



# Artificial Intelligence and Nuclear Weapons: A Commonsense Approach to Understanding Costs and Benefits

Herbert Lin



Artificial intelligence (AI), particularly machine learning (ML), has transformed computing, offering potential benefits in the nuclear enterprise, which encompasses weapons, delivery systems, platforms, and command and control infrastructure. While AI can enhance efficiency in areas like predictive maintenance and operational planning, its integration into the nuclear enterprise poses significant risks, some of which are inherent in the nature of ML. Five principles should guide AI's responsible application in a nuclear weapons context: maintaining meaningful human control in nuclear decision-making processes; evaluating AI risks within a nation's broader nuclear posture; recognizing the challenges of verifying international agreements on AI restrictions; managing risks through self-imposed limitations; and leveraging AI to enhance human oversight. While AI offers opportunities to improve nuclear surety and operational efficiency in areas like planning and predictive maintenance, its deployment must prioritize minimizing catastrophic risks and preserving human judgment in critical decision-making processes.

**A**n internet search conducted on February 26, 2025, using the Perplexity chatbot, for published works since 2022 with the terms “artificial intelligence” and “nuclear weapons,” turned up several hundred hits. Excluding articles written by or describing sentiments of the US military and disregarding applications to arms control and verification, the prevailing sentiment in these hits regarding the application of artificial intelligence (AI) to nuclear weapons systems (such as operations, strategy, doctrine, and command and control) appears to be predominantly negative. This skepticism stems from concerns about detrimental effects on strategic stability and increased risks of escalation, especially of the accidental or inadvertent kind.

These sentiments are not unfounded. But unfortunately they only capture a small portion of what analysts should be talking about regarding the use of AI in a nuclear weapons context. In what follows below, I offer some perspective on AI, on nuclear weapons, and on the broader operational environment within which both AI and nuclear weapons must be understood.

This article begins by reviewing some fundamental characteristics of AI (primarily the machine learning variety) and nuclear weapons (that is, the weapons themselves and entire nuclear enterprise surrounding them), and then outlining scenarios in which AI could increase nuclear risk. This article then proposes five

core principles that may be useful as policymakers think through how to incorporate AI into their nuclear enterprises. Briefly, these principles cover five essential points:

1. Humans remain the most essential element of nuclear command and control. While some nations have committed to human control over nuclear weapons, phrases like “meaningful human control” or “appropriate levels of human judgment” better focus attention on the desired human role.
2. AI's risks and benefits in nuclear command and control cannot be assessed apart from a nation's nuclear posture, which includes force structure, arrangements and infrastructure for nuclear command and control, doctrine, and strategic priorities.
3. International agreements to limit the use of AI in a nuclear weapons context are unlikely to be achieved because of verification challenges, which would remain daunting even with highly intrusive inspections.
4. Nations can mitigate risks from their own AI use in nuclear contexts. Lower-risk applications have minimal stakes and consequences, are commercially pursued, allow time for human review, have clear evaluation criteria, can be isolated if they fail, and have mitigation mechanisms.





5. AI could enhance human control over nuclear weapons in appropriate contexts. Controlling access to nuclear weapons and related functions remains a high security priority for nuclear nations.

## Background and Fundamentals

### On AI

Artificial intelligence (AI) was originally defined in 1955 by John McCarthy, then a professor at Stanford University, as “the science and engineering of making intelligent machines.”<sup>1</sup> AI has evolved through a number of intellectual paradigms since then, the most recent of which is machine learning.

Machine learning (ML) is the art, science, and engineering of enabling computers to perform tasks without explicit instructions, often by generalizing (or “learning”) from patterns in data.<sup>2</sup> This includes so-called “deep learning,” which is inspired by how the human brain is believed to learn. ML seeks to model and understand complex relationships within data and requires large amounts of data from which it can learn. This data can take various forms, including text, images, videos, sensor readings, and more.

A machine learning system (often referred to as a model) operates in two main phases: training and inference. During training, the model learns to identify relationships or structures in its data by adjusting internal numerical values called parameters. These parameters capture patterns that allow the model to generalize to similar data. Once trained, the model performs inference by applying its learned parameters to new inputs to make predictions or derive insights. This process involves using what the model has learned during training to process unseen data and generate meaningful outputs.

Machine learning has enabled computers to perform tasks that were difficult or impossible to perform before and has been described, fairly and correctly, as a revolution in artificial intelligence. Nevertheless, it is essential to keep five points in mind.

First, machine learning is still a way of programming computers. It does free humans from the need to code

instructions explicitly to solve a particular task, but “learning from data” is just another way of programming the computers on which the ML systems run.<sup>3</sup> With respect to their fundamental architecture, those computers are not significantly different in principle than the computers that have dominated the universe of digital technology since World War II.

Second, the internal operations of advanced ML systems are in general incomprehensible to human beings.<sup>4</sup> Because conventional programming does involve the explicit coding of instructions, it is possible at least in principle to follow the computer’s path from input to output and thereby to obtain some understanding of what has happened. But as an advanced ML system learns, a look inside the computer would reveal a vast number of parameters changing—all of the knowledge in the ML system is contained in this set of numbers, which establish the “rules” that drive how the system makes inferences. And in general, a human examining these numbers cannot generate comprehensible representations of those rules. Thus, it is difficult or impossible to understand in a meaningful way how the system makes an inference given a certain input or query.

Third, machine learning is at its root statistics. Machine learning derives its power from the power of statistical reasoning, and machine learning does not provide capabilities that statistics cannot provide.<sup>5</sup> In particular, the adage from Statistics 101 is applicable to machine learning—“correlation is not causation.” This point is fundamentally important to a concern about the explainability of statistical (or ML) conclusions: Statistics or ML may be able to provide a statistical or correlative explanation for a particular conclusion, but neither will ever provide mechanistic or causal explanations without the use of other techniques.<sup>6</sup> In other words, they will never be able to explain *why* a conclusion has been reached. For many purposes, statistical explanations are entirely adequate; for others, causal explanations are necessary.

Fourth, the adage “garbage in, garbage out”—coined in the earliest days of computing—still applies to machine learning.<sup>7</sup> Give any system bad inputs, and bad outputs will often emerge. But the critical

---

1 Teneo AI, “Homage to John McCarthy, the Father of Artificial Intelligence (AI),” <https://www.teneo.ai/blog/homage-to-john-mccarthy-the-father-of-artificial-intelligence-ai>.

2 Sara Brown, “Machine Learning, Explained,” MIT Sloan School of Management, April 21, 2021, <https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained>.

3 Brown, “Machine Learning, Explained.”

4 Will Knight, “The Dark Secret at the Heart of AI,” *MIT Technology Review*, April 11, 2017, <https://www.technologyreview.com/2017/04/11/5113/the-dark-secret-at-the-heart-of-ai/>.

5 Danilo Bzdok, Naomi Altman, and Martin Krzywinski, “Statistics Versus Machines Learning,” *Nature Methods* 15 (April 2018): 233–34, <https://doi.org/10.1038/nmeth.4642>.

6 Kaleb Leetaru, “A Reminder that Machine Learning Is About Correlations Not Causation,” *Forbes*, January 15, 2019, <https://www.forbes.com/sites/kalebleetaru/2019/01/15/a-reminder-that-machine-learning-is-about-correlations-not-causation/>.

7 DevX, “Garbage In, Garbage Out: Definition, Examples,” <https://www.devx.com/terms/garbage-in-garbage-out/>.

caveat is that you can recognize the bad outputs only if you have some way of testing any given output against some ground truth. An ML system trained on biased data, for example, will tend to generate results that are biased.<sup>8</sup> Large language models such as ChatGPT have been trained on a large fraction of the text readily available through the internet; they therefore incorporate all of the biases and limitations that are inherent in this corpus.<sup>9</sup>

Fifth, machine learning, and AI more generally, does not possess the power to alter the nature of reality. ML cannot generate new facts from nothing; rather, it derives insights by analyzing and synthesizing existing data. While AI can uncover patterns or relationships that may not be immediately apparent to humans, this process is not equivalent to creating something entirely new—the results are a reorganization or reinterpretation of what is already known. Additionally, AI is no better equipped than humans to address the elusive “unknown unknowns”—those mysteries that lie beyond the boundaries of current knowledge. Unlike humans, who can sometimes intuit or speculate about such unknowns, AI relies entirely on data and explicit frameworks. Finally, despite its remarkable capabilities, AI remains constrained by the laws of physics, and does not enable humans to transcend or violate these universal principles.

### On Nuclear Weapons and the Nuclear Enterprise

A nation’s nuclear enterprise has many elements, including the nuclear explosive devices (that is, the nuclear weapons themselves); the vehicles that deliver these weapons to their targets (for example, ballistic missiles, cruise missiles, and airplanes); the platforms that support these vehicles (for example, submarines, air bases, and missile silos); and an infrastructure for command, control, and communications that ties all of these elements together.<sup>10</sup> This enterprise is enormously complex, and to manage and interpret the immense information flows needed to operate this enterprise safely and reliably, computers are used everywhere.

To be sure, most of these computers have been programmed conventionally—that is, by a human programmer (or much more likely, a programming team). But since AI is just another way of programming a computer, AI is potentially usable in every place that a computer is used in the nuclear enterprise.

Human beings also play many roles in the nuclear enterprise, ranging from critical to mundane and pedestrian. Consider nuclear command and control (NC2). At the critical end, the most critical role humans play in NC2 is in the decision to “push the button”—that is, to order the use of nuclear weapons. In no known NC2 arrangement does a human literally push a button to send an electrical or radio signal that directly ignites the engines of missiles carrying nuclear weapons.<sup>11</sup> In all known cases, the national leader (or leaders) makes the decision to order the use of nuclear weapons, and that decision is transmitted down a human chain of command that is supposed to ensure that a properly authorized launch order gets to the weapons launchers.<sup>12</sup> This chain of command is also supposed to prevent a launch from happening in the absence of a properly authorized order.<sup>13</sup>

At the mundane end, human beings in all NC2 institutions produce papers and reports about their nuclear weapons, their delivery vehicles, and so on. These people almost certainly produce such papers with the use of word-processing software with features such as “auto-complete,” in which the typing of a few characters leads to a machine-generated suggestion for how to complete the word or phrase. This is one use of AI in NC2, but it is an entirely uncontroversial use.

In between the critical (where the use of AI is absurd on the face of it) and the mundane (where the use of AI is entirely noncontroversial) lie many other possibly helpful uses of AI, though it is not publicly known at this time how any specific nation uses AI for any specific NC2 purpose. One case where AI may well be helpful is predictive maintenance, in which AI provides information used to direct the replacement of parts in delivery vehicles and weapons just before they fail.<sup>14</sup> A second case is in the pre-conflict

8 Bo Cowgill et al., “Biased Programmers? Or Biased Data? A Field Experiment in Optimizing AI Ethics,” arXiv, December 4, 2020, <https://doi.org/10.48550/arXiv.2012.02394>; ML Concepts, “Fairness: Types of Bias,” Google for Developers, <https://developers.google.com/machine-learning/crash-course/fairness/types-of-bias>.

9 OpenAI, “How ChatGPT and Our Foundation Models Are Developed,” <https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-foundation-models-are-developed>.

10 Defense Threat Reduction Agency, “Nuclear Enterprise Directorate,” US Department of Defense, <https://www.dtra.mil/About/Mission/Nuclear-Enterprise/>.

11 Russell Lerman, “The ‘Nuclear Button’ Explained: For Starters, There’s No Button,” *The New York Times*, January 3, 2018, <https://www.nytimes.com/2018/01/03/world/asia/nuclear-button-trump-north-korea.html>.

12 US Department of Defense, “Nuclear Posture Review,” in *2022 National Defense Strategy of the United States of America*, October 27, 2022, <https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.pdf>.

13 US Department of Defense, “Nuclear Posture Review.”

14 Billy Mitchell, “Air Force Selects AI-Enabled Predictive Maintenance Program as System of Record,” *DefenseScoop*, Scoop News Group, May 10, 2023, <https://defensescoop.com/2023/05/10/air-force-selects-ai-enabled-predictive-maintenance-program-as-system-of-record/>.



development of operational plans for using nuclear weapons to maximize their likelihood of success, such as developing the best flight paths for a cruise missile en route to its target.<sup>15</sup> A third possibility is the processing and integration of data from sensors in a particular region that could signal an incoming airborne attack.<sup>16</sup> Still a fourth case is providing options to commanders regarding possible courses of action in the lead-up to and during a conventional or a nuclear war.<sup>17</sup> These are just a few of the possible cases that fall in between the two extremes where the risks of using AI may—or may not—be outweighed by the benefits.

### **Scenarios Where AI Could Increase Nuclear Risk**

No serious proposal has emerged that a computer, AI-enabled or otherwise, should be entrusted with the actual decision to launch nuclear weapons. Thus, the idea of “giving the nuclear launch codes to ChatGPT”—while often imagined—is a misleading caricature of the issue. Nevertheless, the use of AI in certain contexts could still result in increased risks, just as there are also ways to use AI to reduce nuclear risks.

For example, submarines carrying nuclear missiles are generally regarded as survivable in the aftermath of an adversary's first nuclear strike, and therefore capable of inflicting a significant nuclear retaliatory blow if such a strike did occur. Hence, nuclear-armed submarines are usually regarded as enhancing nuclear deterrence and promoting nuclear stability through assured retaliation.<sup>18</sup> But their survivability depends on being able to hide very small vessels in very large oceans. If the adversary were capable of using AI coupled with advanced sensors to localize

submarines, their survivability might be reduced significantly,<sup>19</sup> a consequence that could increase the likelihood of the adversary attempting to conduct a disarming first strike. In this case, the risk arises from the possibility that nuclear-armed adversaries use AI to gain strategic advantages over others.

## **No serious proposal has emerged that a computer, AI-enabled or otherwise, should be entrusted with the actual decision to launch nuclear weapons.**

A second example is that an AI given the task of integrating all early-warning sensor information to determine if an attack is underway would almost certainly have interpreted the November 1979 US NC2 incident as a real attack. In that incident, a training tape was inadvertently and inexplicably loaded into an operational NC2 computer in the United States, which resulted in a false alarm.<sup>20</sup> By definition, a realistic training tape would feed into the NC2 system all of the sensor data that would be expected in a real attack. This procedure would require a decision-making entity that could look outside the sensor data to realize that the warning from a training tape was a false one. Here, one's own use of AI in an NC2 system could increase risk by providing a false sense of confidence to national leaders.

A third possible scenario is based on the idea that nuclear conflict is much more likely to occur as the result of an escalated conventional conflict than to occur as the first step in a conflict,<sup>21</sup> though the latter possibility cannot be dismissed. Thus, imagine that in the future an AI is given the authority to launch nonnuclear attacks on satellites that provide tracking

15 Mary Chestnut et al., “Artificial Intelligence in Nuclear Operations,” Center for Naval Analyses, April 17, 2023, <https://www.cna.org/reports/2023/04/ai-in-nuclear-operations>.

16 David Vergun, “DOD Will Deploy AI-Enabled Detection System to Monitor DC Airspace,” DOD News, US Department of Defense, August 28, 2023, <https://www.defense.gov/News/News-Stories/Article/Article/3507329/dod-will-deploy-ai-enabled-detection-system-to-monitor-dc-airspace/>.

17 “AI Will Transform the Character of Warfare,” The Economist, June 20, 2024, <https://www.economist.com/leaders/2024/06/20/war-and-ai>; International Committee of the Red Cross and Geneva Academy, *Artificial Intelligence and Related Technologies in Military Decision-Making on the Use of Force in Armed Conflicts*, ICRC, Geneva, March 2024, <https://www.geneva-academy.ch/joomlatools-files/docman-files/Artificial%20Intelligence%20And%20Related%20Technologies%20In%20Military%20Decision-Making.pdf>; Ruben Stewart and Georgia Hinds, “Algorithms of War: The Use of Artificial Intelligence in Decision Making in Armed Conflict,” International Committee of the Red Cross, October 24, 2023, <https://blogs.icrc.org/law-and-policy/2023/10/24/algorithms-of-war-use-of-artificial-intelligence-decision-making-armed-conflict/>.

18 US Department of Defense, “America's Nuclear Triad,” <https://www.defense.gov/Multimedia/Experience/Americas-Nuclear-Triad/>.

19 Andrew Reddie and Bethany Goldblum, “Unmanned Underwater Vehicle (UUV) Systems for Submarine Detection,” On the Radar, Center for Strategic and International Studies, July 29, 2019, <https://ontheradar.csis.org/issue-briefs/unmanned-underwater-vehicle-uuv-systems-for-submarine-detection-a-technology-primer/>.

20 Union of Concerned Scientists, “Close Calls with Nuclear Weapons,” January 15, 2015, 4, <https://www.ucsusa.org/resources/close-calls-nuclear-weapons>.

21 Paul Bracken, “The New Logic of Armageddon,” in *The Second Nuclear Age: Strategy, Danger, and the New Power Politics* (St. Martin's Griffin, 2013); Keir A. Lieber and Daryl G. Press, “The Return of Nuclear Escalation,” *Foreign Affairs* 102, no. 6 (December 2023): 45–55, <https://www.foreignaffairs.com/united-states/return-nuclear-escalation>; John K. Warden, “Limited Nuclear War: The 21st Century Challenge for the United States,” in *Livermore Papers on Global Security*, Center for Global Security Research, Lawrence Livermore National Laboratories, July 2018, [https://cgsrc.llnl.gov/sites/cgsrc/files/2024-08/CGSR\\_LP4-FINAL.pdf](https://cgsrc.llnl.gov/sites/cgsrc/files/2024-08/CGSR_LP4-FINAL.pdf).

and targeting data for earthbound conventional weapons. The side losing the satellites could well perceive a strategic threat to its satellites, especially if those satellites were used to support both conventional and nuclear forces.<sup>22</sup> In this case, AI is not integrated into NC2 at all but still adds to nuclear risk.

## **Principles for Thinking About AI and Nuclear Weapons**

The space that lies in between the extremes of “absurd” and “entirely uncontroversial” contains the majority of seriously proposed AI applications, and it is these applications that most demand thoughtful consideration. To help decision-makers think productively about where and where not to use AI in their nuclear enterprises, the following five principles may be useful.

### **Principle 1: Humans continue to be the most important element in nuclear command and control.**

A number of nuclear-armed nations have made commitments to maintain human control over nuclear weapons. The United States has said that “[i]n all cases, [it] will maintain a human ‘in the loop’ for all actions critical to informing and executing decisions by the President to initiate and terminate nuclear weapon employment.”<sup>23</sup> The United Kingdom has said it will “ensure that—regardless of any use of AI in our strategic systems—human political control of our nuclear weapons is maintained at all times.”<sup>24</sup> France has endorsed a statement that it will “maintain human control and involvement for all actions critical to informing and executing sovereign decisions concerning nuclear weapons employment.”<sup>25</sup> In November 2024, the People’s Republic of China (PRC) noted that Presidents Biden and Xi had, together, “stressed the

need to maintain human control, instead of AI, over the decision to use nuclear weapons.”<sup>26</sup>

These statements are useful and helpful, and yet they do not go far enough. The reason is that they do not capture what the role of humans *should* be and what it means to be “in the loop” or “in political control.” For example, consider an entirely automated, AI-enabled, and human-free command and control chain that begins with processing sensor data from satellites and radars and ends with a high-confidence recommendation to the national leader. This construct technically satisfies requirements for a human being in the loop or in political control, but is hardly reassuring because the human being has no information on which to base a decision other than that provided by the AI-enabled system. Even worse, humans often exhibit undue deference to computer-generated conclusions in a phenomenon known as automation bias.<sup>27</sup> These factors suggest that a national leader would have difficulty evaluating the reliability of an AI’s recommendation.

Phrases such as “meaningful human control” or “appropriate levels of human judgment” have been used to address such issues and do provide more reassurance.<sup>28</sup> Even in the absence of hard and fast definitions for “meaningful” and “appropriate,” the use of such terms at least focuses attention on precisely what the roles of humans should be. In addition, human beings, at least today, are needed to bring human virtues to decision-making processes involving the use of nuclear weapons. In such decision-making, wisdom, compassion, and mercy have roles to play—and people without genuine fears of nuclear war or horror at its consequences are probably not the best people to be making decisions about using nuclear weapons.<sup>29</sup> Whether only humans—as opposed to computers—can truly demonstrate traits such as wisdom, compassion, and mercy remains

22 Herb Lin, “Cyber Risks in Selected Nuclear Scenarios,” in *Cyber Threats and Nuclear Weapons* (Stanford University Press, 2021), 108–10.

23 Department of Defense, “2022 Nuclear Posture Review,” in *2022 National Defense Strategy of the United States of America*, October 27, 2022, 13, <https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.pdf>.

24 Ministry of Defence, “Defence Artificial Intelligence Strategy,” GOV.UK, June 15, 2022, <https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy>.

25 Subcommittee for the Treaty on the Non-Proliferation of Nuclear Weapons, “Principles and Responsible Practices for Nuclear Weapon States: Working Paper Submitted by France, the United Kingdom of Great Britain and Northern Ireland and the United States of America,” 2020 Review Conference of the Parties to the Treaty on the Non-Proliferation of Nuclear Weapons, July 29, 2022, NPT/CONF.20/WP.70, <https://documents.un.org/doc/undoc/gen/n22/446/53/pdf/n2244653.pdf>.

26 “Xi, Biden, Confirm Need to Enhance International Cooperation,” Xinhua, 14 November 2024, [https://english.www.gov.cn/news/202411/17/content\\_WS67394d2ac6d0868f4e8ed13c.html](https://english.www.gov.cn/news/202411/17/content_WS67394d2ac6d0868f4e8ed13c.html).

27 Deloitte, “Automation Bias: What Happens when Trust Goes Too Far?,” January 20, 2023, <https://www.deloitte.com/uk/en/services/consulting/research/automation-bias.html>; Bryce Hoffman, “Automation Bias: What It Is and How to Overcome It,” *Forbes*, March 10, 2024, <https://www.forbes.com/sites/brycehoffman/2024/03/10/automation-bias-what-it-is-and-how-to-overcome-it/>.

28 On meaningful control, see, for example, Heather M. Roff and Richard Moyes, “Meaningful Human Control, Artificial Intelligence and Autonomous Weapons,” briefing paper for meeting of experts on Lethal Autonomous Weapons Systems (LAWS), UN Convention on Certain Conventional Weapons, Geneva, April 2016, <https://article36.org/wp-content/uploads/2016/04/MHC-AI-and-AWS-FINAL.pdf>. On appropriate levels of human judgment, see DOD 3000.09 at Department of Defense, “DOD Directive 3000.09,” January 25, 2003, <https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf>.

29 Samuel Zilincik, “The Role of Emotions in Military Strategy,” *Texas National Security Review* 5, no. 2 (2022): 11–25, <https://tnsr.org/2022/01/the-role-of-emotions-in-military-strategy/>.



uncertain in the long term. However, since little progress is currently being made in this area, it seems likely to remain true for the foreseeable future.

## **The nature and extent of appropriate human involvement in such functions is subject to considerable debate and discussion.**

Lastly, the actual scope of commitments to keep humans “in the loop” is not entirely clear. While “the loop” clearly includes the ultimate decision of consequence—namely deciding to launch nuclear weapons—it is not clear whether the definition also includes the assessment of early-warning information that may be used in nuclear decision-making or in making recommendations for targeting or military operations using nuclear weapons. The nature and extent of appropriate human involvement in such functions is subject to considerable debate and discussion.

**Principle 2: The risks and benefits of using AI in nuclear command and control should not be considered in isolation from other aspects of a nation’s nuclear posture.**

Discussions of AI in NC2 are often framed without attention to the nuclear forces that are meant to be controlled. But the posture of these nuclear forces—which includes force structure, arrangements and infrastructure for command and control, doctrine, and strategic priorities—is an essential element of understanding the risks of AI in NC2.<sup>30</sup> For example, concerns have been raised about launch-on-warning scenarios, in which an adversary attacks in an attempt to disarm a nation. The attacked nation launches its missiles “on warning” to prevent their destruction when early-warning sensors indicate that an adversary attack has been initiated but before the adversary’s weapons have detonated. If, however, the indicators of attack are incorrect and in fact no attack is underway, a launch on warning could *initiate* nuclear war by mistake.<sup>31</sup>

AI could be used to speed up the process of resolving conflicting signals. Conversely, AI enhancements could possibly lead to a greater likelihood of erroneous attack indicators under certain circumstances. Unilaterally getting rid of forces that would be vulnerable to a first strike by an adversary—such as silo-based Intercontinental Ballistic Missiles (ICBMs)—would obviously reduce the risk from erroneous attack indicators, because those missiles would not need to be launched on warning of attack.<sup>32</sup> Whether or not to eliminate silo-based ICBMs is beyond the scope of this piece, but the point is a clear example of how force structure might affect the risks of using AI in NC2.

**Principle 3: Nations seeking to reduce AI-enabled threats of preemptive destruction of their strategic forces are very unlikely to achieve acceptable international agreements regarding the use of AI in a nuclear weapons context.**

Adversary use of AI may increase the threat of preemptive destruction of a nation’s strategic forces, thereby reducing stability. To reduce these risks, adversaries would need to refrain from deploying AI-enabled capabilities that improve their ability to conduct a successful first strike. Because these concerns are inevitably reciprocal, risk mitigation implies a commitment by *all* nuclear-armed adversaries to refrain from deploying such capabilities. Thus, many commentators have been drawn to the idea that these nations could come to agreements on AI capabilities in much the same way that they have come to arms control agreements regarding nuclear weapons.<sup>33</sup>

The analogy to nuclear arms control is appealing because any given nation would have much to gain from restraint on the part of its adversaries. However, high-confidence verification of agreed-to limitations of AI capabilities would be entirely unlike anything seen in the history of nuclear arms control.

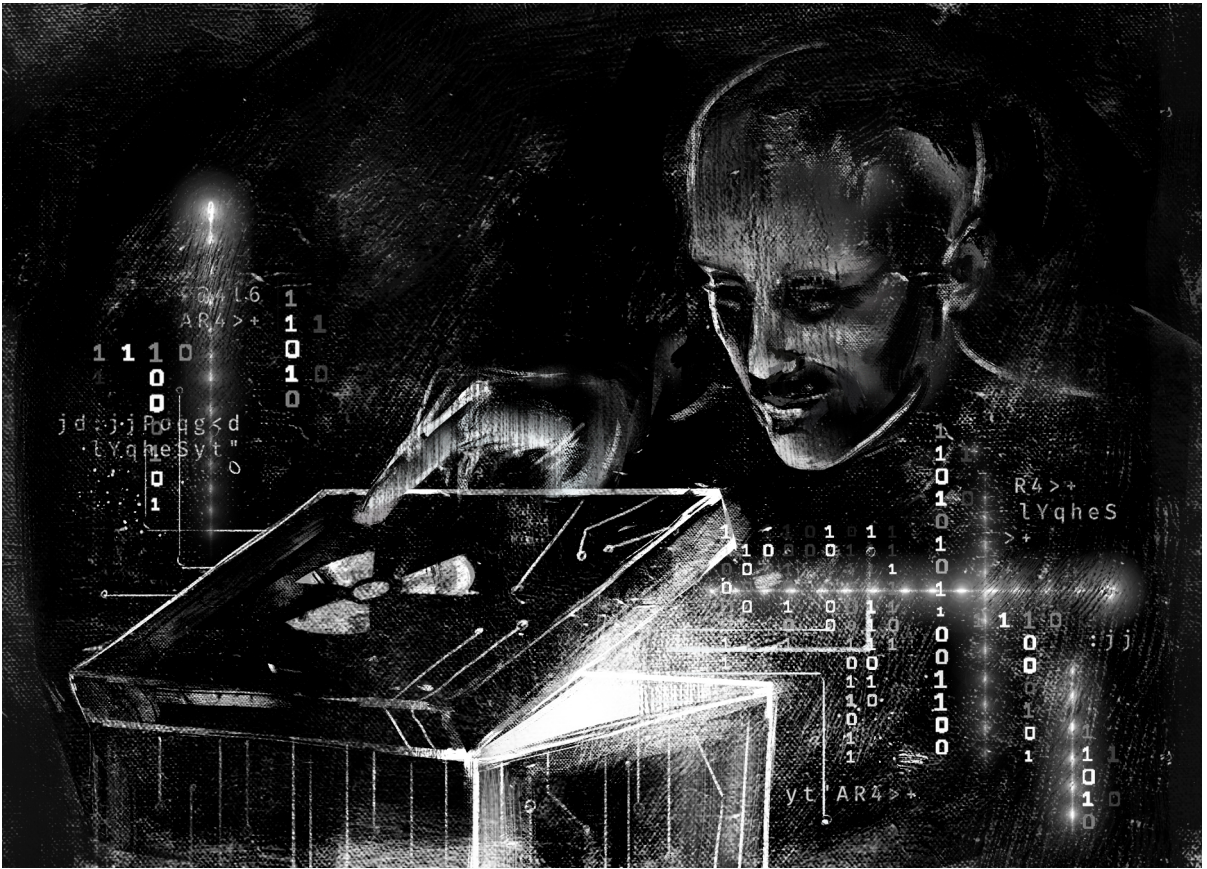
In the history of nuclear arms control, a nation could generally *see* the arms being limited—for example, satellites could view the construction of submarines or missile silos or observe adversary

30 Narang addresses in some detail the relationship between a nation’s nuclear posture and its NC2 arrangements and infrastructure; he is, however, silent on AI per se. See Vipin Narang, *Nuclear Strategy in the Modern Era: Regional Powers and International Conflict* (Princeton, 2014).

31 For historic examples of computer-enabled, near-miss launch-on-warning incidents, see Union of Concerned Scientists, “Close Calls with Nuclear Weapons.”

32 Garrett Hinck and Pranay Vaddi, “Setting a Course Away from the Intercontinental Ballistic Missile,” *War on the Rocks*, February 16, 2021, <https://warontherocks.com/2021/02/setting-a-course-away-from-the-intercontinental-ballistic-missile/>.

33 See, for example, Haleema Saadia, Tynchtykbek Israilov, Ekaterina Mikhalevich, and Jonas Sandbrink, “AI and Nuclear Decisions: Toward an Arms Control Framework,” *Contemporary Security Policy*, March 2025, 1–28, <https://doi.org/10.1080/13523260.2025.2474869>; Zachary Kallenborn, “Giving an AI Control of Nuclear Weapons: What Could Possibly Go Wrong?,” *Bulletin of the Atomic Scientists*, February 1, 2022, <https://thebulletin.org/2022/02/giving-an-ai-control-of-nuclear-weapons-what-could-possibly-go-wrong/>; António Guterres, “Secretary-General’s Remarks to the Security Council on Artificial Intelligence,” United Nations Secretary-General, July 18, 2023, <https://www.un.org/sg/en/content/sg/speeches/2023-07-18/secretary-generals-remarks-the-security-council-artificial-intelligence>; Michael Depp, “The Next Step in Military AI Multilateralism,” *Lawfare*, March 26, 2024, <https://www.lawfaremedia.org/article/the-next-step-in-military-ai-multilateralism>.



missile tests. By contrast, AI efforts today are primarily driven by civilian concerns and they take place behind closed doors, into which satellites have no visibility.<sup>34</sup> Nor do on-site inspections offer much utility—it is impossible to tell what a computer is doing without viewing the code actually running on that computer, and a determined adversary would find it relatively easy to disguise the purpose of any given computer code. The inevitable conclusion is that any international agreement to limit the deployment of AI capabilities for NC2—let alone for any military purpose—would be fundamentally unverifiable.<sup>35</sup>

Furthermore, the benefits that would accrue to a nation entering into such an agreement would have to outweigh the risk of forgoing capabilities, both civilian and military, that might be quite useful. Coupled with the unverifiability of such an agreement, the inevitable conclusion is that nations should not rely on mutual agreements to forgo AI-enabled capabilities in their nuclear enterprises for reducing threats to nuclear stability.

**Principle 4: Nations can reduce the risks emanating from the use of AI in their own nuclear enterprise.**

In contrast to the risks mentioned in principle 3, a second set of risks originates in the inappropriate deployments of AI for a nation's own purposes. Nations are in a much better position to manage these risks through self-restraint and reasoned judgment regarding certain kinds of AI deployments. What follows below are several criteria that could help policymakers decide if a given deployment of AI in a nuclear weapons context is more or less risky.

A deployment of AI in a nuclear weapons context is less risky if one or more of the following conditions are true:

- *The stakes of the application are relatively low and the consequences of mistakes are minimal.* These low-stakes settings provide valuable opportunities to gain experience and refine AI systems without significant risk. For instance, using AI for predictive maintenance—where errors might result in minor

34 See, for example, Mauricio Baker, "Nuclear Arms Control Verification and Lessons for AI Treaties," arXiv, April 8, 2023, <https://doi.org/10.48550/arXiv.2304.04123>; Matthew Mittlesteadt, "AI Verification: Mechanisms to Ensure AI Arms Control Compliance," Center for Security and Emerging Technology, February 2021, <https://doi.org/10.51593/20190020>.

35 A similar conclusion is reached in Megan Lamberth, "Arms Control for Artificial Intelligence," *Texas National Security Review*, May 1, 2023, <https://tnsr.org/2023/05/arms-control-for-artificial-intelligence/>.



inconveniences—offers a safer testing ground compared to employing AI for recommending critical courses of action, where mistakes could have far-reaching consequences. This measured approach allows for gradual learning and improvement before tackling higher-stakes scenarios.

- *The application in question, or something similar, is being pursued in the commercial world.* In such cases, military efforts can leverage these advancements as a foundation. Commercial AI initiatives often serve as close analogues and offer a wealth of experience and established metrics for success that military projects can build on. While civilian efforts provide useful insights, however, they are not always directly applicable to military-specific needs. The unique demands of defense applications require tailored solutions that go beyond what purely civilian projects can offer, which ensures that military systems are designed with their distinct operational contexts in mind.
- *If the application affords humans adequate time to review the application's output.* If it does, the technology can be applied more effectively and with greater confidence. For instance, using AI for preplanned target selection allows decision-makers to carefully evaluate and validate the system's recommendations before taking action. In contrast, relying on AI for real-time target selection introduces higher risks, as decisions must be made instantly without thorough human oversight. By focusing on scenarios where time permits careful review, AI can be integrated more responsibly and with reduced potential for errors.
- *If the application's output can be evaluated against a clear ground truth.* If so, the reliability and effectiveness of the system can be more easily assessed. This condition requires access to comprehensive, accurate, and timely data to ensure that the AI operates within a well-defined framework. Success and failure must be readily identifiable to guide improvements and maintain trust in the system. However, tasks involving subjective assessments, such as determining intent, are inherently risky and less suited for AI applications. For example, AI is better suited for optimization tasks—where outcomes can be measured objectively—than for recommending courses of action (COA), which often involve complex judgment calls.
- *If AI-enabled functionality can be separated from the rest of the system.* Separability ensures that users and systems can continue to operate

competently even if the AI fails or is unavailable. This separation provides a safeguard and allows critical operations to proceed without being entirely dependent on AI. By designing systems with this flexibility, organizations can mitigate risks and maintain functionality while still benefiting from AI where it adds value. Such an approach ensures resilience and prevents overreliance on technology that may not always perform as expected.

- *If mechanisms exist to mitigate the worst consequences of the application's failure.* If so, the risks associated with deploying that application can be managed more effectively. High-risk conditions must be both recognizable and addressable through independent fixes to prevent catastrophic outcomes. For instance, in scenarios such as AI causing the accidental launch of an ICBM, safeguards must be in place to detect the error and intervene independently of the AI system. These mechanisms would provide an essential layer of protection and ensure that even in failure scenarios, the most severe consequences can be avoided. Note that this description does not mean that an AI system could or should be entrusted with launching ICBMs—only that an AI system for launching ICBMs with a way to disable accidentally launched missiles would pose less risk than such a system without such a safeguard.

These criteria are incommensurate; that is, they are not easily comparable using a common metric of value because they address different dimensions of risk and utility. Each factor contributes uniquely to the overall safety and effectiveness of an AI system but cannot be reduced to a single measure for comparison or trade-off purposes. Thus, rather than trying to measure all criteria on a single scale, policymakers can assess the presence or absence of these criteria collectively and draw conclusions about relative risk. In short, the more of these criteria that are present in an AI application, the higher the likelihood that the application makes sense. Conversely, the fewer that are present, the higher the risks and the less advisable it is to use AI for that application.

**Principle 5: In the appropriate context, AI could be employed to enhance human control over nuclear weapons.**

Because AI can be used anywhere computation is used, AI also presents opportunities to improve and enhance human control over nuclear weapons. For example, controlling access to nuclear weapons and related functions is a high security priority for nuclear nations. One important aspect of access

control is the identification of individuals whose personalities and personal circumstances may make it inadvisable to grant such access to them. A second aspect is the use of various technical means to ensure that only the appropriate individuals actually obtain physical access.

## **Because AI can be used anywhere computation is used, AI also presents opportunities to improve and enhance human control over nuclear weapons.**

An example of the first aspect is the Nuclear Weapons Personnel Reliability Program (PRP) of the US Department of Defense, which seeks to ensure that “only those persons who demonstrate reliability will be certified to perform specified duties associated with US nuclear weapons, nuclear command and control (NC2) systems, material, and equipment, and special nuclear material (SNM).”<sup>36</sup> Under the PRP, these persons are “continuously evaluated for adherence to PRP standards in order to maintain PRP status.” Grounds for disqualifying an individual for PRP certification can include adverse information related to personal conduct; emotional, mental, and personality disorders; financial considerations; criminal conduct; substance or drug misuse and drug incidents; alcohol use disorder and alcohol-related incidents; sexual harassment and assault; security violations; or misuse of information technology systems.<sup>37</sup>

Of particular note is the requirement for *continuous* evaluation, which stands in contrast to periodic evaluation. Continuous evaluation requires evaluation of information *as it is received* by the relevant authorities, and given the large quantities of information that are required for continuous evaluation, an AI-enabled continuous evaluation program

to flag items of concern for further human review could increase the effectiveness and efficiency of the evaluation process.

Granting physical access to PRP-authorized individuals is a second aspect of access control for which AI may have some significant value.<sup>38</sup> Known as authentication, this process is essentially one of pattern recognition—a pattern or patterns provided by an individual attempting to gain access is correlated with a library of patterns of behavior known in advance to be associated with that individual (for example, an individual’s gait while walking).<sup>39</sup> AI excels in pattern recognition and may well be suitable for integration into access control mechanisms for nuclear weapons.

## **Conclusion**

In considering the risks that artificial intelligence poses in relation to nuclear weapons, it is essential to take into consideration the specific instance of AI and the specific nuclear weapons application on which that AI might be focused. In some applications, particular AI-enabled systems may enhance human control. In others, particular AI-enabled systems may pose catastrophic and unacceptable risks.

Perhaps the most important imperative in moving forward with AI in the nuclear enterprise is for policymakers in all nations to develop better understandings about the risks and potential benefits of specific applications of AI in a variety of possible use cases. To this end, the nascent movement toward a common commitment for responsible military use of AI is an important step forward. For instance, various nations are involved in informal Track II discussions among the nuclear powers to explore these topics.<sup>40</sup> Such involvement is to be applauded and encouraged.

At a state-to-state level, the Biden administration had announced that it was seeking support from other countries for the adoption of a Political Dec-

36 Department of Defense, “Nuclear Weapons Personnel Reliability Program,” in Department of Defense Manual: Number 5210.42, January 13, 2015, <https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodm/521042m.pdf>.


37 For more on continuous evaluation and Trusted Workforce 2.0, see Aaron Boyd, “Artificial Intelligence Is Helping Evaluate 1.1 Million Security Clearance Holders,” Nextgov/FCW, April 11, 2019, <https://www.nextgov.com/artificial-intelligence/2019/04/artificial-intelligence-helping-evaluate-11-million-security-clearance-holders/156255/>. For USG progress on AI-enhanced continuous evaluation, see Office of Inspector General, DHS Has Made Progress in Implementing an Enhanced Personnel Vetting Program, US Department of Homeland Security, August 8, 2024, <https://www.oig.dhs.gov/sites/default/files/assets/2024-08/OIG-24-43-Aug24.pdf>.

38 Avi Turgeman, “Machine Learning and Behavioral Biometrics: A Match Made in Heaven,” Forbes, January 18, 2018, <https://www.forbes.com/sites/forbestechcouncil/2018/01/18/machine-learning-and-behavioral-biometrics-a-match-made-in-heaven/>.

39 ASEE, “Behavioral Biometrics Authentication: Use Cases and Benefits,” February 28, 2023, <https://cybersecurity.asee.io/blog/what-is-behavioral-biometrics-authentication/>.

40 Ryan Hass and Colin Kahl, “Laying the Groundwork for US-China AI Dialogue,” Brookings Institution, April 5, 2024, <https://www.brookings.edu/articles/laying-the-groundwork-for-us-china-ai-dialogue/>; Heather Williams, “CSIS European Trilateral Track 2 Nuclear Dialogues: 2023 Consensus Statement,” Center for Strategic and International Studies, April 25, 2024, <https://www.csis.org/analysis/csis-european-trilateral-track-2-nuclear-dialogues-6>. Track II discussions involve nongovernmental experts from the respective countries who meet to develop a shared understanding of what may or may not be acceptable to their governments on specific issues. Insights from these dialogues often inform and sometimes influence official communications between the nations on these topics.

laration on Responsible Military Use of Artificial Intelligence and Autonomy.<sup>41</sup> This declaration asserted that nations “should take appropriate measures to ensure the responsible development, deployment, and use of their military AI capabilities, including those enabling autonomous functions and systems.”<sup>42</sup> The most important aspect of the declaration, however, was its use of words such as “responsible,” “appropriate,” “transparent,” “auditable,” “explicit,” and “informed,” the meanings and understandings of which were left to be determined by the individual states using those words. Whether the Trump administration will continue this approach is not yet known at this writing, though its revocation of the Biden executive order on artificial intelligence is not encouraging in this regard.<sup>43</sup>

Many parties have affirmed support for the Treaty on the Prohibition of Nuclear Weapons and the ultimate objective of a world without nuclear weapons.<sup>44</sup> Whether or not one supports the abolition of nuclear weapons, it should be clear that no nation has an interest in AI that operates to create catastrophic risks, or that runs away from human control. This paper is intended to provide some insights into how to prevent these things from happening.

**Herbert Lin** is a senior research scholar and research fellow at Stanford University, with interests at the intersection of national security and emerging technologies. He is also director of the Stanford Emerging Technology Review (<http://setr.stanford.edu>). Additionally, he is Chief Scientist Emeritus for the Computer Science and Telecommunications Board of the National Academies and serves on the Science and Security Board of the Bulletin of Atomic Scientists. Lin was a member of President Obama’s Commission on Enhancing National Cybersecurity (2016) and the Aspen Commission on Information Disorder (2020). He was also a professional staff member and staff scientist for the House Armed Services Committee, where his portfolio included defense policy and arms control issues. He received his doctorate in physics from MIT.

**Image:** “Nuclear mushroom cloud as a reflection off a robot helmet, profile view,” image generated by Adobe’s Firefly, April 7, 2025.<sup>45</sup>

41 US Department of State, “Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy,” November 27, 2024, <https://www.state.gov/bureau-of-arms-control-deterrence-and-stability/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/>.

42 US Department of State, “Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy,” n.d., <https://2021-2025.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/>.

43 The Biden Executive Order on AI was EO 14110 (“Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence”). It was issued on November 1, 2023, and can be found at <https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence>. The Trump revocation was the subject of Executive Order 14179, issued on January 31, 2025, and can be found at <https://www.federalregister.gov/documents/2025/01/31/2025-02172/removing-barriers-to-american-leadership-in-artificial-intelligence>.

44 “Treaty on the Prohibition of Nuclear Weapons,” United Nations Treaty Series, United Nations, New York (XXVI–9), opened for signature August 9, 2017, <https://treaties.un.org/doc/Publication/MTDSG/Volume%20II/Chapter%20XXVI/XXVI-9.en.pdf>.

45 For the image, see <https://firefly.adobe.com/>



