largely stems from the problem of brittleness as encountering unrecognized data often results in misinterpretation by the ML system as it attempts to categorize the data within the boundaries it does understand. Therefore, not only is it hard to identify when ML systems are not working as expected, but their degree of confidence may engender unwarranted trust in the machine.

AI systems in early warning or predictive forecasting would face immense, complex tasks. There is a very real chance that these machines would be brittle enough to fail at its desired performance outside the laboratory setting.

---

[47] Horowitz and Scharre, ‘AI and International Stability’, 7.
[48] Boulanin et al, 2020, 121; Hruby and Miller, 2019, 7
[49] Hruby and Miller, 2019, 7
[50] Dear, ‘Artificial Intelligence and Decision-Making’, 23.
[51] Ibid.

4.1.4 Human Bias in the Machine

One of the main arguments supporting the potential integration of ML systems with nuclear command and control states that decision-making processes stand to benefit from eliminating human error and bias. However, ironically, AI systems themselves are built and trained with human bias present, which clouds their own decision-making abilities.[52]

The code and algorithms that form the foundation of 'objective' and autonomous ML systems are written by humans - coders, developers, programmers, engineers - and it is also humans who train the program to run according to their interests and expectations. However well meaning, these humans unintentionally integrate human bias into their work; it is inevitable. Moreover, AI becomes further biased when it is trained using historical data, which is a product of its circumstances. For example: "machine learning algorithms designed to aid in criminal risk assessments… [are learning] racial bias from historical data, which reflects racial biases in the American criminal justice system."[53] This conclusion was reinforced when a study "of a commonly used tool to identify criminal recidivism found that the algorithm was 45 percent more likely to give higher risk scores to black than to white defendants".[54]

Another example of human bias occurred when Amazon used an AI system to filter résumés of potential job candidates. Evidently, the system was found to have a significant bias toward male applicants, presumably because it trained by observing desirable patterns in résumés of previously successful employees, most of which were male.[55] As a result, the AI determined that gender was a significant factor of candidate success and learned to place a higher value on male candidates. This not only reflected the male dominated nature of the industry at the time, but also perpetuated it.[56]

These examples illustrate that human bias can become integrated into AI systems by humans and/or training data, and then it is further ingrained and perpetuated by the AI system until it is detected. So, even in a best-case scenario where a ML system accurately interprets quality, relevant data and overcomes brittleness or other technical issues, it can still over- or under-evaluate the information it is assessing due to its trained human biases.

One can easily imagine similar problems occurring in AI systems trained to observe the actions of adversarial nations for the purpose of early warning or predictive forecasting. Human bias could sneak into the system, and the data relating to certain actors or variables could cease to accurately reflect what is occurring in the real world. In that case, extra weight could be

---

[52]'Human-Machine Teaming (JCN 1/18)', Joint Concept Note, 2018, https://www.gov.uk/government/publications/human-machine-teaming-jcn-118.
[53] Horowitz et al., 'Policy Roundtable'.
[54]Ibid.
[55]Jeffrey Dastin, 'Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women', *Reuters*, 10 October 2018, sec. Retail, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.
[56] Ibid.

given to standard activity, making a relatively normal action appear threatening. This could result in an AI warning of a potential incoming nuclear attack. A U.S. president and military commanders would be expecting a balanced, fair calculus when in truth the machine is perpetuating the kind of bias that could lead to catastrophic miscalculation and further the risk of responding to false positives.

These concerns become all the more pertinent during crisis situations such as the current ongoing conflict in Ukraine. A hypothetical ML integration system designed to assess the probability of Russian aggression against the West could act in unpredictable ways and give faulty information. The previously mentioned issues surrounding brittleness and the 'Black Box' could lead the system to misinterpret signating information such as troop movement, launch indicators, chatter, etc. A trained bias could then compound this misinterpretation, over-valuing the possibility of Russian aggression. For example, the system could perceive certain rhetoric or troop deployment as overly significant indicators of an imminent attack. While we hope that even in these cases military planners would review the information and hesitate to act on it, I don't believe one should so easily discount the power of confirmation bias or the chaos and 'fog of war' in a crisis scenario. Seemingly accurate information provided during an emergency could easily be acted on.

4.2 Problems with Human-Machine Interaction

This section explores the impact more automation could have on the human element of the machine-human team. This involves both how we treat the AI and also what the use of such intelligent machines means for our own thought processes.

Understanding why something has *not* happened is an inherently arduous task. Determining why nuclear weapons have not been used, whether it is due to more normative factors like the nuclear taboo or factors linked to structural realism such as deterrence, is difficult. In all likelihood a complex combination of factors have contributed to the non-use of nuclear weapons since WW2. Nonetheless, I would argue a key factor has been human *uncertainty* and our *lack* of knowledge. How we make decisions matters.

Human limitations and emotions, despite all their dangers, seem to have played a key role in preventing nuclear weapon use. The consequences of potential mistakes are so great at the nuclear level that individuals facing decisions at key moments often chose to risk their own lives and their teams rather than deploy nuclear weapons because they were not convinced of what appeared to be an incoming attack.[57] This is not to downplay their courage or training, but when facing seemingly reliable information of a nuclear attack, these individuals decided not to

[57]Nicola Davis, 'Soviet Submarine Officer Who Averted Nuclear War Honoured with Prize', *The Guardian*, 27 October 2017, sec. Science, https://www.theguardian.com/science/2017/oct/27/vasili-arkhipov-soviet-submarine-captain-who-averted-nuclear-war-awarded-future-of-life-prize. ; Pavel Aksenov, 'Stanislav Petrov: The Man Who May Have Saved the World', *BBC News*, 26 September 2013, sec. Europe, https://www.bbc.com/news/world-europe-24280831.

act when others may have, and in some cases almost did. Whether it was intuition, reasonging, fear, guilt, panic, or a combination of these and other very human reactions, their decision to question the situation and hold back on reacting may have saved millions, maybe billions of lives.

Between 1945-2017 there have been “37 different known episodes [linked to close inadvertent use], including 25 alleged nuclear crises and twelve technical incidents.”[58] On the surface, it appears that the systems designed to prevent nuclear use worked. It has even been argued that “those in charge of nuclear weapons have been responsible, prudent, and careful… [and] ‘close calls’ have ranged in fact from ‘not-so-close’ to ‘very distant.’”[59] While I would be far more hesitant to put such strong faith in the systems designed to prevent inadvertent use, the fact remains that we haven’t used the weapons since the end of WW2. Luck may play a role here, but too many years with too many ‘close calls’ have occurred for us to completely discount the nuclear decision making systems. As the current decision-makers, we must be doing something right. One part of this success preventing inadvertent use is explicit mistrust of computer warning systems. In both the Petrov and NORAD cases, computers demonstrated high confidence of incoming attacks, and human operators didn’t trust these false alarms. Their doubt and skepticism was paramount. Regardless of one’s stance in this discussion, any evolving technology that could even potentially change how nuclear decisions are made demands thorough scrutiny as there are a number of unfortunate ways the situation could become more dangerous as a result.

While the previous section looked at technical flaws within ML systems and how they could impact the quality and reliability of their outputs, this section looks at how increased automation impacts the human element. This includes issues such as an over reliance on ML systems, the increasing speed of warfare and AI systems, and the growth of misplaced confidence in military commanders when backed by powerful machine intelligences. Each issue will be outlined and their possible impact on nuclear decision making explored. The eventual conclusion is that integrating ML with NC3 could result in a paradigm shift in nuclear decision-making.

4.2.1 Automation Bias

One key issue is the development of automation bias – the “phenomenon whereby humans over-rely on a system and assume that the information provided by the system is correct.”[60] Bias also exists in the other direction with anover-mistrust of machines known as the “trust gap”.[61]

---

[58] Bruno Tertrais, "On The Brink—Really? Revisiting Nuclear Close Calls Since 1945," *The Washington Quarterly* 40, no. 2 (2017): 51, https://doi.org/10.1080/0163660X.2017.1328922
[59] Ibid.
[60] Boulanin et al, 2020: 114
[61] Michael C. Horowitz, ‘Trust, Confidence, and Organizational Decisions about AI Adoption: The Impact for US Defen’, Minerva Research Initiative, 2020. ; Hruby and Miller, ‘Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapon Systems’, 8.

Michael Horowitiz suggests there are three stages of trust towards artificial intelligence technologies:

1. ‘Technology Hype’, or “inflated expectations of how a given technology will change the world”
2. The ‘Trust Gap’, “or the inability to trust machines to do the work of people, in addition to the unwillingness to deploy or properly use these systems”[62]
3. ‘Overconfidence’ or automation bias.

The claim that humans progressively move through these stages was supported when “psychologists… demonstrated that humans are slow to trust the information derived from algorithms (e.g., radar data and facial recognition software), but as the reliability of the information improves so the propensity to trust machines increases – even in cases where evidence emerges that suggests a machine’s judgment is incorrect.”[63]

Automation bias has “been recorded in a variety of areas, including medical decision-support systems, flight simulators, air traffic control, and even “making friendly-enemy engagement decisions” in shooting-related tasks.”[64] Given the multitude of technical flaws outlined earlier, one can imagine the problem with an overreliance on ML systems. And yet, the ‘trust gap’ is also problematic and should not be adopted as a desirable stance. Instead, there should be strong cognizance of the human tendency to trust the machine to the point of assuming it is always correct or more capable than its human counterpart.

In the context of inadvertent nuclear use, automation bias is a substantial problem when considering integrating ML with NC3 because it exacerbates both the time crunch of a crisis scenario and the ‘black box’ problem. Human operators would be unable to check the math behind the AI’s decision, nor would they have time to even try. Additionally, “operators might therefore be more likely to over-trust the system and not see a need to verify the information that it provides.”[65] Not only does this increase the likelihood that the aforementioned technical flaws of ML systems will go unnoticed, it also represents an effective pre-delegation of authority to these machine intelligences. Automation bias, and the “unwarranted confidence in and reliance on machines… in the pre-delegation of the use of force during a crisis or conflict, let alone during nuclear brinkmanship, might inadvertently compromise states’ ability to control escalation.”[66]

The pre-delegation of authority and automation bias are most evident in the Patriot fratricides (or friendly fire incidents) during the 2003 Iraq War. Three out of twelve successful engagements involved fratricides. This included “two incidents in which Patriots shot down

---

[62]Horowitz, ‘Trust, Confidence, and Organizational Decisions about AI Adoption’.
[63]Johnson, ‘Delegating Strategic Decision-Making to Machines’, 444.
[64]Ruhl, ‘Autonomous Weapon Systems & Military AI: Cause Area Investigation’, 17.
[65] Boulanin et al, 2020: 115
[66]Johnson, ‘Rethinking Nuclear Deterrence in the Age of Artificial Intelligence’.

friendly aircraft, killing the pilots, and a third incident in which an F-16 fired on a Patriot."[67] In the cases where the Patriots fired and caused the fratricides the AI guided machines were wrong, and in trusting them their operators succumbed to automation bias. In both cases, the errors in the Patriot systems harken back to the brittleness technical issue. In the first, the system misidentified a friendly fighter as a missile[68], and the second instance involved the Patriot systems tracking an incoming "ghost" missile that wasn't there. Unfortunately, a nearby friendly fighter was in the wrong place at the wrong time and the Patriot's seeker locked on and killed the pilots.[69]

In the end, the AI driven systems were wrong and people were killed, illustrating that grave "problems can arise when human users don't anticipate these moments of brittleness".[70] One aspect of this is ensuring that automation does not result in the undue pre-delegation of authority to a ML enhanced system. Ensuring that the faults and strengths of these machines are understood is critical for their safe use.

4.2.2 Machine Speed

As previously stated, a key potential benefit to increasing the automation of military systems is that they can perform their assigned functions at blinding speeds that drastically outpace human operators. The risk here is that this increased processing speed could push "the speed of war to 'machine-speed' because autonomous systems can process information and make decisions more quickly than humans."[71] Aggressive reactions and decisions could take place in nano-seconds - dubbed 'hyperwar' in the West and 'battlefield singularity' in China.

Thomas Schelling's concern that "the premium on haste" is "the greatest source of danger that peace will explode into all out war" is echoed here when discussing machine speed and nuclear decision making. There is a concern that, in an effort to maintain battlefield advantage, states will pursue machine speed in their military operations and risk losing control of their machines, effectively pre-delegating authority to their AI systems to act in their stead. As of now, these systems are brittle and flawed, lacking the ability to think critically or work outside of the box. They are "set up to rapidly act on advantages they see developing on the battlefield" and could easily "miss de-escalatory signals,"[72] echoing R.K. Bett's conclusion that states often "stumble into [war] out of misperception, miscalculation and fear of losing if they fail to strike first."[73]

---

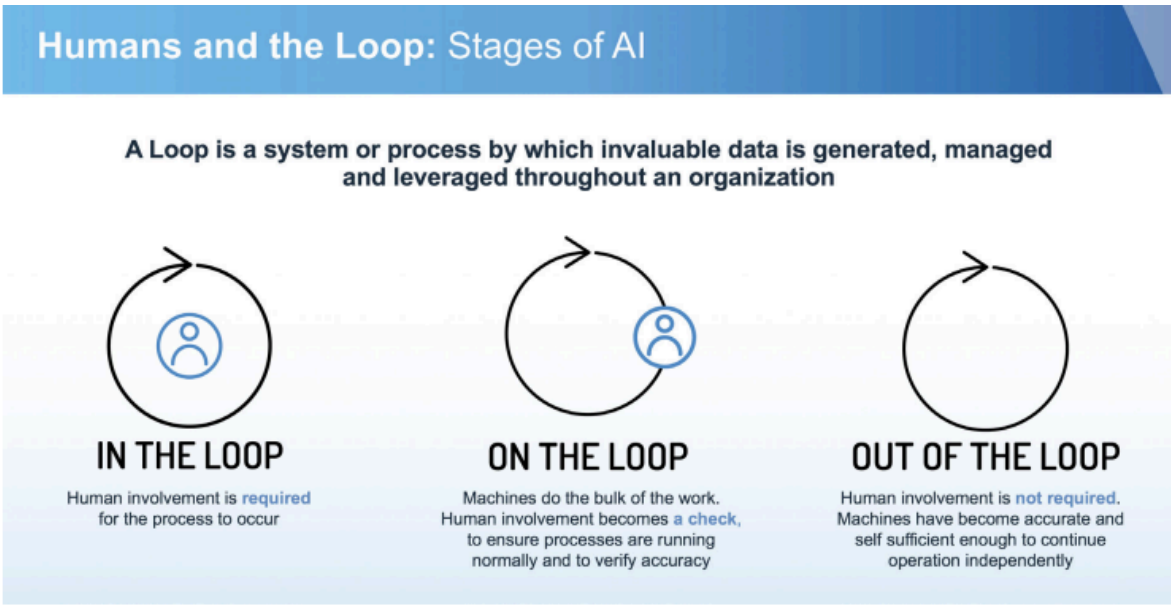
[67] Scharre 2019, 143
[68] Ibid., 139.
[69] Ibid., 143.
[70] Ibid., 162.
[71] Ruhl, 'Autonomous Weapon Systems & Military AI: Cause Area Investigation', 14.
[72] Wong et al., 'Deterrence in the Age of Thinking Machines', 66.
[73] Richard K. Betts, "Realism Is an Attitude, Not a Doctrine," The National Interest, August 2015, https://nationalinterest.org/feature/realism-attitude-not-doctrine-13659.

Increasing automation and integrating ML systems with nuclear command would apply all of the issues surrounding machine speed to nuclear decision-making, impacting and changing one of our most consequential weapons systems. Intimidating as that is, the reality of modern warfare could mean that working towards this level of haste may be necessary in order to maintain a competitive deterrence system. Other modern developments - such as the advent of hypersonic weapons or the potential for an offshore attack - have already shaved down a nuclear weapons strike to mere minutes. While distance matters even in an automated world,[74] a lightning fast first strike can only be deterred when a response at the same speed is feasible. Nonetheless, even if the need for speed in nuclear deterrence is unavoidable in our current security climate, it is imperative that we understand these problems and devise solutions or mitigating processes to prevent inadvertent nuclear use.

4.2.3 Out-of-the-loop

Machine Speed is a key aspect of automation that affects nuclear decision making, but it often does so by impacting the placement of the human supervision element in the decision making loop. As outlined in the below image, “there are three types of human supervision: a human can be in-the-loop, meaning a human will make final decisions; the human can be on-the-loop, supervising the system and data being generated; or the human can be out-of-the loop for full autonomy.”[75]

Roles for Humans and Machines in Decision Making



Source: www.datacenterdynamics.com/en/opinions/path-ai-connected-government.

[74]Wong et al., ‘Deterrence in the Age of Thinking Machines’, xi.
[75] Hruby and Miller, 2019, 9

The primary problem is that "AI systems operating at machine-speed could push the pace of combat to a point where the actions of machine [actors] surpass the (cognitive and physical) ability of human decision-makers to control (or even comprehend) events."[76] One proposed method for addressing the flaws of ML systems is ensuring that a human operator remains within-the-loop and able to control or stop automated military systems from acting in a manner that is unaligned with their objectives. However, as Andreas Matthias explained, any impactful human control is likely "impossible when the machine has an informational advantage over the operator ... [or] when the machine cannot be controlled by a human in real-time due to its processing speed and the multitude of operational variables."[77]

A 2020 RAND report explored the question: 'How might deterrence be affected by the proliferation of AI and autonomous systems?' by constructing and experimenting with a wargame to simulate how actors would make decisions in a conflict if automation became far more prevalent. A key insight from their experimentation was that "the differences in the ways two sides configure their human versus machine decisionmaking and their manned versus unmanned presence could affect escalatory dynamics during a crisis."[78] The following figure demonstrates that escalation is affected by whether the machine or the human makes the decision. (The RAND report also considered the effects of humans remaining physically present at the outset of conflict scenarios, but that is outside the scope of this report). A key takeaway here was that escalation was harder to control or prevent when automated machines were the primary decision-makers.[79]

Human and Machine Configurations and Potential Escalatory Dynamics

Physical Presence	Decisionmaking: Primarily Human	Decisionmaking: Primarily Machine
Human	Lower escalatory dynamic Higher cost of miscalculation	Higher escalatory dynamic Higher cost of miscalculation
Machine	Lower escalatory dynamic Lower cost of miscalculation	Higher escalatory dynamic Lower cost of miscalculation

Note. Reprinted from "Deterrence in the Age of Thinking Machines," by Wong et al, RAND Corporation, 27 January 2020, page 64.

[76]Johnson, 'Delegating Strategic Decision-Making to Machines', 459.
[77] Wong et al., 'Deterrence in the Age of Thinking Machines', xi.
[78]Ibid., 63.
[79]Ibid., 64.

Human-machine warfighting and 'teaming' has been progressively undertaken and its importance and impact will increase as it is attempted in more scenarios.The further humans are removed from the loop, the more reliant we become on possibly faulty technology that does not make decisions like humans do. This can become a strength but, as already discussed, it can easily become a danger instead. Nonetheless, many influential figures are still not opposed to pushing the human element further out. Gen. Terrence J. O'Shaughnessy, commander of NORAD stated that "What we have to get away from is … 'human in-the-loop,' or sometimes 'the human is the loop.'"[80] By doing so we can attempt to leverage the speed and power of automated systems while theoretically still ensuring that human hands guide the ML power technology. Regardless of how we manage this change in the human-machine relationship, the further humans are taken from the loop, the more the risk increases that a technical flaw will impact nuclear decision making. The question remains; at what point will the inflexibility of these systems, and the resulting escalatory potential, outweigh the advantages in speed that they offer?

This possible gap in reliability is a clear problem when deploying ML systems in complex safety critical environments like nuclear command. However, this almost alien decision further complicates issues due to the general human desire to defer to such machines as though they are human. Work done in this area has demonstrated that "humans are predisposed to treat machines (i.e., automated decision support aids) that share task-orientated responsibilities as 'team members,' and in many cases exhibit similar in-group favoritism as humans do with one another."[81] In fact, work by James Johnson showed that instead of constraining the brittleness or flaws of ML systems, keeping a human in the loop can "lead to similar psychological effects that occur when humans share responsibilities with other humans, whereby '[[social loafing]]' arises – the tendency of humans to seek ways to reduce their own effort when working redundantly within a group than when they work individually on a task."[82]

In the end, however powerful and useful, the decision-making process of ML enhanced systems can result in unexpected deviations or instances of misalignment. Despite their inherently inhuman nature, humans have the tendency to treat ML systems as colleagues rather than instruments, and expect human-like, aligned responses from them.

4.2.4 Misplaced Confidence

Similar to automation bias, or at least its outcome, misplaced confidence occurs when a system that seems superior and capable of providing a military advantage can create a sense of

[80]Jackson Barnett, 'AI Needs Humans "on the Loop" Not "in the Loop" for Nuke Detection, General Says', FedScoop, 14 February 2020, https://www.fedscoop.com/ai-should-have-human-on-the-loop-not-in-the-loop-when-it-comes-to-nuke-detection-general-says/.
[81]Johnson, 'Delegating Strategic Decision-Making to Machines', 445.
[82]Ibid.

confidence in military leaders that encourages them to take aggressive, risky, or drastic action. This is also linked to the importance of uncertainty in nuclear decision making.[83]

An overzealous commander believing themselves to have perfect information along with the advantage of machine speed might act aggressively if they believed the need was great enough or the chance of reprisal was low enough, thinking that the strength of the AI system would guarantee a successful operation. This problem is twofold; not only could overconfidence risk increasing the likelihood of aggressive actions in general, but it would also make a commander more susceptible to automation bias. Both are inherently escalatory and raise the risk of inadvertent nuclear weapons use. This kind of "overconfidence, caused or exacerbated by automation bias in the ability of AI systems to predict escalation and gauge intentions – and deter and counter threats more broadly – could embolden a state (especially in asymmetric information situations) to contemplate belligerent or provocative behavior; it might otherwise have thought too risky".[84]

5. Analysis of Potential Solutions

The following section covers solutions designed to address the problems outlined in this report. They are not necessarily unique to the issue of integration, nor are they being suggested for the first time in this report. Nonetheless, here they are evaluated through the combined lens of nuclear strategy and ML technology. In doing so I aim to continue this discussion by adding my thoughts on what could work best and how funders could use this assessment to aid them in nuclear or nearterm AI risk reduction efforts.

In general, I find the first two possibilities in 5.1 unsatisfactory as they fail to take into consideration the wide range of implications of the technical and psychological issues around AI integration on nuclear security.The three solutions in 5.2 are more likely to be successful as they take into consideration the fact that we may not fix the technical flaws within current AI technology. They either pursue stopgap measures to prevent inadvertent use or they attempt to address the psychological side of the problem. In doing so, they work with factors that we can control: policy and people. In 5.3 I present the complicated solution of keeping humans in the loop which, while impactful to pursue, cannot address the problems outlined here by itself.

5.1 Unsatisfactory Solutions

5.1.1 Do not integrate AI with Nuclear Command at all

While this solution would have a high degree of impact on preventing inadvertent use as a result of ML integration, it is not tractable due to the pressures to increase automation with NC3.

---

[83] Johnson, 'Delegating Strategic Decision-Making to Machines', 451.
[84] Ibid., 450.

At this time, no state or actor is advocating for completely autonomous nuclear weapons systems where the human element is completely out of the loop. This is important, and more states should follow the UK and U.S's examples of committing to this publicly in official strategy documents. Nevertheless the temptation may remain "for countries that feel relatively insecure about their nuclear arsenal, the potential benefits [of full automation] in terms of deterrence capability may outweigh the risks."[85]

Nevertheless, not integrating AI into any stages of NC3 at all is not a realistic option. As argued in the modernization section, it is extremely likely that the U.S. government is already pursuing some degree of AI integration at some stages, rendering this suggestion irrelevant. Additionally, there are real benefits to integration at some stages and this should be at least explored. Improving one's early warning system, if done correctly could help mitigate the dangers faced by nuclear weapons on the road to disarmament.

Beside providing better analysis, AI integration could help ensure robust communication in nuclear command that helps reduce uncertainty during a crisis.[86] Additionally, AI run cyber security could help secure key systems related to deterrence[87] and the safer and less vulnerable these weapons are, the less aggressive state may need to be. This is the same reason submarine based nuclear weapons are often touted as the current pinnacle of deterrence. They cannot be found and thus can reliably ensure a second-strike. Therefore, one can generally act with a greater degree of confidence and with less aggression because of their deterrent work. AI run cyber security could hopefully do the same, and in all honestly may simply be required in the face of AI run offensive cyber actions. Given the security climate, not pursuing AI integration could result in a lopsided or asymmetrical environment that itself encourages the kind of nuclear crisis or coercion that increases the risk of nuclear use.

There is still a lot of value in criticizing or supporting what appears to be an inevitable policy. Advocating for the elimination of nuclear silos and landlocked nuclear weapons may be difficult for a number of reasons but researchers and activists should not necessarily stop. The same can be said about advocating for zero ML integration to NC3. I personally do not suggest this solution for funders and grantmakers attempting to maximize impact per dollar spent.

5.1.2 Improve AI to eliminate technical problems

Attempts to 'just make it better' may fix some of the technical problems eventually but the perfect system will never exist. While solving key technical issues would have a large degree of impact, andmany AI researchers are exploring this route, it seems like an unlikely achievement. This in and of itself isn't a failure. Humans are not perfect either, but problems like brittleness or trained AI bias are especially dangerous because they could rear their heads suddenly and without warning and because of the 'black box' issue we might not be able to routinely check for these problems nor will we always be able to test for them.

---

[85] Boulanin et al, 2020: 109
[86] Hruby and Miller, 2019, 12
[87] Ibid., 13.

‘Normal Accident Theory’ “suggests that: as system complexity increases, the risk of accidents increases as well and that some level of accidents are inevitable in complex systems.”[88] Therefore,the risk of accidents in complex defence systems that incorporate autonomy may therefore be higher.”[89] The reality is that “even with simulations that test millions of scenarios, fully testing all the possible scenarios a complex autonomous system might encounter is effectively impossible. There are simply too many possible interactions between the system and its environment and even within the system itself.”[90]

Furthermore, even if we implement safety features designed to stop accidents from occurring, it is often these very features that result in deadly errors and accidents. Two recent examples of this include Lion Air Flight 610 on October 29, 2018, and Ethiopian Airlines Flight 302 on March 10, 2019. There was a safety feature that resulted in the planes crashing.[91] The point is that not only are accidents normal and difficult to prevent in complex systems, but that even attempts to combat them can result in accidents themselves.

Perhaps the math works out that humans fail catastrophically more often than AI would (even if the AI is brittle or misaligned). Nonetheless, when combined with the issues forming in the human-machine relationship, the dangers of AI integration could go undiagnosed due to the implicit trust that the AI system is objectively better than its human counterparts. If automation increases and more nuclear decision-making is pre-delegated to machine systems, these problems will grow and risk inadvertent use in this safety critical environment.

Additionally, the consequences of human failure are mitigated by human uncertainty and the supervision and control of peers and superiors. Throughout our entire history, humans worked together. Switching from human to AI is a clear paradigm shift for both decision making and safety culture, drastically shifting our reactions and approaches to problems in crisis.

This is not an outright rejection of ML in NC3 but rather a claim that we cannot rely on technological improvements to remove the possibility of accidents.

---

[88] Michael C. Horowitz, “When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability,” *Journal of Strategic Studies* 42, no. 6 (September 19, 2019): 764–88, https://doi.org/10.1080/01402390.2019.1621174

[89]Ruhl, ‘Autonomous Weapon Systems & Military AI: Cause Area Investigation’, 21.

[90] Scharre, 2019, 149.

[91]Mina Kaji, Amanda Maile, and Gio Benitez, ‘1 Year after the Ethiopian Air Flight 302 Crash, Questions Remain as to When Boeing’s 737 Max Will Fly Again’, ABC News, accessed 2 October 2022,

5.2 Potentially Beneficial Solutions

5.2.1 Update Nuclear Posture to Reflect Changing Paradigm

One of the greatest hurdles to mitigating the challenge of ML in NC3 is the fact that the technical problems discussed above may simply be unsolvable in the near-term. Approaching this problem from a non-technical angle may allow us to bypass the inherent technical flaws within ML systems by mitigating the dangers without necessarily resolving these technical problems. Specifically, policy can be restructured to mitigate the potential problem of ML integration with NC3.

Although technical solutions may be developed to lessen the impact of these limitations, the sheer complexity of AI machines and the pressure to integrate them with military systems mean that the problems these technical issues pose must be addressed now. The heavy incentive for states to start incorporating the technology as soon as possible means they may have to accept and implement imperfect systems into critical roles in nuclear command. In the end, however, the real problem is not the AI per se, but a rushed integration of AI with nuclear systems that does not fully take into consideration the heightened risks posed by the technical limitations of current AI technology and the complexity of nuclear security.

One solution is adopting a nuclear policy that expands the decision-making time for launching a nuclear weapon, such as by moving away from launch-on-warning (LOW) strategies or by “de-alerting” silo- based intercontinental ballistic missiles. The LOW strategy keeps missiles alert and constantly ready to fire so they can be launched before the first impact of an incoming attack. By legally increasing the appropriate amount of time to ready nuclear weapons for use, these types of policies would allow leaders more opportunity to assess the nuclear security-related information provided by the AI. In a sense, this would forcibly elongate the ‘loop’ so that regardless of lightning fast AI assessments, a specific no-first-use policy and a restriction on a rapid launch could lead to a less dangerous AI system or even make it effectively impossible for inadvertent use to occur.

Creating shifts in nuclear posture that increase decisions making time or move away from LOW strategies will result in innate advantages for reducing nuclear risk before we even consider their influence on ML integration with nuclear command. The risks of LOW can be broken down into two categories: the intense pressure under which the decisions are made, and the often-questionable quality of the information used to make decisions. As the many close calls and cases of near nuclear use demonstrate, LOW is already a dangerous prospect whose instance of responding before impact allows for the possibility of inadvertent use and catastrophe. The idea of moving away from a LOW posture is not a new idea but it made all the

more relevant when considering the potential dangers associated with ML integration with nuclear command systems.

Ultimately, integrating ML into nuclear command would act as a compounding factor or threat multiplier for inadvertent use if done improperly. On the one hand, a working ML system could provide better information faster than our current systems do. This would rescue the chance of inadvertent use and potentially increase the amount of time decision makers have to determine whether they will respond or whether it is a false alarm. This could help reduce the risks associated with LOW strategies. However, while possible, these benefits rely on a number of assumptions regarding the ability of incredibly complex systems to work under immense pressure.

The risks outlined in this report should give one pause when considering the helpfulness of ML systems for reducing nuclear risk. This is not to say they have no place in risk reduction — on the contrary, finding the proper balance for safe AI integration should be pursued in the quest for nuclear risk reduction. Nonetheless, increasing automation within the systems that are integral to managing a LOW nuclear doctrine means potentially falling prey to the various technical and human-machine relationship issues outlined in this document. While we may be able to address some of these problems through technical or training based means, changing policy to make it nearly impossible to launch nuclear weapons within a few minutes could effectively mitigate the greatest risks associated with ML integration. While the danger of inadvertent use would still exist, it would be drastically reduced by this kind of a shift in posture. Still, it should be noted that moving away from Launch-On-Warning is unlikely as there is heavy domestic political pressure to maintain ICBMs and other more static nuclear forces.

An alternative solution is the commitment to move away from LOW during periods of peace and stability. Originally proposed by Podvig, this idea encourages nuclear weapon states to “introduce a policy of keeping their forces off alert most of the time”.[92] This would help reduce the chance of peacetime miscalculations. While escalation in times of stability is not as likely as it is in crisis scenarios, Barrett, Baum, and Hostetler demonstrated that half of all false alarm cases occurred during periods of low tension.[93] Crises are still far more dangerous overall as half of all false alarms and inadvertent use scenarios occur during a comparably tiny amount of time when compared to periods of lower-tension. Nonetheless, this approach would effectively end peacetime inadvertent use risk and present a possible confidence-building-measure that could help increase crisis stability by enabling both informal and more structured talks. Additionally, it would signal the universal desire for safety and stability and a general lack of interest in conducting a first strike with these weapons.

---

[92]Pavel Podvig, ‘Reducing the Risk of an Accidental Launch’, *Science & Global Security* 14, no. 2–3 (1 December 2006): 89, https://doi.org/10.1080/08929880600992990.

[93]Anthony M. Barrett, Seth D. Baum, and Kelly Hostetler, ‘Analyzing and Reducing the Risks of Inadvertent Nuclear War Between the United States and Russia’, *Science & Global Security* 21, no. 2 (2013): 127.

And yet, changing nuclear doctrine is no easy feat even in the best of environments. Given the current war in Ukraine and revamped tensions over Taiwan, advocating for changes to doctrine at this time is perhaps more difficult than ever before. Successfully updating posture could effectively negate the risk of inadvertent use, but the tractability of this is incredibly low in the current international security climate.

5.2.2 Update and ensure adequate training

How we train the human element will be crucial to ensuring the safe integration of the human-machine team. This suggestion covers the training of all human components, whatever role they may play in the decisions making or support process.

Paradoxically, the more autonomous a machine, the more training is required for the humans involved with it. This training needs to reflect functionality as well as the machine-human relationship and the inherent flaws within the technology.[94] In order to maintain the benefits of 'human uncertainty' in nuclear decision making, training should embed a healthy degree of skepticism or doubt toward the militarized AI in its human operators. Properly done, this balanced training will avoid creating a 'trust gap' or an automation bias by highlighting both the potential benefits and risks of ML integration.

With the understanding that AI is flawed, perhaps the next crisis will reflect the one in the 1980s where Stanislav Petrov doubted an early warning system telling him, with the highest level of confidence, that there was an incoming U.S. nuclear attack.[95] Petrov's uncertainty saved the day. Conversely, the operators of the 2003 Patriot systems that killed friendly aircraft were found to have a culture of "trusting the system without question."[96] To properly manage the flaws of powerful ML systems, people need to be trained to understand "the boundaries of the system - what it can and cannot do. The user can either steer the system away from situations outside the bounds of its design or knowingly account for and accept the risks of failure."[97] Both the Petrov case and the Patriot fratricides outline the importance of instilling the proper amount of confidence, and skepticism, when training operators of automated military systems.

A different but linked idea is creating organizational and bureaucratic solutions to address the technical problem of militarized AI. The SUBSAFE program is a "continuous process of quality assurance and quality control applied across the entire submarine's life cycle."[98] Between 1915 and 1963, the U.S. lost an average of one submarine every three years to non-combat losses; since the program was established in 1963 not a single SUBSAFE

---

[94] Boulanin et al, 2020, 128
[95] Stanislav Petrov, interviewed in Vasilyev, Yuri (2004), 'On the Brink', The Moscow News, 29 May, http://www.brightstarsound.com/world_hero/the_moscow_news.html.
[96] Scharre 2019, 144
[97] Ibid., 146.
[98] Ibid., 161-162.

certified submarine has been lost.[99] This is even more impressive when considering both the increased complexity of modern submarines and their operating environments. This seemingly counteracts Normal Accident Theory.

Perhaps this kind of safety culture can act as a model for handling militarized AI systems. According to Paul Scharre, these lessons are already applied to the operation of the navy Aegis combat system, which was not trusted by its human operators.[100] They understood

---

[99] Scharre 2019, 162.
[100] Ibid.

Navigating our relationship with machines, especially in a military context, is integral to safeguarding human lives. When considering this challenge at the already contentious nuclear level, our concern expands from preserving human lives to preserving human civilization. Undoubtedly, any development that alters decision making around nuclear weapons use demands our close attention. AI and ML stand to enable a significant increase in automation throughout Nuclear Command, Control, and Communications (NC3). Only time will reveal the exact way this will unfold, but early warning and decision support systems seem to be likely candidates for high levels of automation.

Integrating ML systems could increase the accuracy of these systems while also potentially removing human related errors. However, given their importance in detection, and their history with false positives, anything that impacts early warning and decision support systems requires the utmost scrutiny even if the change stands to potentially improve the safety of such systems.

This report is designed to lay out the problem, analyze possible solutions, and present funding opportunities designed to support impactful projects and developments. However, in a manner more akin to the academic paper, the through line in this report highlights that the automation ML fessably brings to NC3 could have substantial impacts on nuclear decision making. By increasing automation, we are effectively ‘pre-delegating’ authority to machine intelligences that are inevitably flawed systems despite the advantages they could bring. In one sense, their imperfection does not in-of-itself derail any argument in favor of ML integration. They would either work alongside, or replace, human operators who are also both flawed and limited in what they can do. However, the true danger of relying on flawed machines becomes apparent when we consider how automation will affect the human-machine relationship.

The very purpose of developing AI is for its theoretical ability to perform better than humans - to be faster, stronger, and continuously vigilant. However, overconfidence in their abilities could result in over deployment of the technology and premature ‘pre-delegation’ of responsibility. This not only sets the stage for technical flaws within the modern and nearterm ML systems to increase the risk of catastrophes, it also removes humanity from this process. While humans are far from perfect, we’ve managed to avoid inadvertent nuclear weapons use and nuclear war for just under a century. I argue that human weakness itself played a key role in preventing inadvertent use of nuclear weapons. Therefore, the increasing automation that

accompanies ML integration with NC3 could represent a dramatic shift in nuclear weapons decision making processes and this exacerbates pre-existing risks around inadvertent use.

1.2 Possible interventions?

A philanthropist interested in reducing nuclear risk stemming from a poorly planned integration of ML with NC3 could support efforts to increase research and discussion on this topic, and/or efforts to provide stronger empirical support of these concepts and possibilities.

Possible interventions include funding new experimental wargaming efforts, funding research at key think tanks, and increasing international AI governance efforts. While far from an exhausted list, some of the following think tanks and research groups are likely impactful candidates:

1. The United Nations Institute for Disarmament Research
2. Stockholm International Peace Research Institute
3. RAND Corporation
4. The Centre for Strategic and International Studies
5. The Nuclear Threat Initiative

I look at each intervention in more detail at the end of the document.

1.3 Target Audience

This report is designed to outline the effect that integrating AI/ML with NC3 could have for the risk of inadvertent use. It is also primarily designed for an audience already somewhat versed in nuclear deterrence literature with an understanding of what crisis stability/instability entails. Nonetheless, this report does delve into some of these topics insofar as they directly interact with the main argument. What this report is not is a deepdive into the computer science behind AI technology, nor is it a psychological piece of research designed to act as the definitive piece on the human-machine relationship.

This work aims to build off insights from both the fields of nuclear deterrence and AI safety. In synthesizing the literature of these two fields I hope to inform nuclear grantmakers of the risks involved in this integration and to provide direction for further research or other efforts to mitigate this problem.

1.4 Scope of Analysis

This work focuses on the U.S. and allied Western states: the U.K., France, and Israel. The reason for this is threefold:

1) Systems between different powers differ and of the nuclear powers, I am best positioned to examine western nuclear command.

2) The problems addressed are related to the technology itself: how it could potentially impact the human-machine relationship and affect nuclear decision-making. These issues are not state specific, so omitting other nuclear states from the scope of this report is not a significant shortcoming.
3) Cooperative efforts to reduce nuclear risk are arguably preferable, but unilateral action matters and can make a positive impact on nuclear stability.

In this case, any effort to mitigate the negative side effects of AI integration with nuclear command will likely lead to a net-positive outcome in terms of safety for all actors. Therefore, one can primarily focus on the U.S. and still provide reasonable suggestions for reducing nuclear risk.

There is also the question of alarmism. Some thinkers have rightly pointed out that discussions on emerging technologies often amounts to a dangerous form of alarmism.# Historically, other technologies that were predicted to change the nature of warfare, such as chemical weapons, failed to live up to these expectations.# Other times, “even when technologies do have significant strategic consequences, they often take decades to emerge, as the invention of airplanes and tanks illustrates.”# The notable exception to this was the advent of nuclear weapons. The undeniability of their sheer destructive power has dominated international security and great power interactions since their conception. While the details of how AI will impact both the world at large, and the military context specifically, rests uncertain, there are reasons to be concerned that it will be more akin to nuclear weapons than other overhyped technologies. AI has been likened to electricity; “like electricity brings objects all around us to life with power, so too will AI bring them to life with intelligence.”# Others have stated that it will be “the biggest geopolitical revolution in human history.”# The potentially massive impact of strategic military AI, and the relatively small amount of philanthropic directed to this issue, highlight the need to outline potential dangers.

1.5 Outline of Report

This report starts with defining what is meant when discussing AI and NC3. Second, it then outlines the pressures to integrate AI with NC3, evidence to support this claim, and how this modernization could take place. Third, the report portrays the mechanisms that could lead to the risk of increased nuclear use risk: technical flaws within AI systems and problems surrounding the human-machine relationship . Forth, the report briefly analyzes a variety of solutions to these risks and outlines which seem most likely to have a positive impact on risk reduction. Fifth and finally, it provides a list of tractable funding opportunities.

2. Defining Key Aspects of the Discussion

2.1 Artificial Intelligence and Machine Learning

Automation, to varying degrees, has been a part of deterrence and command/control since the dawn of nuclear weapons.# “Modern AI” as it is publicly imagined is a relatively new addition. The Soviet Perimetr system, known as the ‘Dead Hand’ in the West, is an example of a more extreme form of automation with nuclear weapons. This was an automated NC3 system designed to react in a situation in which a potential nuclear detonation was detected and communication with national leadership was dead. The system could interpret this as an indication of a nuclear attack against Russia,# and give authority and ability to a human operator in a hardened bunker to launch nuclear missiles in response.# This operator would likely have very little information other than what the Perimetr system provided and would have to decide whether or not to trust its determination.

There is disagreement as to how developed this Perimetr system actually was.# Nonetheless, it acts as an example of the potential pitfalls associated with linking a ‘Doomsday Device’ to a machine intelligence. Fortunately, both the U.S. and U.K. have made formal declarations that humans will always retain political control and remain in the decision-making loop when nuclear weapons are concerned.#

The term “AI” invokes a myriad of different definitions, from killer robots to programs that classify images of dogs. What brings these computational processes together under the umbrella term of artificial intelligence is their ability to solve problems or perform functions that traditionally require human levels of cognition.#

Machine learning (ML) is an important subfield of this research as these systems learn “by finding statistical relationships in past data.”# To do so, they are trained using large data sets of real-world information. For example, image classifiers are shown millions of images of a specific type and form.# From there, they can look at new images and use their trained knowledge to determine what they are seeing. There are also a number of ways one can train an ML system. Training can be supervised, where data sets are pre-labeled by humans, or unsupervised, where the AI finds “hidden patterns or data groupings without the need for human intervention.”#

With this in mind, while ML is the overarching driver behind the renaissance of intelligent machines, its own subset of deep learning is arguably the main way in which AI will be used in nuclear command. Deep learning differs from machine learning through the number, or depth, of its neural network layers, which allows it to automate much of the training process and requires less human involvement in its training.# This automation allows the system to use unstructured or unlabeled data to train itself by finding patterns within larger datasets that are not curated by

humans. This makes them extremely useful for recognizing patterns and managing and assessing data for “systems that the armed forces use for intelligence, strategic stability and nuclear risk.”#

However, because the ways in which AI could be integrated with NC3 could be incredibly varied and broad, I will generally use the ML terminology when discussing AI in this report. To put this in context, one study found that potentially 39% of the subsystems that make up NC3, could be integrated with ML.# While deep learning may be crucial to the success of AI within command systems, it is but one, very important, subset of ML, and therefore this report will focus on the idea of machine learning.

Machine learning is the key to the advent of intelligent machines within nuclear command, and the autonomy it enables is arguably the most important benefit. Autonomy, or ‘machine autonomy,’ “can be defined as the ability of a machine to execute a task or tasks without human input, using interactions of computer programming with the environment.”#

By removing human operators from the decision making loop in certain instances, one can better leverage both the operating speed, and the skill of ML systems to find hidden patterns in large complicated data sets. This presents a large strategic advantage within a military context where there is a premium on haste and reliable information. For the purpose of this report, AI integration will be generally discussed in terms of ML, automation, and its ability to assess data and provide analysis.

2.2 Nuclear Command, Control, and Communications (NC3)

NC3 is “the combination of warning, communication, and weapon systems–as well as human analysts, decision-makers, and operators–involved in ordering and executing nuclear strikes, as well as preventing unauthorized use of nuclear weapons”.#

New and more powerful ML systems could be used to improve the speed and quality of assessment completed by NC3. ML systems’ ability to find correlations by continuously sorting through large amounts of data with an objective eye is particularly relevant to early-warning systems and pre-launch detection activities within the nuclear security field.

The other core reason for focusing on early warning and decision-support systems within NC3 is their susceptibility and influence on the possibility of inadvertent use. Since the inception of nuclear weapons, there has been a plethora of false alarms and false positives due to both technical and human error.# A now classic example of this occurred in the Soviet Union in September 1983 when Lieutenant Colonel Stanislav Yevgrafovich Petrov’s early warning system falsely detected five incoming U.S. Minuteman intercontinental ballistic missiles

that “the automation was powerful and they respected it - they even recognized there was a

(ICBMs). The system confirmed the attack with the highest level of confidence with a probability factor of two.# Despite this, he had reservations about the system’s capability and accuracy and realized that the incoming attack did not fit Soviet strategic doctrine as it was far too small in scale.# Ultimately he dedicated to report it as a false alarm and his intuition was correct. The machine had made the wrong call and Petrov’s skepticism and critical thinking combined likely contributed to preventing Soviet retaliation.

On the other side of the Iron Curtain, NORAD was not free from human and technical error resulting in false alarm situations. In November 1979 an early warning system was accidentally fed test scenario data designed to simulate an incoming Soviet nuclear attack. Lucky radar was about to confirm that this was a mistake and in 1980 NORAD “changed its rules and standards regarding the evidence needed to support a launch on warning.”# However, this isn’t the only example as less than a year later a faulty computer chip caused the early warning system to detect what looked like an incoming Soviet attack.# At 02:26 on 3 June 1980, National Security Advisor Zbigniew Brzezinski received a telephone call informing him that 220 missiles had been fired at the U.S, which was then confirmed in another call with the number of missiles being raised to 2,200.# At the literal last minute before he was about to inform President Carter did Brzezinski receive the final call telling him it was a false alarm that had been caused by a faulty computer chip. In one sense these stories demonstrate that even in the face of complex and flawed systems, organizational safety measures can prevent inadvertent use. However, they also speak to the frightening ease with which we arrive at the potential brink of nuclear use when even a small mistake is made.

Integrating ML systems could increase the accuracy of these systems while also potentially removing the human source of these errors. However, given their importance in detection, and their history with false positives, anything that impacts early warning and decision support systems requires the utmost scrutiny even if the change stands to potentially improve the safety of such systems.

3. Pressures to Integrate AI with NC3 and the Shape It Could Take:

3.1 Pressures to integrate ML with nuclear systems

Despite the outlined dangers of AI integration, there are pressures and incentives to use ML systems within NC3. As previously stated, automation has always played a role in nuclear strategy and deterrence. ML could take the level of automation to new heights and potentially change nuclear weapons decision-making.

place for it - but that didn't mean they were surrendering their human decision-making to the machine."[101] The teachings from the SUBSAFE program and modern Aegis system act as key examples for how we can instill a healthy degree of skepticism in any instances where ML is integrated with NC3. One can respect both the power *and* the flaws of automated military technology.

This suggestion will have low to medium impact as where and how the training occurs matter greatly, and errors will still occur, but it is highly tractable as militaries have a vested interest and desire to ensure their operators have any relevant training.

5.2.3 Confidence Building Measures

Confidence building measures (CBM) could create international norms around appropriate AI integration and reduce uncertainty between nations on acceptable military use of AI.

CBMs could start as unilateral declarations that nuclear launch decisions will always remain under human control. That first step could bring other powers to the table and potentially result in similar declarations if these CBMs align with their strategic interest. The UK government has already made this declaration in their June 2022 Defence Artificial Intelligence Strategy.[102] Additionally, the U.S. government recently followed suit in their 2022 National Defence Strategy.[103] While we can't know the exact implications of this statement, this kind of action is conducive to reducing tensions and uncertainty.

Going further, an international dialogue about AI integration with NC3 would build confidence and also allow states to share common concerns and determine accepted best practices. Recent work looking at the viability of controlling militarized AI through international measures and norms saw "that while past strategies to contain and control nuclear weapons cannot and should not be taken as blueprints, these historical lessons remain essential in designing any future efforts to responsibly contain military AI."[104]

The Anti-Ballistic Missile (ABM) Treaty was a key example of a de-escalatory CBM that allowed both Cold War powers to demonstrate their commitment to maintaining stability.. By agreeing to limit the development of missile defense systems, both sides signaled that circumventing Mutually Assured Destruction (MAD) and even winning a nuclear war would be too destabilizing and dangerous. While adversarial nations may often feel the need to consistently doubt and fear each other's intentions and actions, these CBMs help cut through

---

[101] Ibid., 168.
[102]'Defence Artificial Intelligence Strategy', 59.
[103] U.S. Department of Defence. 'National Defence Strategy, 2022, 49, 'https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.PDF
[104] Matthijs M. Maas, 'How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons', *Contemporary Security Policy* 40, no. 3 (3 July 2019): 287, https://doi.org/10.1080/13523260.2019.1576464.

the ‘security dilemma’, reduce uncertainty between states regarding military uses of AI, and mitigate the proverbial ‘noise’ around decision-making.

Naturally states are limited on how much they can share about their ML developments to keep adversaries from taking advantage of the information to manipulate or attack the systems. Nonetheless, this kind of cooperative effort should be explored to create a degree of confidence between states and allow experts within and outside government to determine and spread best safety practices.

The impact of CBMs is complicated to assess; individual measures’ impacts are immensely varied. However,even low impact measures are important to reducing the risk of inadvertent nuclear use. Lowering tensions and reducing uncertainty are precursors to changing posture or preventing a race to bottom in AI safety. The tractability of this is again variable due to changing international environments, but at this time any multilateral agreements are unlikely due to high-tensions around the war in Ukraine. Conversely, this could also be the time when the tractability of low-cost CBMs go up due to a serious need to reduce tensions between Russia and the U.S. While the war in Ukraine undoubtedly makes cooperation difficult, we could also see states coming to the table at a high-level of diplomacy to make low-cost CBMs designed to reduce tension around nuclear weapons as a first step towards a potential return to a semblance of normality. All that being said, the tractability of CBMs is incredibly variable.

5.3 The Complicated Humans in-the-loop Solution

While it is ultimately useful to keep humans in-the-loop, and human oversight is still required to some degree, there are multiple issues with this approach to control and limit automation. First, machine speed in highly competitive scenarios will effectively remove humans from the equation. The speed of modern warfare increased by hypersonic weapons, cyber warfare, and other new technologies could simply require autonomous systems to be given the pre-delegated authority to act without meaningful human oversight. Demanding these systems to always have a human in-the-loop would effectively negate the reason for their use in the first place. Of course, it can be argued that this is completely acceptable and that we should not race to the bottom of AI safety and risk nuclear catastrophe in order to stay allegedly equal to our military adversaries. Regardless of where one falls in that debate, machine-speed makes keeping a human in-the-loop difficult.

Secondly, keeping a human in-the-loop does not address the ‘black box’ problem. Meaningful human control is negated when we don’t know why the ML system is making its choices. While one can evaluate the outcomes, and this may be enough in many scenarios, not knowing if the ML system is actually working as intended opens us up to unpredictable ML outcomes. The human operator may be able to control for obviously wrong outcomes, but brittleness could lead to incorrect assessments proposed with a high level of confidence that could be undetected until a catastrophe arises during a crisis.

Finally, automation bias can make the human element effectively irrelevant. If the bias develops, a human in-the-loop who is inclined to defer to the machine intelligence will be inherently less likely to act as a good check on its actions. Therefore, due to the problems listed here, keeping humans in-the-loop cannot arguably be relied on to prevent an increase in risk.

However, despite not being the ultimate answer to the issues outlined in this report, attempting to maintain human control over AI-powered nuclear C3 by preserving a human position somewhere within the decision making loop is still worth pursuing. And yet, depending on the situation, the human may need to be in, on, or even out of the loop, so while this approach is respectable , keeping humans in-the-loop is not a final catchall solution.

6. Potential Funding Avenues

The early and relatively underexplored nature of this topic means that there are significant funding opportunities around expanding literature and study, while also looking for ways to improve the empirical strength of these ideas.

6.1 Wargame Funding

- Funding the creation, use, and research of wargames designed to demonstrate nuclear decision making using various forms of AI decision support
  - Wargaming could add a form of empirical evidence to the discussion of how machine learning could impact nuclear deterrence and/or nuclear decision making
  - Funding could be used to create the game, execute it, and sustain thorough research based on the results.
- The use of wargaming could assist with two satisfactory solutions listed above.
  - It could provide evidence and insight into the areas where nuclear and strategy doctrine is lacking when taking into account how automation could be integrated with NC3. While wargaming is clearly limited by the medium of the game, it can demonstrate how humans and machines could interact in a crisis.
  - Alongside its implication for policy and doctrine, the observations regarding human participants and their actions could also help determine where and how training could be improved. Clear examples of automation bias or situations where pre-delegating authority to machine intelligence is problematic can inform human operators in real-world situations.
- Additionally, while keeping humans in-the-loop may have been critiqued as a non-successful solution to the problems of integration, wargaming could identify best practices as to where and when humans should or must be kept within the loop of automated systems.
  - My critique of the in-the-loop solution was based around the fact that the perceived need for machine speed and the whole reason one would want

automation, could push humans out-of-the-loop. Therefore, trying to always have a human play the role of final call in an ML integrated system might simply not be feasible and thus cannot solve the problems outlined here by itself. However, ensuring human oversight when possible or where required will still be crucial to the safe use of ML systems. With this in mind, wargaming can determine where we must endeavor to maintain human control or final say as well as where human control could be detrimental to stability.

- A 2020 RAND report explored the question; 'How might deterrence be affected by the proliferation of AI and autonomous systems?' They did this by constructing and experimenting with a wargame to simulate how actors would make decisions in a conflict if automation became far more prevalent.
  - While focusing on a different aspect of militarized AI, their insights are undoubtedly valuable when looking at complex military operating environments, the machine-human relationship, and how deterrence could be impacted by AI.
  - They effectively added empirical evidence to a subject that is classified and somewhat speculative in nature.
  - One very important takeaway for the subject of inadvertent use is that in their wargame "the U.S. and Japan players set their air defense systems to be fully autonomous…however, when North Korea unexpectedly launched a missile over Japan, the AI in the system not only shot down the missile but launched counterbattery fire that hit North Korea."[105] Neither player had intended to strike at North Korea but the machine-speed of conflict resulted in an inadvertent use of aggression before they could stop it from occurring.
  - There is currently very limited efforts to simulate how ML systems will impact deterrence or NC3 and funding here could provide strong empirical evidence and insights into what the future of nuclear deterrence will look like in a world with powerful militarized machine intelligences.
- The Nuclear Threat Initiative created the wargame/tabletop exercise: "Strengthening Global Systems to Prevent and Respond to High-Consequence Biological Threats".
  - It was designed to provide insight to high-consequence biological events and to determine how we might better improve prevention and response capabilities.
  - While not as directly linked to this report as the RAND wargaming was, these kinds of exercises are designed to explore crises which are by their nature, and luckily, a rarity. However, despite their rarity, failure to navigate them can result in immense suffering or even global catastrophic risk.
- I recognize the inherent limitations of wargaming and caution falling prey to any bias in favor of them.

[105]Wong et al., 'Deterrence in the Age of Thinking Machines', 52.

        - They are, as their name suggests, games and not a pure reflection of the real world. Rules and structure are designed to provide an accurate simulation but there will always be flaws and failures. One key one is that in wargaming participants in the RAND game were often overly aggressive, likely due to the artificial nature of the game.[106]
    - Nonetheless, I believe that wargaming can be immensely useful for preparing for crisis events. We cannot know the future, but we can prepare for it.
        - While we can always look back at past real-world events for guidance and analysis, what we may face in the future likely won't be properly represented by what has happened before. Therefore, working to understand how these events will unfold is an undeniably useful tool for planners when accompanied by analysis and research.
    - To that end, funding for wargaming on the integration of ML with nuclear command is made all the more important by the flaws of wargaming. No one game will be a perfect exploration of this kind of crisis. Therefore, a wide range of wargaming and analysis needs to take place to ensure that, when taken all together, we can paint an accurate picture of this nearterm risk.

6.2 Targeting Funding for Influential Think Tanks

- Targeted funding for think tanks could promote diverse research in this area. I suggest focusing on think tanks that often work alongside the defense and policy establishment of nuclear armed states.
    - Impact on this issue could be found by funding research on the impact of AI on NC3.
    - By targeting specific think tanks in the U.S., UK, France, etc., one could reach and influence policy makers and other key individuals who rely on the research from these institutions.
        - While far from an exhausted list, some of the following think tanks and research groups are likely impactful candidates:
            - The United Nations Institute for Disarmament Research
            - Stockholm International Peace Research Institute
            - RAND Corporation
            - The Centre for Strategic and International Studies
            - The Nuclear Threat Initiative
        - Future work should attempt to determine the ability for each organization to impact the risk of inadvertent nuclear use. While each organization may be effective limited funding toward nuclear

[106]Wong et al., 'Deterrence in the Age of Thinking Machines', 50.

risk reduction at this time[107] may require an effective impact evaluation.

- Although targeting the funding can ensure the greatest impact per dollar spent, more generalized funding for this research would ensure a wide range of thought on the subject. This includes:
  - funding researchers who are critical of governments, nuclear weapons, and AI.
  - funding early career individuals who require the runway provided by grants and paid opportunities to develop their expertise in the field for the long term.

6.3 Funding for AI Governance and CBMs

- Funding AI Governance efforts and CBMs that take into account the strategic impact machine learning could have if/when integrated with NC3.
  - Much of the international focus on autonomous weapons is on 'slaughter bots'
  - Increasing funding for efforts to promote international cooperation and declaratory statements between states on ML integration with their nuclear command systems could
    - lower risk by creating dialogue and removing a degree of uncertainty
    - help reorganized the debate surrounding autonomous weapons to factor in the more "boring" lines of risk and the strategic impacts that ML could have
- These measure could take a multitude of shapes:
  - A conference of experts from different states and from the private sector
    - Recent work done by Christian Ruhl and Founder Pledge explored autonomous weapons and militarized AI, and outlined how funding track II dialogues and workshops could mitigate risks associated with autonomous weapons and great power conflict.[108]
    - These unofficial and non-governmental efforts can bypass barriers to state cooperation and allow for experts from both sides of an adversarial dynamic to build relationships, share points of view, and take insights and best practices back home.
  - Research aimed at creating and evaluating various international governance efforts for arms control and AI.
    - Not all governance is created equal, nor can individual successful accomplishment be directly applied to a novel issue. Determining

[107]Bryan Bender, '"A Big Blow": Washington's Arms Controllers Brace for Loss of Their Biggest Backer', POLITICO, accessed 2 October 2022, https://www.politico.com/news/2021/07/19/washington-arms-controllers-nuclear-weapons-500126.

[108]Ruhl, 'Autonomous Weapon Systems & Military AI: Cause Area Investigation'.

which CBMs would have the greatest impact on mitigating the risks of AI integration with nuclear command will require in-depth study and gathering of experts at workshops and conferences.

- The ideal outcome of this funding would be novel treaties and other legally binding documents that can be signed and ratified by states. While funding cannot be directly applied to this level of official diplomacy (track I), behind the scenes funding for experts and unofficial meetings builds a core intellectual base on which formal diplomacy could occur.
  - In extreme examples such as treaties outlawing certain uses of militarized AI, non-proliferation work could prove valuable even without key AI actors.
    - Categorizing militarized AI at the same level as other banned weapons, combined with the non-proliferation efforts around nuclear weapons, could impact how autonomous ML systems within the military are seen, and how the general public/domestic audiences understand the potential danger of ML within state miliaties.

7. Conclusion

To conclude this report, I will briefly touch on its core ideas and findings. ML integration with NC3 has the potential to drastically alter our decision making process and this demands scrutiny due to the safety critical requirements of nuclear security. Additionally, there is the possibility that moving from human to machine intelligence could eliminate key aspects of human nature that have helped prevent nuclear weapon use since 1945.

To address this, we should aim for solutions that affect the policy and people involved with nuclear decision making as the technical aspects of this problem may not be solvable for near term considerations. If done correctly, increasing the automation of NC3 through ML integration could significantly decrease the chance of inadvertent nuclear use. However, if rushed and implemented in a manner that does not consider the wider implications that ML could have for nuclear security, integration risks raising the risk of use in dangerously subtle manner that may go unnoticed until catastrophe strikes. Therefore, it is crucial to address this problem now while it is still in its malleable infancy.

8. References

Aksenov, Pavel. ‘Stanislav Petrov: The Man Who May Have Saved the World’. *BBC News*, 26 September 2013, sec. Europe. https://www.bbc.com/news/world-europe-24280831.

Amodei, Dario, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. ‘Concrete Problems in AI Safety’. arXiv, 25 July 2016. https://doi.org/10.48550/arXiv.1606.06565.

Arms Control Association. ‘U.S. Nuclear Modernization Programs.’ https://www.armscontrol.org/factsheets/USNuclearModernization

Barrett, Anthony. ‘False Alarms, True Dangers?: Current and Future Risks of Inadvertent U.S.-Russian Nuclear War’. Santa Monica, CA: RAND Corporation, 2016. https://www.rand.org/pubs/perspectives/PE191.html.

Barrett, Anthony M., Seth D. Baum, and Kelly Hostetler. ‘Analyzing and Reducing the Risks of Inadvertent Nuclear War Between the United States and Russia’. *Science & Global Security* 21, no. 2 (2013): 106.

Barnett, Jackson. ‘AI Needs Humans “on the Loop” Not “in the Loop” for Nuke Detection, General Says’. FedScoop, 14 February 2020. https://www.fedscoop.com/ai-should-have-human-on-the-loop-not-in-the-loop-when-it-comes-to-nuke-detection-general-says/.

Bender, Bryan. ‘“A Big Blow”: Washington’s Arms Controllers Brace for Loss of Their Biggest Backer’. POLITICO. Accessed 2 October 2022. https://www.politico.com/news/2021/07/19/washington-arms-controllers-nuclear-weapons-500126.

Bleicher, Ariel. ‘Demystifying the Black Box That Is AI’. Scientific American, 2017. https://www.scientificamerican.com/article/demystifying-the-black-box-that-is-ai/.

Richard K. Betts, “Realism Is an Attitude, Not a Doctrine,” The National Interest, August 2015, https://nationalinterest.org/feature/realism-attitude-not-doctrine-13659.

Boulanin, Vincent, Lora Saalman, Petr Topychkanov, Fei Su, and Moa Peldán Carlsson. ‘Artificial Intelligence, Strategic Stability and Nuclear Risk’. SIPRI, June 2020. https://www.sipri.org/publications/2020/other-publications/artificial-intelligence-strategic-stability-and-nuclear-risk.

Dario, Jack. ‘Faulty Reward Functions in the Wild’. OpenAI, 22 December 2016. https://openai.com/blog/faulty-reward-functions/.

Dastin, Jeffrey. ‘Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women’. *Reuters*, 10 October 2018, sec. Retail. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G.

Davis, Nicola. ‘Soviet Submarine Officer Who Averted Nuclear War Honoured with Prize’. *The Guardian*, 27 October 2017, sec. Science. https://www.theguardian.com/science/2017/oct/27/vasili-arkhipov-soviet-submarine-captain-who-averted-nuclear-war-awarded-future-of-life-prize.

Dear, Keith. ‘Artificial Intelligence and Decision-Making’. *The RUSI Journal* 164, no. 5–6 (19 September 2019): 18–25. https://doi.org/10.1080/03071847.2019.1693801.

Drum, Kevin. ‘Tech World: Welcome to the Digital Revolution’. *Foreign Affairs*, July/August 2018. https://www.foreignaffairs.com/articles/world/2018-06-14/tech-world.

Flournoy, Michèle A, Avril Haines, and Gabrielle Chefitz. ‘Building Trust through Testing: Adapting DOD’s Test & Evaluation, Validation & Verification (TEVV) Enterprise for Machine Learning Systems, Including Deep Learning Systems’. WestExec Advisors, 2020.

Ganguli, Deep, Danny Hernandez, Liane Lovitt, Nova DasSarma, Tom Henighan, Andy Jones, Nicholas Joseph, et al. ‘Predictability and Surprise in Large Generative Models’. arXiv, 15 February 2022. http://arxiv.org/abs/2202.07785.

GOV.UK. ‘Defence Artificial Intelligence Strategy’, 2022. https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy.

Harper, Jon. ‘Nuclear Command, Control, Comms Under Scrutiny’. *Center For Strategiv Deterrence Studies: News and Abalysis*, no. 1357 (2019): 7–8.

Heaven, Douglas. ‘Why Deep-Learning AIs Are so Easy to Fool’. *Nature* 574, no. 7777 (9 October 2019): 163–66. https://doi.org/10.1038/d41586-019-03013-5.

Horowitz, Michael, and Paul Scharre. ‘AI and International Stability: Risks and Confidence-Building Measures’. Technology & National Security. Center for a New American Security, 2021. https://www.cnas.org/publications/reports/ai-and-international-stability-risks-and-confidence-building-measures.

Horowitz, Michael C., Lauren Kahn, Christian Ruhl, Missy Cummings, Erik Lin-Greenberg, Paul Scharre, and Rebecca Slayton. ‘Policy Roundtable: Artificial Intelligence and International Security’. Texas National Security Review, 2020. https://tnsr.org/roundtable/policy-roundtable-artificial-intelligence-and-international-security/.

Horowitz, Michael C. ‘Trust, Confidence, and Organizational Decisions about AI Adoption: The Impact for US Defen’. Minerva Research Initiative. Accessed 2 October 2022. https://minerva.defense.gov/Owl-In-the-Olive-Tree/Owl_View/Article/2328498/trust-confidence-and-organizational-decisions-about-ai-adoption-the-impact-for/https%3A%2F%2Fminerva.defense.gov%2FOwl-In-the-Olive-Tree%2FOwl_View%2FArticle%2F2328498%2Ftrust-confidence-and-organizational-decisions-about-ai-adoption-the-impact-for%2F.

Horowitz, Michael C. ‘When Speed Kills: Lethal Autonomous Weapon Systems, Deterrence and Stability’. *Journal of Strategic Studies* 42, no. 6 (19 September 2019): 764–88. https://doi.org/10.1080/01402390.2019.1621174.

Hruby, Jill, and Nina Miller. ‘Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapon Systems’. NTI, 2019.

https://www.nti.org/analysis/articles/assessing-and-managing-the-benefits-and-risks-of-artificial-intelligence-in-nuclear-weapon-systems/.

Hua, Shin-Shin. 'Machine Learning Weapons and International Humanitarian Law: Rethinking Meaningful Human Control'. *Georgetown Journal of International Law* 51, no. 1 (2020 2019): 117–46.

'Human-Machine Teaming (JCN 1/18)'. Joint Concept Note, 2018. https://www.gov.uk/government/publications/human-machine-teaming-jcn-118.

IBM. 'What Is Machine Learning?', 2020. https://www.ibm.com/cloud/learn/machine-learning.

Johnson, James. 'Delegating Strategic Decision-Making to Machines: Dr. Strangelove Redux?' *Journal of Strategic Studies* 45, no. 3 (16 April 2022): 439–77. https://doi.org/10.1080/01402390.2020.1759038.

Johnson, James. 'Rethinking Nuclear Deterrence in the Age of Artificial Intelligence'. Modern War Institute, 28 January 2021. https://mwi.usma.edu/rethinking-nuclear-deterrence-in-the-age-of-artificial-intelligence/.

Jones, Nate, and Peter J. Scoblic. 'The Week the World Almost Ended'. *Slate*, 7 June 2017. https://slate.com/news-and-politics/2017/06/able-archer-almost-started-a-nuclear-war-with-russia-in-1983.html.

Kaji, Mina, Amanda Maile, and Gio Benitez. '1 Year after the Ethiopian Air Flight 302 Crash, Questions Remain as to When Boeing's 737 Max Will Fly Again'. ABC News. Accessed 2 October 2022. https://abcnews.go.com/Politics/year-ethiopian-air-flight-302-crash-questions-remain/story?id=69469775.

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

Lewis, Patricia, Benoît Pelopidas, and Heather Williams. 'Too Close for Comfort: Cases of Near Nuclear Use and Options for Policy'. Chatham House, 28 April 2014. https://www.chathamhouse.org/2014/04/too-close-comfort-cases-near-nuclear-use-and-options-policy.

Maas, Matthijs M. and Matteucci, Kayla and Cooke, Di, Military Artificial Intelligence as Contributor to Global Catastrophic Risk (May 22, 2022). Cambridge Conference on Catastrophic Risk 2020, Available at SSRN: https://ssrn.com/abstract=4115010 or http://dx.doi.org/10.2139/ssrn.4115010

Maas, Matthijs M. 'How Viable Is International Arms Control for Military Artificial Intelligence? Three Lessons from Nuclear Weapons'. *Contemporary Security Policy* 40, no. 3 (3 July 2019): 285–311. https://doi.org/10.1080/13523260.2019.1576464.

Miller, Philip Reiner, Alexa Wehsener, and M. Nina. ‘When Machine Learning Comes to Nuclear Communication Systems’. C4ISRNet, 1 May 2020. https://www.c4isrnet.com/thought-leadership/2020/04/30/when-machine-learning-comes-to-nuclear-communication-systems/.

Podvig, Pavel. ‘Reducing the Risk of an Accidental Launch’. *Science & Global Security* 14, no. 2–3 (1 December 2006): 75–115. https://doi.org/10.1080/08929880600992990.

Reiner, Philip, and Alexa Wehsener. ‘The Real Value of Artificial Intelligence in Nuclear Command and Control’. War on the Rocks, 4 November 2019. https://warontherocks.com/2019/11/the-real-value-of-artificial-intelligence-in-nuclear-command-and-control/.

Ruhl, Christian. ‘Autonomous Weapon Systems & Military AI: Cause Area Investigation’. Founders Pledge, May 2022. https://docs.google.com/document/d/1hZfuxAp4yhsjmYXBNHtHZxeYabVK1kwalTl711cyW0A/edit?usp=sharing&usp=embed_facebook.

Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War*. W. W. Norton & Company, 2019.

Sechser, Todd S., Neil Narang, and Caitlin Talmadge. ‘Emerging Technologies and Strategic Stability in Peacetime, Crisis, and War’. *Journal of Strategic Studies* 42, no. 6 (19 September 2019): 727–35. https://doi.org/10.1080/01402390.2019.1626725.

Stanislav Petrov, interviewed in Vasilyev, Yuri (2004), ‘On the Brink’, The Moscow News, 29 May, http://www.brightstarsound.com/world_hero/the_moscow_news.html.

Stoutland, Page O. “Artificial Intelligence and the Modernization of US Nuclear Forces.” Edited by Vincent Boulanin. *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk: Volume I Euro-Atlantic Perspectives*. Stockholm International Peace Research Institute, 2019. http://www.jstor.org/stable/resrep24525.13.

Tertrais, Bruno. ‘“On The Brink”—Really? Revisiting Nuclear Close Calls Since 1945’. *The Washington Quarterly* 40, no. 2 (3 April 2017): 51–66. https://doi.org/10.1080/0163660X.2017.1328922.

U.S. Department of Defence. ‘National Defence Strategy, 2022. ’https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.PDF

Wong, Yuna Huh, John Yurchak, Robert W. Button, Aaron B. Frank, Burgess Laird, Osonde A. Osoba, Randall Steeb, Benjamin N. Harris, and Sebastian Joon Bae. ‘Deterrence in the Age of Thinking Machines’. RAND Corporation, 27 January 2020. https://www.rand.org/pubs/research_reports/RR2797.html.