FAS FEDERATION OF AMERICAN SCIENTISTS

JUNE 2025

Artificial Intelligence, and Nuclear Command, Control, and Communications

Current Status and Future Risks

GLOBAL RISK
FEDERATION OF AMERICAN SCIENTISTS

ABOUT FAS

The **Federation of American Scientists (FAS)** is an independent, nonpartisan think tank that brings together members of the science and policy communities to collaborate on mitigating global catastrophic threats. Founded in November 1945 as the Federation of Atomic Scientists by scientists who built the first atomic bombs during the Manhattan Project, FAS is devoted to the belief that scientists, engineers, and other technically trained people have the ethical obligation to ensure that the technological fruits of their intellect and labor are applied to the benefit of humankind. In 1946, FAS rebranded as the Federation of American Scientists to broaden its focus to prevent global catastrophes.

Since its founding, FAS has served as an influential source of information and rigorous, evidence-based analysis of issues related to national security. Specifically, FAS works to reduce the spread and number of nuclear weapons, prevent nuclear and radiological terrorism, promote high standards for the safety and security of nuclear energy, illuminate government secrecy practices, and prevent the use of biological and chemical weapons.

FAS can be reached at **fas@fas.org**.

COVER IMAGE: NUCLEAR MISSILE LAUNCH KEYS 1996 VIA WIKIMEDIA COMMONS.

ABOUT THIS REPORT

This report summarizes the key findings and insights from the February 2025 roundtable convened by the Federation of American Scientists Global Risk Program, funded by the Future of Life Institute. This event brought together nuclear policy experts, artificial intelligence specialists, government officials, and academics to examine the intersection of artificial intelligence and nuclear command, control, and communications (NC3) systems. Through this discussion, our experts identified potential risks, benefits, and policy considerations surrounding the integration of AI technologies into NC3 frameworks. The enclosed material is a summary of those findings intended for policymakers and those interested in how we can protect humanity from the catastrophic risks.

Global Risk Program at FAS

The Global Risk Program focuses on addressing and preventing the events and threats that could permanently cripple or destroy humanity. Among them: nuclear war, the next global pandemic, biological attack, and even a collision with a massive near-earth object. Humanity must proactively develop and pursue sound policies to protect against these dangers, including through global cooperation. Find out more at our website: fas.org/issue/global-risk

Funding

This report and the associated workshop were made possible through the generous support of the Future of Life Institute and are part of a wider series on "AI and Global Risk". This series of events will culminate in a global summit in Spring 2026. The views expressed in this report are those of the authors and do not necessarily reflect the positions of the funders or participants.

Acknowledgements

Special thanks to Global Risk Program Associate Elliott Gunnell and Herbert Scoville Fellow Allie Maloney for their contribution to the event and this report.

FEBRUARY 2025 ARTIFICIAL INTELLIGENCE AND NUCLEAR COMMAND, CONTROL AND COMMUNICATION

On February 27th, the Federation of American Scientists, in partnership with the Future of Life Institute, convened an all-day unclassified workshop on the intersection between artificial intelligence and nuclear command, control and communications (NC3) systems. Few topics are as important to national security, or as poorly understood outside of very narrow communities. This meeting was the first of a series of AI/national security events being organized by FAS and FLI, leading up to a Global Risks summit in the spring of 2026.

There is growing interest in integrating AI into the U.S. nuclear command, control, and communications (NC3) architecture—but this interest is outpacing a clear understanding of the associated risks. Critically, there is no existing government guidance to manage or constrain future AI/NC3 integration. This lack of policy creates a strategic vacuum at a moment when foundational decisions are looming. At present, the U.S. NC3 system is not ready for rapid AI adoption—offering a narrow but valuable window to define the terms of integration before deployment pressures escalate. Moreover, key assumptions driving integration—such as reduced decision time and enhanced stability—remain untested against alternative approaches that may achieve similar benefits with less risk. Policymakers must act now to establish rigorous frameworks, assess trade-offs, and ensure that any move toward AI-enabled NC3 is deliberate, accountable, and strategically sound.

EVENT DETAILS

The FAS workshop was attended by 30 members of the nuclear and AI policy community, including from the executive and legislative branches, academia, think-tanks, funders, and the national security community. The event was divided into stand-alone panels on the definitions of AI and NC3, the current research on where AI and NC3 might pose unknown or high risks, and a discussion of how humans interact with AI and machine systems overall and what that means to the future of AI/NC3 integration. A 4th discussion was held on China's current approach on AI nuclear interactions informed by two academic experts on China's approach to nuclear and AI policy.

FINDINGS

Workshops attendees shared a broad understanding that AI remains a relatively immature set of technical capabilities but they are already integrated into parts of the NC2 and NC3 systems, mainly in the areas of early warning and signals processing. There was also a broader sense among attendees that AI would likely continue to be applied into military and even nuclear-related activities incrementally over time. Many of these were described by one presenter as low risk areas like logistics and predictive maintenance that should pose little if any risk to stability or nuclear control. The workshop presenters also noted that AI outputs used outside of time pressures that can be validated by humans notionally posed lower concerns than those outputs designed to support a rapid decision and that could not be subject to human validation.

There was discussion and some concern around the potential application of AI into any role related to nuclear decision making, including from launch authority all the way to decision support, analysis and even applications for real time situational analysis. It was noted that there was, at least to the attendees, no process or metrics yet in place to assess proposals to expand the AI/NC3 nexus. Attendees noted that four of the major five nuclear weapon possessing states had agreed to political statements that humans and not machines should control nuclear decision making, but there was not agreed understanding or details about what that means in practice, and how the further integration of AI into NC3 might compromise those intentions.

There is presently no known US agreed guidance or metrics for assessing possible areas of AI/NC3 integration, or for comparing them against other non-AI or even doctrinal changes that could achieve similar or better results. One key objective discussed by attendees was the prospect for AI/NC3 work to enhance nuclear decision time for the National Command Authority. However, it was noted that decision time can be created through a variety of means that may involve more or less risk that AI integration. This is an area ready for broader conceptual work both inside and outside of government and that can be done outside the confines of classification.

It was noted that US NC3 system modernization was still at a relatively early stage, with two important implications for possible AI integration. The first was that much of the US system is not modern enough to support immediate AI systems in operations. This led to the second observation; that there was a window of opportunity to think about how best to approach the issue of modernization with an eye toward where AI might be a positive addition and where AI would pose real or even unacceptable levels of risk. Including these issues in the design parameters for US NC3 remains a valuable goal to be explored. At the same time, NC3 adjacent programs, including training, simulations, war gaming and other intelligence related functions might be ripe for the use of AI, including LLMs. The risks and limitations of doing so, however, are not well understood or explored, creating risks for the nuclear complex if not properly developed and implemented.

There was a valuable discussion with experts on China's nuclear program and strategic thinking as well about the potential application of AI into Chinese NC3. The main question posed was whether conceptually or programmatically anything China was or might do in this arena would have a significant influence on whether the US should adjust its own analysis of AI/nuclear integration.

Lastly, there was a valuable discussion on the ways in which humans respond to machines and how this should be a greater consideration as countries consider whether and how to expand reliance on AI in key areas. It was noted that there is considerable literature on the extent to which human beings assume machine generated outputs are more reliable that human ones, and how humans will follow machine generated instructions or directions even when there are obvious or observable reasons not to do so. There is also a risk of human capabilities being degraded over time as they rely more on machines, and concern that human skills must be maintained and even enhanced over time or the risk of loss of control and greater dependence on AI systems may degrade the standard of effective human control over nuclear-decision making.

RECOMMENDATIONS

Based on the workshop discussions and subsequent analysis, we recommend the following actions to address the challenges at the AI-NC3 nexus:

1. **ESTABLISH CLEAR GOVERNANCE FRAMEWORKS**
   Systems within Nuclear Command, Control, and Communication, in coordination with relevant agencies, should develop explicit guidelines and metrics for evaluating proposed AI integration that emphasize human control with 'human in the loop' in the systems.
2. **PRIORITIZE HUMAN CONTROL**
   Codify and operationalize the principle of meaningful human control through specific technical requirements, training protocols, and operational doctrines. As NC3 systems are modernized the delineation of human controls versus machine derived inputs in the system. Siloing where AI is able to be integrated, making off limits the most sensitive and critical components in NC3 like launch authority.
3. **CONDUCT COMPARATIVE ANALYSES**
   Before implementing AI solutions, rigorously assess whether alternative approaches—including doctrinal changes, organizational restructuring, or simpler technological upgrades—could achieve similar benefits with lower risk profiles.

4. **CREATE INTERNATIONAL DIALOGUE MECHANISMS**
   Establish dedicated bilateral and multilateral channels to discuss AI-NC3 issues with other nuclear powers, focusing initially on confidence-building measures and shared terminology before pursuing more ambitious agreements.
5. **INVEST IN HUMAN EXPERTISE**
   Maintain and strengthen human capabilities within the exciting nuclear operations. Invest in specialized training programs that develop critical thinking skills specifically for working with AI systems under high-stress conditions.

CONTENTS

KEY FINDINGS FROM ROUNDTABLE

On February 27th, the Federation of American Scientists, in partnership with the Future of Life Institute, convened an all-day unclassified workshop on the intersection between artificial intelligence and nuclear command, control and communications (NC3) systems in Washington DC. Attended by 30 members of the nuclear and AI policy community from the government, think tanks, academia, and philanthropy, the purpose of the roundtable was to discuss risks, integration frameworks, and potential innovations in the AI and NC3 national security realm.

This discussion is at a pressing time as new AI technologies develop and all nuclear weapons states undergo nuclear modernization. While four of the major five nuclear weapon possessing states had agreed to political statements that humans should control nuclear decision making,there is little understanding of what "meaningful human control" comprises. On top of this the United STates has no official guidance for assessing possible areas of AI/NC3 integration, or for comparing them against other non-AI or even doctrinal changes that could achieve similar or better results. Due to this gap, it is an important window of opportunity for policy makers to voice their concerns of risk, analyze doctrine, and propose frameworks and regulations for Artificial intelligence integration in NC2 and NC3.

DEFINING THE STATE OF PLAY

The workshop focused on specific applications of current and predicted artificial intelligence models. The National Security Commission on Artificial Intelligence's definition for AI as "the ability of a computer system to solve the problems and to perform tasks that have traditionally required human intelligence to solve," is a broad category and still leaves a myriad of possible models. For example, classic AI is based upon a programmed rule based system. A more complex machine learning model learns from training data, finding patterns in that data to teach itself how to adapt. Within this, models can become even more complex, using multiple layers of neural networks for deeper analytics. The newest models like ChatGPT and Claude are forms of Generative AI, which produce new media. Currently, the United States utilizes classic AI within multiple nodes of nuclear forces, like in fusing data from multiple command centers for tactical early warning, and is likely exploring how to use some machine and deep learning capabilities. It is unclear if it is considering Generative AI usage.

Participants divided nodes into which AI could be integrated in the vast network of NC3. Maintenance and force management was a common node in which consensus was that there was little risk and even that AI could meaningfully improve outcomes, through risk reduction from nuclear surety. This node was used as a common example of beneficial AI use. Other low-stakes nodes included early warning, where classic AI transmutes multiple streams of data.

Within communications, AI could secure identity & signal verification in communication and eventually could enable operations within comms denied environments. Within Intelligence, surveillance and reconnaissance (ISR), it could combine data from sensors, satellites, and human intelligence to increase ground force awareness. In these nodes, perceptions of risk amounted when generative AI integration was considered. There was a concern that generative data could increase risk of incorrect outcomes or influence human judgment.

Participants identified higher risk nodes included planning, targeting, and decision support/making. Planning consists of creating war plans so that menus of options can be presented. These plans are developed through the analysis of current capabilities, adversary capabilities, geography, and potential threats. Experts noted that for AI to enhance planning, it would require significant amounts of data on these factors. This data, to reduce risk, would have to be based on *ground truth* rather than theory or biased information. An example of AI use in targeting and delivery would be to identify targets, and ensuring the delivery at the proper speed and height. Within this iteration of the roundtable, less conversation focused on this node, though it will be important for further discussion.

Conclusions were that AI use is most beneficial in low-stakes scenarios with clear metrics, availability of real data and human oversight and most risky in high-stakes applications, with less ground truth data and where human oversight is less viable. It was also commonly agreed that there is a need for more psychological research on how humans react to certain framings, option patterns, and styles of information presented with or without AI analyses

Because of the over 200 programs of record present in the NC3 network, conversations focused on areas of the highest risk. This mostly centered around AI for decision enhancement. While some participants agreed that expanding the decision making window through AI enhancement would be beneficial, this was not without risk. Risky forms of AI decision enhancement included adversarial behavior analysis and prediction, analysis of red team perceptions of blue force awareness, and understanding past adversary stances. Another concern surrounding "lengthening the window" was unintended decision time compression. AI integration could actually reduce the time that leaders have for decision making because the AI analyses encourages quicker decision making, thus increasing the tempo of conflicts.

RISKS OF EMERGING TECHNOLOGY

Moral questions arose among participants who wondered: If humans do not have control over this dangerous technology, what do they control? AI agents cannot feel the consequences of a nuclear attack. AI systems are also advancing at rapid paces. Some participants were concerned at the idea of autonomous agents taking over the system. Proposed security considerations included air gapping so that ai agents can not propagate into other systems, and maintaining certain hardware designs. However, the idea was presented that an AI agent could take over and *prevent* a nuclear launch.

The AI data drawbacks are amplified in a highly volatile environment of global nuclear politics. Nuclear weapons are an existential risk; even small scale use of nuclear weapons would cause massive human health and climatic effects. Whether or where these technologies are implanted, perceptions among states of capabilities are of higher consequences. States may resort to worst-case assumptions about their adversary's integration with AI, believing that they have automated launches connected to early warning systems. These kinds of assumptions in a highly volatile environment could lead to escalation due to misperception.

Cyber security drawbacks are amplified as this technology is actively emerging. The intel on how the models will fail and the circumstances under which they will break down is limited. AI is in the early stages of introspection where AI can evaluate itself for transparency, so the risk of data poisoning by an adversary is still high. Systems are still vulnerable to brute force manipulation and hackers can jailbreak even 'secure' models almost instantly to produce potentially harmful results. As states and even nonstate actors invest in cyber hacking, they will be assisted by artificial intelligence for more efficient hacking.

Participants also discussed the risk of AI when there is entanglement of conventional and strategic forces. There is a risk that AI could propagate into various systems, models need not adopt agency to bleed into other nodes. Due to entanglement between conventional and non conventional forces and nodes within NC3, AI analyses or models could be used for unintended purposes as AI use becomes part of more norms. In an adversarial context, AI enabled or enhanced attacks on conventional systems have strategic implications.

While concerns of AI often focus on how the technology could adapt agentic qualities, the concern on how humans will apply and adapt to machine learning may pose inconspicuous challenges. First, humans often appear to have automation bias, trusting an algorithm over human data, assuming the latter is less fallible. If humans become overly reliant on automation, this could result in skill and critical thinking degradation, resulting in panic if the AI system fails. On the other hand, there were some thoughts that AI could be beneficial because it would provide a role in mitigating some internal human bias, and present more balanced fact based analysis.Outside of integration, AI proliferation could lead to a psychological manipulation of a population which results in confusion that endangers

operational success. Participants noted the danger to the US, whose population is increasingly more susceptible to mass manipulation, particularly with more Americans operating on the internet and social media platforms.

GOVERNMENT & MILITARY APPLICATIONS

Participants who had government experience noted there are limiting factors to integration due to the legacy infrastructure of systems. Much of existing systems within the military and government writ large are operating on 'legacy' systems that are extremely outdated. Bringing these systems into the digital age is a challenge, not to mention then integrating these systems with AI. While miltech startups try to break into the industry, they have high market caps but little product to show for it, thus they are outside the legacy military industry. Without clear congressional spending appropriations, there is little hope that non-existent products can enter into the market at present time.

Those same participants also noted the limiting factors due to politics, bureaucracy and normative structures. They said that NC3 is something that people are not willing to take risks, therefore it's logical to assume that most people are at consensus that integrating AI into those systems would have a net negative effect on both perceptions about security and practical implications of AI fallibility. National leaders are on the whole older, so perhaps they would be less likely to trust or integrate AI into systems. Some in the room disagreed saying that leaders are often more accepting of AI than the people working in the lower rungs of the institution. Service and nuclear doctrine would greatly influence decisions on integration, implementation, and outcomes. Some proposed further studying how service members view AI and how influences to doctrine could influence the motivations behind proliferation of AI tech.

Technology advancement requires capital, both human and monetary. The landscape for development includes private sector, academic, and governmental investment. There is a general inquiry among the participants about how the distribution of knowledge to build AI systems is spread across both China and the US. United States policy on immigration may impact high skilled labor from China and India. There is expected legislation to secure national labs from possible foreign nationals who could bring back information to adversary countries.

As AI becomes more integrated in national security systems, potentially there will be increasing compartmentalization of the data used to build, train and create these models and systems at the source. But China has had a longer history of prioritizing the development of AI. Most of the creators of Deep Seek were educated in China. Conversation turned to the future of AGI, and whether the US would pursue a "Manhattan Project" for this technology, knowing it could greatly shift global dynamics.

CONSIDERATIONS FOR CHINESE NUCLEAR POWER

China's force structure and posture could be the reason the PLA is motivated to integrate AI into NC3 architecture. Both China and the United States assume the other is pursuing the AI nuclear nexus to their potential detriment. Some experts hypothesized that China was growing its arsenal because it is worried about the survivability of forces. AI use by the United States for sensor and image detection could threaten China's survivability of their mobile launched systems. However, China could also possibly predict, intercept, and track US seekers to better hide its mobile launched systems.

China's future developments could also depend on the force structure of its SSBNs and development in quantum ocean transparency technologies. The Chinese SSBNs are low priority for China, because their technology is not as developed, resulting in louder engines that are easier to find. As the United States invests in quantum computing, oceans may become more transparent to the US, first negatively affecting Chinese SSBN security. China is

influenced by creating more uncertainty for the adversary and having the ability to coerce their opponents into taking greater risks.

Inter CCP dynamics muddy waters when it comes to predicting leadership perspectives on artificial intelligence. The predominant consensus in China, albeit with little public information, is that China is moving towards a formalized intelligent warfare that harmonizes systems towards an increasingly centralized command structure. Technology is being used to forge China into a greater combat power. Chinese perceptions about AI are divided into four camps: concerned, optimistic, risk acceptant, and risk resistant.

President Xi may depart from the PLA's positive views on AI or he could see AI has an opportunity to engage internationally without limiting its nuclear buildup. President Xi would need to be willing to trust AI judgements, since he has consolidated power. China is increasingly growing its technical workforce and this is indicated in Deepseek researchers being majority chinese. Participants suggested using track 2 dialogues as a way to educate experts and broaden China's thinking on risk perception.

MOVING FORWARD

The expert community should work to provide definitions for congress including what "meaningful human control means." While there was vibrant discussion surrounding areas of high risk and low risk, defining priorities would allow the community to put more specificity into the dangers or benefits of AI to nodes in the middle of that spectrum of the NC3 system.

In order to gain broader support throughout the political spectrum, framing may need to be shifted from ethical dilemmas to the fallibility of AI. Artificial Intelligence may be a viable avenue to pursue arms control or risk reduction with other countries. A lot of the discussions surrounding AI can be boiled down to nuclear doctrine and strategy, though so major changes to these may be necessary to reduce the motivation to proliferate new technologies.

UNDERSTANDING ARTIFICIAL INTELLIGENCE, NUCLEAR COMMAND, CONTROL, AND COMMUNICATION, AND THEIR INTERACTION (EVENT PRE-READ)

JON B. WOLFSTHAL[1] AND DR. OLIVER L. STEPHENSON, PHD[2]

To effectively analyze the intersection of nuclear command, control and communications (NC3) and artificial intelligence (AI), it is critical to define terms. The challenge of managing the AI-nuclear nexus is real, as NC3 and nuclear operations are highly specialized and in many cases highly classified. At the same time AI technology is new to many policy actors and analysts, rapidly evolving, and highly technical in nature. This discussion paper offers a concise overview of NC3, AI, and their intersections to encourage deeper analysis and dialogue on the United States' potential integration of AI into NC3.

NUCLEAR COMMAND, CONTROL AND COMMUNICATIONS (NC3)

The United States controls its arsenal of over 3700 nuclear weapons through a complex web of technical systems and operations known collectively as nuclear command, control and communications (NC3). NC3 is actually a system of systems designed to make sure nuclear weapons will be used only at the direction of the authorized national command authority (the President or his/her legal successor) and that American nuclear weapons will never be used unless properly authorized.

Underpinning this "Always/Never" mission is a web of information collection, processing and transmission capabilities tasked with detecting any relevant threats to the United States or its allies and interests, presenting that information in a form digestible to policy makers, and enabling the President and the US military to remain in active and reliable communication to all facets of the US nuclear complex, as needed. The Nuclear Matters Handbook produced by the Department of Defense states "U.S. command, control, and communications is necessary to ensure the authorized employment and/or termination of nuclear weapons operations, to secure against accidental, inadvertent, or unauthorized access, and to prevent the loss of control, theft, or unauthorized use of U.S. nuclear weapons. The President's ability to exercise authorities is ensured by NC3."[3]

As described in a 2019 report by Dr. Jeffrey Lawson, these NC3 sub systems—and there are over 250 separate programs, combinations of equipment, sensors, or capabilities—were developed and maintained to provide critical functions, including:

- Situation monitoring—including the ability to collect intelligence, assess a threat, provide tactical warning and attack assessment to decision makers, and give them updates on the readiness levels of US forces. This includes gathering and sharing information on friendly forces, adversary forces, and potential targets, as well as global events of interest. The hardware requirements to do all of this are broad and demanding.
- Planning—developing war plans, including the use of nuclear weapons, so as to minimize decision time in a crisis or conflict. This may also entail adaptively planning responses during a crisis.
- Decision-making—senior military and political leaders assessing the situation, consulting in real time, and considering the use of nuclear weapons in certain scenarios.

1 Jon B. Wolfthal is the Director of the Global Risk Program at the Federation of American Scientists in Washington, D.C.
2 Dr. Oliver Stephenson, PhD is the Associate Director AI and Emerging Technology Policy at the Federation of American Scientists in Washington, D.C.
3 Nuclear Matters Handbook 2000 (revised), Chapter 2 Nuclear Weapons Employment Policy, Planning and NC3. https://www.acq.osd.mil/ncbdp/nm/NMHB2020rev/

- Force management—assigning, maintaining, training, and supporting nuclear weapons, nuclear delivery vehicles, and support forces; maintaining force readiness; and ordering the dispersal or deployment of nuclear forces.
- Force direction—NC3 enables NC2 by ensuring the accurate transmission of messages executing lawful strike orders with US strategic nuclear forces or terminating operations at the end of a conflict.[4]

A serious and methodical examination of these essential components of nuclear operations and how each system and subsystem might be affected (improved, compromised or complicated) by the further integration of AI systems is an essential part of ensuring that the United States and its allies can continue to effectively manage the challenge of safely and effectively maintaining a nuclear arsenal.

ARTIFICIAL INTELLIGENCE

Artificial Intelligence is defined by the National Security Commission on Artificial Intelligence, simply, as "[t]he ability of a computer system to solve problems and to perform tasks that have traditionally required human intelligence to solve."[5]

However, this simple definition leaves out much of the complexity associated with AI and its integration into military and nuclear matters. In the context of NC3, AI generally refers to machine-learning-based systems that learn from large volumes of data to carry out particular tasks. This contrasts with more traditional, rules-based computer systems, which rely primarily on if/then statements written by humans.

Because machine-learning-based AI models acquire their behavior from data, they can exhibit outcomes or reasoning pathways that designers did not explicitly anticipate—raising important questions about control, oversight, and accountability in high-stakes situations. There are many ways to classify AI systems, but one popular division is between predictive and generative AI.[6]

Predictive AI refers to systems that perform tasks such as classification, anomaly detection, or forecasting. In an NC3 context, relevant examples might include systems for detecting military hardware in satellite imagery or identifying anomalies in early warning data. Generative AI, by contrast, refers to systems that can create new media—including text, audio, and video. Prominent examples include OpenAI's ChatGPT and Anthropic's Claude, which can generate human-like text in response to user prompts, from detailed explanations to creative writing. It remains unclear, based on public reporting, whether generative AI systems are currently incorporated into NC3.

However, one can easily envision future use cases ranging from generating realistic training scenarios to offering recommendations about military action. AI is rapidly evolving, and this progress will increasingly blur the line between predictive and generative systems. We already see AI models capable of sophisticated reasoning—such as OpenAI's o-series models which can solve complex programming tasks and advanced mathematical problems. Looking ahead, we can expect increasingly autonomous AI "agents" that integrate both generative and predictive elements to formulate multi-stage plans and adapt to unforeseen circumstances.

While AI has made tremendous progress in recent years, several well-known drawbacks and vulnerabilities remain:

- High Data Requirements: Achieving high performance typically demands vast amounts of training data.[7]

4 Jeffrey Larsen, "Nuclear Command, Control, and Communications: US Country Profile", NAPSNet Special Reports, August 22, 2019, https://nautilus.org/napsnet/napsnet-special-reports/nuclear-command-control-and-communications-us-co untry-profile/
5 National Security Commission on Artificial Intelligence Final Report, 2021, https://reports.nscai.gov/final-report/
6 IBM, Generative AI vs. predictive AI: What's the difference?, https://www.ibm.com/think/topics/generative-ai-vs-predictive-ai-whats-the-difference
7 Villalobos, Pablo. 'Will We Run Out of Data? Limits of LLM Scaling Based on Human-Generated Data'. Epoch AI, June 6 2024. https://epochai.org/blog/will-we-run-out-of-data-limits-of-llm-scaling-based-on-human-generated-data.

- Adversarial Vulnerabilities: AI systems can be fooled by seemingly insignificant changes that are imperceptible to humans.[8] Attackers can also inject or modify training data in subtle ways that degrade AI performance or cause systematic errors, an attack known as data poisoning.[9]
- Lack of Explainability: Even AI experts often cannot fully account for why certain systems produce particular outputs.[10]
- Reliability Issues: AI systems can behave in unexpected ways. Large language models, for example, may generate false or misleading information (sometimes called "hallucinations").[11] AI systems also often struggle when real-world conditions differ from those seen in training (known as "distribution shift" or "model drift").[12]
- Emerging Issues with AI Agents: As AI systems gain more independence in decision-making, they can sometimes "game" the goals or rules set for them, pursuing unexpected methods that fulfill instructions on paper but undermine their intended purpose.[13] More concerning is the idea of "sleeper agents," where an AI system appears benign but can switch to harmful actions under certain hidden triggers embedded by an attacker.[14]

Media coverage highlighting AI breakthroughs has created substantial enthusiasm about applying AI to nearly every facet of the economy. Nevertheless, these issues continue to pose serious challenges for using AI in safety-critical arenas—particularly those as sensitive as nuclear command, control, and communications.

In assessing the application of AI for nuclear command, control and communication it remains essential to focus not only on the speed and benefits that can come from such systems, but the flaws and risks that come with such systems. In the end, the reality that machines cannot feel or understand the human and societal consequences of their recommendations and decisions must be a primary factor when thinking about AI and nuclear systems. Accountability and the ability to attribute decisions and automation back to responsible officials or decisions has also been a hallmark of nuclear operations and steps that might dilute or eliminate that should be viewed with great caution and concern.

THE INTERSECTION OF AI AND NC3

Some systems and roles within NC3 are well-suited for AI integration and arguably include little or no risk to the always/never challenge. Indeed, there are clearly applications where predictive models can enhance (and potentially already are enhancing) control, reduce the risks of accidental or authorized use, and further solidify effective control of the US arsenal. Recent analysis by the Institute for Security and Technology (IST)[15] highlights how NC3 systems—both legacy and modern—are rapidly evolving alongside computing, communications, and related technologies. As AI research and adoption expand, governments and militaries are now considering how advanced tools might be incorporated into the NC3 mission. However, questions remain mostly abstract, and discussions often overlook the adversarial dynamics and vulnerabilities that such technologies could introduce into NC3.

8 viso.ai. Attack Methods: What Is Adversarial Machine Learning?. December 2 2023. https://viso.ai/deep-learning/adversarial-machine-learning/
9 IBM. What is Data Poisoning?. https://www.ibm.com/think/topics/data-poisoning
10 IBM. What is AI interpretability?. https://www.ibm.com/think/topics/interpretability
11 IBM. What are AI hallucinations?. https://www.ibm.com/think/topics/ai-hallucinations
12 IBM. What is Model Drift?. https://www.ibm.com/think/topics/model-drift
13 Google DeepMind. Specification gaming: the flip side of AI ingenuity. April 21 2020. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/
14 Anthropic. 'Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training'. 14 January 2024.https://www.anthropic.com/research/sleeper-agents-training-deceptive-llms-that-persist-through-safety-trai ning.
15 Wehsener, Alexa, Andrew W Reddie, Leah Walker, and Philip J Reiner. 'AI-NC3 Integration in an Adversarial Context'. 2023. https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Context.pdf

Historically, arms control efforts have focused on nuclear warheads and delivery vehicles rather than the underlying digital, maritime, or space-based infrastructure. This gap becomes more pressing as next-generation NC3 systems are increasingly integrated with conventional command and control, potentially blurring the lines between nuclear and non-nuclear domains and raising the risk of inadvertent escalation.

The potential security and stability implications of further AI and NC3 integration have been identified, but requires deeper engagement and policy refinement to ensure robust oversight and accountability of a system which is operated by the military but crucial to safeguarding the entire nation. This set of questions is all the more important because NC3 and AI integration is not some hypothetical issue—the US military and combatant commands have been using narrow AI tools for years to collect and process signals intelligence and early warning to provide rapid analysis of global developments as rapidly as possible. This includes areas both outside and within the NC3 process. Understanding what has been done, how it has worked, and how those lessons (good and bad) may be applied to future systems is a fundamental step in developing sound policy.

At the same time, it must be acknowledged that there are areas where the further integration of AI systems have unknown implications and are based on data that may not be fully reliable, accurate or realistic. This lack of certainty and an inability to define both risk and benefit needs to be a component in and of itself in determining whether and how to consider the application of AI in key systems. And lastly, there are areas where NC3 and AI integration carry clear risks that must be weighed against potential gains. This is true both within a domestic only context, but also in a broader interaction with potential nuclear adversaries.[16]

Public discussion of AIxNC3 has largely been limited to the statement of high level principals. For example, the 2021 final report of the National Security Commission on Artificial Intelligence recommended that the U.S. should

> "... clearly and publicly affirm existing U.S. policy that only human beings can authorize employment of nuclear weapons and seek similar commitments from Russia and China."[17]

More recently, the FY25 National Defense Authorization Act affirmed that

> "... [i]t is the policy of the United States that the use of artificial intelligence efforts should not compromise the integrity of nuclear safeguards, whether through the functionality of weapons systems, the validation of communications from command authorities, or the principle of requiring positive human actions in execution of decisions by the President with respect to the employment of nuclear weapons."[18]

However, there is as of now no national consensus or indeed broader public discourse on when and how AI should be further integrated into the NC3 mission. This is not surprising since both the NC3 and AI communities are small, involve specific arcane and often technical knowledge that demand time and attention that may not be forthcoming from key public officials. Given the awesome responsibility that comes with managing America's nuclear arsenal and the consequences of risk in this area, a broader and sustained discussion about the benefits and risks associated with AI/NC3 integration is warranted.

Beyond these points, there are several clear drivers that will demand public discussion and assessment of how deeply AI and NC3 should be integrated. The first is simply the "shiny new object" dynamic—AI is a capable and powerful technology in certain roles. There is a constant assessment of how US NC3 can be improved and enhanced that some analysts are pushing for greater reliance and even delegation of nuclear decision making to automated systems.[19]

16 This point is explained in fine detail in the aforementioned IST report from 2023: Wehsener, Alexa, Andrew W Reddie, Leah Walker, and Philip J Reiner. 'AI-NC3 Integration in an Adversarial Context', 2023. https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Cont ext.pdf

17 National Security Commission on Artificial Intelligence Final Report, 2021, https://reports.nscai.gov/final-report/

18 United States Congress, Servicemember Quality of Life Improvement and National Defense Authorization Act for Fiscal Year 2025, Section 1638 https://docs.house.gov/billsthisweek/20241209/RCP_HR5009_xml%5b89%5d.pdf

19 Lowther, A and McGiffen, C, America Needs a Dead Hand More Than Ever, War on the Rocks, March 28, 2024, https://warontherocks.com/2024/03/america-needs-a-dead-hand-more-than-ever/

The second factor is that, as computing increasingly depends on AI for basic functions, the technology may blend by default with core NC3 systems over time. Lastly, there are areas where those responsible for protecting the always/never capabilities of NC3 see substantial advantages from the application of AI into NC3. For example, STRATCOM Commander, General Anthony J. Cotton has noted the potential value of AI for secure communications in NC3 missions.[20]

Stakeholders bring diverse backgrounds to this discussion, and not all are experts in AI's technical intricacies. Even so, bringing in a range of voices to develop a framework for accountability—complete with clear metrics for assessing risk and benefits—is firmly in the national interest. A sustainable process for guiding AI–NC3 decisions will ensure that innovations in AI strengthen, rather than undermine, the security provided by America's nuclear command architecture.

SOME POTENTIAL DISCUSSION QUESTIONS

- What are potential applications of predictive and generative AI systems to NC3?
- What are the principal benefits and risks of these applications?
- What are the potential applications of emerging AI capabilities (e.g. AI agents)?
- Which applications of AI to the NC3 nexus should be prohibited?
- What is the role of different areas of the US Government on AIxNC3 integration?
- What are the appropriate standards for human judgement and control in nuclear command and control?
- At what point does reliance on AI for nuclear operations and decision making compromise the principle of effective human control over nuclear use?
- What should the US Government be advocating for on the international stage on AIxNC3integration? Are international agreements possible or desirable?
- What verification methods could be used to check compliance with any international agreement?

20 Hadley. G, 2024 AI 'Will Enhance' Nuclear Command and Control, Says STRATCOM Boss', Air & Space Force Magazine. https://www.airandspaceforces.com/stratcom-boss-ai-nuclear-command-control/

HOW WILL HUMAN-AI INTERACTIONS AFFECT NUCLEAR STABILITY (EVENT PRE-READ)

DR. ANTHONY AGUIRRE, PHD[21]

Humans interact differently with each other than they do with machines. Even the belief that one is dealing with a machine instead of a person can alter how actions and information are processed. This dynamic sometimes appears as automation bias, where algorithms are assumed to be less fallible than humans, resulting in an operator incorrectly deferring to an automated recommendation. It can also involve worst-case assumptions about an adversary's reliance on automation—for example, fearing that an adversary's AI-driven early warning system might automatically trigger a nuclear launch, leaving little room for human judgment or de-escalation. Understanding fundamental issues with human-AI interactions, and how the United States perceives major adversary actions in an AI-enhanced environment, is critical to assessing the AI–nuclear nexus.

INTRODUCTION

On September 26, 1983, Lieutenant Colonel Stanislav Petrov was on duty at the Soviet Union's Oko early warning system command center, responsible for detecting any incoming U.S. intercontinental ballistic missiles. The Oko system suddenly reported multiple launches heading toward the USSR, suggesting an imminent nuclear strike. Despite intense pressure and the system's alarming data, Petrov chose not to report an attack to his superiors, relying on his judgment and intuition that the alarm was likely a false reading due to technical glitches. His decision to classify the warning as a false alarm is widely credited with preventing a potentially catastrophic nuclear escalation during a period of heightened Cold War tensions. In this case, a human was able to overcome over-reliance on technology when others may not have. In the AI-nuclear era, this will prove particularly difficult.

AI-HUMAN INTERACTIONS AND AUTOMATION BIAS

There is a long and documented history of humans having greater trust and reliance on technological outputs over their best judgements and intuitions. This problem has become more acute with the rise of automated systems since the computer revolution of the 1950s. This phenomenon can be summed up as automation bias, defined as the tendency for people to over-rely on automated systems and their outputs, often prioritizing machine-generated information over human judgment—even when that information may be flawed or incorrect.[22]

In some cases this bias is well earned as machines can execute many tasks faster and more reliably, over time and at scale, than human beings. However, over trust or bias for technical outputs - or confidence in an output one believes to be produced by machine over humans - is a serious challenge as one combines military and lethal systems with advanced technology and Automation.

21 Dr. Anthony Aguirre, PhD is the Executive Director and Board Treasurer of the Future of Life Institute based in Campbell, California.
22 In some cases, automation bias may swing in the other direction, towards distrust and underutilization: Skepticism toward AI can lead humans to ignore or underuse potentially valuable analysis. Striking a Balance between healthy skepticism and appropriate trust is critical.

AUTOMATION BIAS ISSUES IN THE CONTEXT OF NUCLEAR STABILITY

As nuclear-states consider and further integrate technology, automation and AI into nuclear command, control and communication there is a risk that human judgement, decision making and second guessing will erode leaving humans as part of a decision making loop they no longer control or even fully understand. This is an acute issue when one considers the challenge of automation bias and over-reliance. This applies to a broad range of automated and possibly AI-aided or enhanced NC3 functions ranging from training and simulation, real-time monitoring and decision support. As noted above in one acute case, it is a hard and a rare thing for a human being to decide that what the machines are saying is wrong. When one applies this challenge to a systemic level - including enabling LLM to develop nuclear crisis scenarios due to a lack of real world crises, and then train itself on these models to support human decision making and judgement, there are clear risks introduced into the complex system of nuclear decision making.

Due to corollary advances in nuclear technology, policymakers are in some ways dealing with shorter windows to respond to a nuclear escalation. As modern nuclear launch vehicles (e.g. intercontinental ballistic missiles (ICBMs), submarine-launched ballistic missile (SLBMs)) deliver payloads in a matter of minutes, and missiles with ambiguous payload tips are becoming more common, it is unlikely there would be enough time to independently verify inferences, conclusions, and recommendations made by AI systems.

In addition, the literature on cognitive load has demonstrated that in time-pressured high-stakes situations decision-makers are more likely to rely on technology in other domains. It is hard to think of a more high-stakes situation than a nuclear crisis. Furthermore, nuclear escalation scenarios are most likely to take place with adversaries with whom we already have incredibly low levels of trust.

Finally, in the absence of specialized AI training and education on the limits of AI systems, personnel may misinterpret AI outputs, leading to flawed strategic assessments further up the command chain.

Hence, while there has been considerable focus on ensuring that there are 'humans in the loop' (i.e., the final decision is made by a human authority), this may prove to be challenging in practice. If an AI system claims that a nuclear weapon has been launched by an adversary, it may prove unlikely that human agents would oppose this conclusion. This problem of automation bias has already been demonstrated in other domains, making the problem of ensuring 'meaningful human control' over AI systems is incredibly difficult.

These dynamics raise some broader implications around nuclear command and control that you could discuss. These include:

- Chain of Command: Clear evidence on how AI will inform and indeed direct human decisions remains murky, raising questions about accountability in nuclear operations.
- Compliance with International Law: The vast majority of international law, including treaties, case laws, and resolutions call for human involvement and accountability in the decision to employ a nuclear weapon. Human-AI dynamics make it complicated to demonstrate adherence to this body of law.

OTHER STRUCTURAL EFFECTS OF HUMAN-AI INTERACTIONS ON NUCLEAR STABILITY

While automation bias and corollary issues are the most urgent concerns as human-AI interactions are related to some nuclear stability, it is also key to consider some more structural effects of human-AI interactions:

- Worst-case assumptions may interpret AI-driven actions—or even the mere capability to deploy AI—through a lens of mistrust, assuming adversaries might rely on automation in ways that increase the likelihood of a first strike.
- Human Signal vs. AI Noise: It will become an increasingly hard challenge discerning actual adversary intent from the "noise" of AI exercises, testing, and routine system updates. Misinterpretations can heighten the risk of escalation.

DISCUSSION QUESTIONS

- What elements of AI-NC3 integration are most likely to be relevant to issues of automation bias?
- What measures can be used to reduce automation bias for key principals in a time-constrained context such as a nuclear escalation scenario?
- How can we establish a clear set of guidelines around chain of command and compliance with domestic and international law cognizant of human-AI dynamics?
- Considering the potentially serious implications of human-AI interactions in the nuclear setting, what should our bar be for integrating generative AI into the various components and sub-components of NC3?

EMERGING AREAS OF AI-NUCLEAR CONSENSUS ARE SOME APPLICATIONS MORE CONCERNING THAN OTHERS? (EVENT PRE-READ)

JON B. WOLFSTHAL

Issues related to nuclear weapons are often seen through a theological lens. Therefore, it should not surprise anyone that when experts consider the further integration of artificial intelligence into nuclear command, control, and communications that opinions land largely along traditional theological lines. These range from strong support for advancing this technology as part of the nuclear deterrence mission to extreme skepticism mirroring that often applied to the role, value, and even morality of nuclear weapons and the ability of human beings to effectively control them and other advanced technology.

A funny thing happens, however, when experts and practitioners talk about specific applications as opposed to broader concepts and themes—debates tend to become less theological and more constructive. It is therefore very useful when discussing the AI-nuclear nexus to engage in specific discussions about applications, making it possible to find agreement about which areas hold high risk, low risk, no risk or unknown and possibly unknowable risk.

For example, an analyst skeptical of introducing increased automation into nuclear operations in theory may have very little, if any, concern about using AI systems to predict when components in large complex systems or nuclear weapons themselves might fail in order to do predictive maintenance and/or surveillance.[23]

At the same time, there is broad concern about introducing automation into any component of the NC3 complex that might effectively erode human control over the decision to launch nuclear weapons.There has been some excellent unclassified work assessing areas where AI may be integrated into NC3 capabilities. One 2023 reportlays out potential areas of integration in the following way:[24]

> Areas of Opportunity: AI Integration in NC3
>
> - Nuclear weapons security
> - Survivability (decreasing the effectiveness of jamming)
> - Integrated air defenses
> - Navigation assistance, particularly in GPS denied environments
> - Improved targeting data, including meteorological data, with AI increasing precision and speeding up analysis of additional factors needed for targeting decisions.
> - Planning, distribution, responsibility and effectiveness
> - Image signal processing for better detection of weapon movements or launches, specifically early warning or other emerging targets that humans would not have conceived of
> - Cyber offense and defense
> - Decision support"

Each of these areas may be attractive from an operator's perspective. However, they may also introduce instabilities, risks, or even unacceptable levels of uncertainty. The impact depends on several factors: how and where the capabilities are applied, what benefit or problem the integration is trying to address, and whether or not these are being adopted internally or by an adversary. Thus, simplistic assessments about whether AI/NC3 integration is

23 Hruby, Assessing and Managing the Benefits and Risks of Artificial Intelligence in Nuclear-Weapon Systems, January 2021, NTI, https://www.nti.org/news/new-nti-paper-assesses-benefits-and-risks-of-ai-in-nuclear-weapon-systems/

24 Wehsener, et al, "AI NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence Building Measures", Institute for Security and Technology, February 2023 - https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an-Adversarial-Cont ext.pdf

"good or bad" need to be replaced with more nuanced and specific understandings of how these systems interact and how they support the broader mission of deterrence and preventing nuclear use.

Thus, instead of trying to identify each and every component of the NC3 complex and determining the specific level of risk tolerance associated with introducing enhanced automation or AI, it may be more productive for those in the broader policy and analytical community to work from the other direction and identify those high-risk areas that most, if not all, informed observers can agree pose unacceptable or unknowable risks from increased reliance on AI.From there, discussions in the public policy and civil society realm can pursue sustainable approaches that seek to protect the most dangerous or unknowable risks from being introduced into nuclear operations.

There are several clear examples where AI-NC3 integration will raise concerns across a broad cross section of analysts, and potentially even among a broader array of nations. The clearest example is whether or not artificial intelligence or automation should have the ability or authority to employ nuclear weapons. The United States, France, United Kingdom, and China have all made various statements that human beings should always be in effective control of nuclear weapons. They are indications that Russian officials think among similar lines, even if they have not agreed to diplomatic initiatives to that effect. While it is not fully clear what "effective control" actually means and at what point the application of artificial intelligence into NC3 could erode that level of control, the fact that multiple countries agree that nuclear decision-making should always remain in human hands demonstrates that areas of agreement and even consensus are possible.

This then raises the issue of what specific applications of AI-NC3 integration should be viewed with grave concern or skepticism.[25] This analysis has to consider that there are certain things that have been done by the United States that would not concern the United States but may concern its adversaries. Likewise, actions by China or Russia to integrate artificial intelligence into certain nuclear operations may seem logical and stabilizing to them, but could cause considerable instability or concern here in the United States.

One example of this would be the ability of China to integrate large numbers of underwater unmanned vehicles (UUVs) and combine them with enhanced AI modeling of ocean dynamics, temperature, and acoustic signatures to increase the vulnerability of US ballistic missile submarines. It's reasonable to expect that the United States has been engaged in research on this area to help enhance its own anti-submarine capabilities.[26] However, strategists can easily recognize that anything that undermines the ability of nuclear weapon states to have confidence in the reliability of their submarine ballistic missile capabilities could be inherently destabilizing. Yet it is hard to imagine the President asking the Navy to stop enhancing its ability to find and target Russian or Chinese nuclear submarines.

It is necessary to ask what other areas of nuclear operations are both attractive areas for AI enhancements and could have significant impacts on stability and human control of nuclear decision making. Some of these applications may be small and subtle – such as allowing AI to assess an adversary's operations and infer intent. Even seemingly minor design choices in automated systems, like how information is visually or linguistically presented, could shape critical decisions. For example, a green box stating "missile fueling identified" conveys a vastly different sense of urgency than a red flashing box declaring "missile launch preparation detected." Given the life-and-death consequences of nuclear decision-making, information provided to key leaders has always been handled with extreme care. Delegating this responsibility to AI decision support systems introduces unpredictable risks.

Other potential applications of AI discussed in the open literature include AI provided nuclear-decision support for the national command authority. How would or should a President assess options being provided by a very capable AI system as compared with key military and civilian leaders? Such officials are often confirmed or promoted by the Congress, providing a critical part of the national system of checks and balances. While imperfect, they have

25 An excellent discussion of these issues was done by Michael Depp and Paul Scharre in Artificial Intelligence and Nuclear Stability, January 2024, War on the Rocks, https://warontherocks.com/2024/01/artificial-intelligence-and-nuclear-stability/

26 Accessed on May 14th, 2025: DARPA "ACTUV: Anti-Submarine Warfare (ASW) Continuous Trail Unmanned Vessel" https://www.darpa.mil/research/programs/anti-submarine-warfare

certain intangible democratic advantages over an AI supported agent. How then should AI decision support be employed, if at all?

It is also not too far a leap to think about how the United States might seek to use its growing AI capabilities for integrated battle management. For decades, Russia—and more recently, China— have expressed concern about the ability of the United States to use its conventional precision, strike, missile defenses, and its nuclear weapons as part of a "splendid strike." A splendid strike would involve the United States using conventional capabilities to destroy or disarm a significant portion of an adversary's nuclear weapons, and use missile defenses to limit the damage that could be brought to bear in response. Even if imperfect, such a strike would leave the United States with a far greater number of nuclear weapons, which it could use to potentially blackmail or coerce its adversary into submission.

A key limitation of a potential splendid strike has been Russia's reliance on mobile missiles and, increasingly over the past decade, China's deployment of nuclear weapons on strategic submarines or bombers. However, if the United States were willing and able to employ highly effective targeting technology integrated through rapid AI-driven planning, its perceived ability to actually pursue and implement a splendid strike would likely increase. How then should the United States view the temptation of such a program, and how would it on the other hand, view the use by Russia or China (when its arsenal grows to include a larger number of nuclear weapons) to do something along similar lines?

Beyond specific application of AI in NC3, broader concerns arise about human control over AI counterparts. Not long ago, the idea of AI systems autonomously learning, adapting, and making complex decisions was confined to science fiction and Hollywood. Today, however, we are seeing major AI companies start to develop AI agents capable of planning and executing basic tasks autonomously. While these agents remain rudimentary, it is possible to imagine—at least conceptually— powerful AI agents behaving outside of their design parameters with potentially disastrous consequences. Already, AI systems have shown signs of being able to deceive their users in certain controlled situations,[27] and AI capabilities are progressing rapidly.[28] These 'over-the-horizon' risks may be closer than they appear, underscoring the need for careful deliberation as both advocates and skeptics work toward sustainable frameworks for AI's role in nuclear command and control.

SOME POTENTIAL DISCUSSION QUESTIONS

- What uses of AI in NC3 are widely agreed to be acceptable by the United States?
- What uses of AI in NC3 are widely agreed to be unacceptable by the United States?
- Which uses of AI in NC3 by adversaries would present the most serious concerns?
- How can our understanding of human-computer interaction be used to better understand potential risks and benefits of AI NC3 integration?

27 Anthropic, Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training, Jan 14 2024, https://www.anthropic.com/research/sleeper-agents-training-deceptive-llms-that-persist-through-safety-trai ning; Vox, The new follow-up to ChatGPT is scarily good at deception, Sept. 14 2024, https://www.vox.com/future-perfect/371827/openai-chatgpt-artificial-intelligence-ai-risk-strawberry; Meinke et al., Frontier Models are Capable of In-context Scheming, https://arxiv.org/abs/2412.04984

28 Our World in Data, Test scores of AI systems on various capabilities relative to human performance, https://ourworldindata.org/grapher/test-scores-ai-capabilities-relative-human-performance

WANT TO LEARN MORE ABOUT OUR AI X NEXUS SERIES?

Please visit fas.org to learn more about upcoming events publications, and Global Summiy 2026.

FAS FEDERATION OF AMERICAN SCIENTISTS

ABOUT THE FEDERATION OF AMERICAN SCIENTISTS

The Federation of American Scientists is dedicated to democratizing the policymaking process by working with new and expert voices across the science and technology community, helping to develop actionable policies that can improve the lives of all Americans. For more about the Federation of American Scientists, visit **FAS.org.**