Marcellus
POLICY ANALYSIS

Fall 2024 - Marcellus Policy Analysis

Modernizing Arms Control: The Case for Codifying Oversight in AI and Nuclear Command Policy

By Sofia Guerra

EXECUTIVE SUMMARY

The rapid integration of artificial intelligence (AI) into nuclear command, control, and communications (NC3) systems presents both opportunities and risks. While AI can enhance decision-making and operational efficiency, it also increases vulnerabilities, such as automation bias, miscalculation risks, and compressed decision timelines. In the U.S.-China context— marked by mutual mistrust and strategic competition—these risks are particularly acute and heighten the potential for unintended escalation.

This paper assesses the limitations of current safeguards and argues that codifying "human-in the-loop" (HITL) oversight into U.S. law is essential but insufficient. To address the broader risks, a multifaceted strategy is proposed: passing legislation to prohibit fully autonomous nuclear weapons systems, investing in AI safety research, and pursuing confidence-building measures with China. Complementary multilateral initiatives, such as joint missile notification systems and agreements on AI governance, are also critical to stabilizing the evolving security environment.

By addressing these challenges with a comprehensive governance framework, the United States can lead efforts to mitigate the risks of AI-driven escalation while fostering stability in U.S.-China relations. This approach balances innovation with security, ensuring that technological advancements in NC3 systems enhance rather than undermine global stability.

Sofia Guerra is a Government Relations Associate at Win Without War, where she engages Congress on nuclear risk reduction, security assistance, and Pentagon budgeting. A Mexican-Salvadoran-American native of Daly City, California, she is passionate about making American foreign policy work for transnational security. She graduated from Amherst College with a degree in Political Science and Asian Languages and Civilizations in 2022. In her free time, you can find her checking out concerts, practicing her Arabic, or running around D.C.

Introduction

Amid heightened tensions over territorial encroachment, economic protectionism, and nuclear expansion, the United States has actively pursued nuclear stability talks with the People's Republic of China. A notable development in this effort is the emphasis on maintaining human control over nuclear weapons systems — a modest but significant entry point for wider arms control discussions. This emphasis reflects a growing urgency to address emerging technologies, such as advanced AI, as shared global risks. The dissolution of longstanding nuclear arms control treaties, which historically focused on limiting arsenals rather than their operational frameworks, further underscores the need for innovative approaches to strategic stability. In this context, President Joseph Biden and President Xi Jinping publicly agreed that AI should not control nuclear weapons, yet diplomatic progress has been stalled by Beijing's insistence on preconditions, including U.S. concessions on Taiwan.

Given these challenges—and the heightened risk of nuclear miscalculation in a potential Taiwan conflict—the United States must take proactive steps to ensure nuclear stability. The competitive integration of AI into nuclear command, control, and communications (NC3) systems magnifies this urgency, as both nations prioritize technological superiority over mutual stability. The risks of automation bias, compressed decision timelines, and opaque AI decision-making exacerbate the dangers of an arms race dynamic, where technological breakthroughs outpace governance measures.

One viable confidence-building measure (CBM) to lay the groundwork for more in-depth conversations with China about nuclear safeguards and strategic stability is legislative action within the United States. Senator Ed Markey and Representative Ted Lieu have introduced a bill to mandate human oversight over all processes informing and executing a nuclear launch. Short of changing the United States' position on Taiwan as a precondition for bringing China to the table, this legislative effort would prevent the use of congressional funds for fully autonomous AI-enabled NC3 systems, ensuring human oversight remains at the core of nuclear decision-making.

This analysis explores the key technological, strategic, and policy dimensions of human oversight in nuclear decision-making. It begins by examining the technological developments behind AI's recent boom and their implications for NC3, focusing on the dual-edged nature of its data-processing capabilities. Next, the essay considers the geopolitical complexities of U.S.- China relations, including relevant AI-enabled technologies that shape 21st century warmaking for the two parties. including how AI governance efforts intersect with broader strategic tensions. Finally, it outlines concrete policy recommendations for pursuing human oversight or "human in-the-loop" codification in federal law, fostering international collaboration, and advancing AI safety measures to ensure responsible innovation.

As the backbone of a nation's nuclear arsenal, NC3 systems comprise a vast, classified network of sensors, communication links, and decision-making frameworks that integrate conventional and nuclear operations. These systems embody the "always/never" paradox: nuclear forces must always be ready for use if needed but never deployed in error. While integrating AI promises enhanced early-warning capabilities and situational awareness, it also introduces unprecedented risks to system reliability and stability. Addressing these challenges will require balancing innovation with security to ensure that technological advancements strengthen, rather than undermine, global strategic stability.

Emergent Properties and Technological Limits of Advanced AI

It is well known that the range and depth of AI applications has experienced a boom in recent years, owing much to key developments in computing power, data, and novel algorithms in the last decade. In 2012, convolutional neural network AlexNet's success showcased the ability to process vast amounts of complex, image-based data.[1] Just five years later, the transformer-based computer architecture style revolutionized natural language processing, paving the way for advanced AI like GPT-3 in 2020.[2] In parallel, the introduction of powerful hardware accelerators like graphic processing units in 2016 expanded access to high-level computing for AI. These developments have contributed to a 350 million fold increase in the amount of computing power used to train leading AI.[3] Growing data availability, due to platforms like Facebook and Twitter in the early 2010s, further accelerated model training capabilities.[4] Finally,

strategic investments by tech giants and, increasingly, the U.S. Department of Defense (DoD) have ensured rapid advancement. This unprecedented pace of AI development, fueled by international competition and corporate innovation, has drawn significant interest from the U.S. military as it seeks to modernize its capabilities in response to evolving geopolitical threats.

Most leading AI models employ transformer architecture due to their superior ability to handle large amounts of data and contextual information. Unlike traditional AI systems that rely on explicitly programmed rules, frontier AI models learn patterns and make decisions based on vast amounts of data processed through multiple layers of computation. Each layer extracts increasingly abstract features from the data, but the relationships and transformations within these layers are not easily interpretable. For instance, while a model might correctly identify a missile launch in satellite imagery, the precise features or correlations it used to arrive at that conclusion are often inscrutable even to its developers. This opacity is further compounded by the sheer scale and complexity of these systems, which involve billions of parameters interacting in non-linear ways.

Lack of transparency in AI systems exacerbates risks of data manipulation. This problem is not merely one of technical inconvenience; it creates vulnerabilities to adversarial manipulation, such as data poisoning, where attackers corrupt training datasets to produce escalatory behaviors. The reliance on black-box systems also means that errors in data ingestion—such as mislabeled inputs or misinterpreted signals—can propagate through decision chains without correction, further increasing risks during crises. Fortunately, risks of data poisoning can be addressed at the manufacturing level, for example by incorporating anomaly detection, federated learning techniques, and adversarial training. Even more difficult to tackle than these tactical risks are negative emergent properties of AI, a strategic challenge tied to the fundamental limitations and uncertainties of AI.

A few characteristics of advanced AI have presented stubborn challenges to reliable integration into high-risk systems — phenomena that make AI systems unpredictable under real-world conditions. Brittleness refers to an AI system's inability to handle scenarios outside its training or design parameters.[5] Despite the massive amounts of data ingested, datasets are often limited in scope. For example, a computer vision model trained to identify vehicles on well-lit highways may fail to detect the same vehicles under poor lighting or unusual weather conditions. Similarly, in NC3, brittleness could undermine reliability during crises. In an ambiguous early-warning scenario, such as unexpected radar reflections from atmospheric disturbances, a brittle system might misclassify benign anomalies as hostile missile launches, creating a risk of unnecessary escalation.

Overfitting, on the other hand, occurs when an AI model becomes overly fixated on patterns in its training data, leading to incorrect generalizations for new or unseen data. For instance, a machine learning model designed to detect phishing emails might erroneously classify legitimate emails as threats if they share superficial linguistic patterns with its training data. In the NC3 context, overfitting could cause an AI system to mistake routine adversarial actions, such as scheduled military exercises, for imminent strikes based on a pattern of resemblance. This risk is particularly concerning as it could intensify human cognitive biases, where decision-makers rely on AI outputs without questioning their validity, further complicating nuclear decision-making.[6] This property, in addition to other factors like data bias, can exacerbate hallucinations in advanced AI, or the generation of false data or conclusions. In a nuclear context, this could look like the detection of an attack where none exists or failing to identify actual threats.[7]

Advanced AI's fundamental reliance on vast quantities of high-quality data compounds these challenges in the nuclear domain.[8] Real-world nuclear strike data is scarce or nonexistent, relying on synthetic data, simulations, or proxy datasets, which may introduce biases, inaccuracies, or untested assumptions into AI models. Analysts and policymakers, relying blindly on AI outputs, may overestimate their validity. For instance, the fratricides involving Patriot missile systems during the 2003 Iraq invasion underscore how automation bias and misclassification errors can lead to catastrophic outcomes.[9] These vulnerabilities are even more pronounced in nuclear decision-support contexts, where technical and institutional structures are already strained by compressed timelines and the complexities of multi-domain operations.[10]

Addressing the flaws' root causes is complex and, in many cases, only partially surmountable.

Improvements in AI interpretability — a field that focuses on making decision-making processes more transparent, understandable, and verifiable, offer one avenue to mitigate these risks. For example, techniques that "open the black box" of machine learning algorithms or allow operators to open the "lid" of each data-processing "node" in computer architecture parlance, may allow human operators to verify the system's outputs.[11] Despite advancements, interpretability remains limited in its ability to predict emergent failures like hallucinations, as these behaviors often arise from complex, non-linear interactions among billions of model parameters.

Additionally, cutting-edge interpretability techniques typically require additional computational resources and time, potentially lowering its operational utility in high-stakes environments. Manually verifying the types of data processed — and ensuring data has not been poisoned — in generating a particular piece of information further takes time that can impede the system's promptness and reverse some of the very advantage for which it was installed. Conversely, forgoing interpretability to maintain speed risks reliance on opaque AI systems, which may produce outputs that are difficult to trust or verify. Balancing the need for AI transparency with the imperative for swift decision-making is essential to ensure both the reliability and responsiveness of AI-enabled NC3.

For the U.S. military, the integration of AI into NC3 promises tactical advantages, especially in enhancing intelligence, surveillance, and reconnaissance (ISR) capabilities. At the 2024 DoD Intelligence Information System Conference, U.S. Strategic Commander General Anthony Cotton emphasized that advanced AI and robust data analytics capabilities enhance situational awareness — military parlance for the ability to perceive, understand, and predict the dynamic elements of a given operational environment— provide decision-making advantages, and improve "integration of conventional and nuclear capabilities, strengthening deterrence."[12] Multi-domain integration has been a cornerstone of the U.S. military's modernization plans, as exemplified by the U.S. Combined Joint All-Domain Command & Control (CJADC2) architecture connecting and managing space, ground, air, sea, and cyber sensors into a unified system. Advanced AI plays a key role in accelerating the fusion of data streams of satellite imagery, radar data, and cyber intelligence, offering significant operational advantages but potentially amplifying risks in NC3 systems.

Leveraging AI reduces the human burden of searching, processing, and analyzing data, enabling commanders to develop a coherent operational picture and make better-informed decisions in crises. However, fused data streams add another layer of ambiguity and complexity, making AI more prone to error in NC3. CJADC2's reliance on AI decision-support tools — each major combatant command has been equipped with the same software — is intended to accelerate data processing but inadvertently overlook escalation risks. A significant challenge of the fused data sources is the integration of controls for conventional and nuclear weapons systems, which can blur distinctions and elevate the risk of miscalculation or unintended escalation. For instance, a conventional strike might be misinterpreted as a preemptive attack on nuclear assets.[13] This risk is especially pronounced in any scenario where quick action is prioritized over the deliberate decision-making processes necessary for strategic stability, such as during defensive or offensive counter-air operations. While the Defense Department has introduced system redundancy and error correction to mitigate these risks military contexts often lack the clear judgment and robust real-world data sets needed for AI-enabled systems to perform reliably.[14]

Strategic competition between the United States and China further propels global investment in AI development, to the extent that fully autonomous systems — conventional and nuclear — are increasingly feasible. Although current NC3 already incorporates some automation aided by older AI models, improvements in generative AI and image and language processing allow for them to operate beyond predefined rules and adapt to new inputs dynamically and without human oversight.[15] Semi-autonomous systems require operator approval for critical decisions, and AI-assisted systems focus on enhancing human decision-making by processing complex data more effectively. This spectrum of autonomy underscores the necessity of tailored governance strategies for each category, particularly given the risks and opportunities unique to NC3 contexts. While no country has officially announced plans

to fully automate nuclear systems using advanced AI, Russia's Perimeter system—a semi-autonomous retaliatory capability— offers a historical example of pre-delegated authority in nuclear decision-making. Further, a deepening arms race dynamic with China has weakened some experts' assessment of U.S. second-strike capability (i.e., the ability to launch a devastating nuclear strike after absorbing a first strike), meaning an American "dead hand" meant to automatically retaliate if U.S. leadership were to become incapacitated could become reality.[16]

US-China Relations and Strategic Stability

The integration of advanced AI into nuclear command and control is unfolding within the broader context of U.S.-China strategic competition. As the world's two largest economic powers, U.S. and Chinese policies on AI integration and nuclear command not only shape their bilateral security dynamic but also influence global norms around strategic stability and arms control. Beijing accuses Washington of invoking outdated Cold War paradigms to justify what it sees as a containment strategy and in response, U.S. officials argue that China's aspirations to annex Taiwan and adopt advanced military capabilities represent a bid to redefine global strategic stability. In 2022, Assistant Secretary of Defense for Indo-Pacific Affairs Ely Ratner remarked "if China is our pacing challenge, Taiwan is our pacing scenario," a framing that underscores an invasion of Taiwan as a critical flashpoint guiding military modernization efforts.[17]

Congressional discussions and DoD priorities increasingly reflect this competition, centering on deterring an invasion by matching China in area-denial systems — including unmanned systems, anti-satellite weapons, and hypersonic systems. Following revelations of China's nuclear weapons development in 2021, the 2023 Strategic Posture Commission report — referenced in subsequent years' Congressional budget proceedings— recommended expanding the deployment of nuclear warheads and bolstering industrial capacity for nuclear weapons production. That same year, testimony before the House Armed Services Committee emphasized the utility of area denial systems like AI-enabled swarm robotics and intelligent, redundant ISR platforms for decision support as cost-effective items to acquire or transfer to Taiwan as a deterrent to an invasion.[18]

This competition shapes a security spiral dynamic, where actions perceived as defensive by one side are interpreted as escalatory by the other, with AI playing a compounding role.

Like the United States, the integration of advanced AI-enabled technology into its military doctrine is central to China's modernization mission, albeit with a few distinctions. The People's Liberation Army pursues an "intelligentization" strategy leveraging AI to enhance decision-making, battlefield awareness, and operational efficiency in NC3, modeling its multi-domain precision warfare doctrine after U.S. CJADC2.[19] Unlike the United States, PLA writings have focused on AI enabled systems with more autonomous, task-specific decision-support capabilities, aiming to directly influence the decision-making processes of key individuals and public opinion to annex Taiwan without waging a difficult conventional war.[20] This focus underscores China's broader strategy to mitigate perceived military vulnerabilities, leveraging AI to offset U.S.

technological advantages.
Much of the reasoning for an ambitious emphasis on "intelligent" systems centers around a perceived capabilities gap, particularly in areas like missile defense, ISR dominance, rapid decision-making enabled by advanced analytics, and long-range precision-strike capabilities. Since withdrawing from the Anti-Ballistic Missile Treaty in 2002, U.S. investments in missile defense systems designed to neutralize ballistic threats have prompted China to develop countermeasures such as hypersonic glide vehicles and nuclear-armed multiple independently targetable reentry vehicles (MIRVs) to penetrate these defenses.[21] China views U.S. ISR advancements, driven by superior AI training and computing power, as facilitating battlefield awareness and decision-making superiority that pressures the PLA to enhance its own ISR and command systems.[22] Similarly, Washington's deployment of long-range precision-strike programs in East Asia and pursuit of medium and intermediate range missiles, are viewed as a significant threat to China's second-strike capability, driving further investments in survivability and counter-capabilities, especially since prevailing wisdom has framed the United States as a frequent benefactor of nuclear coercion.[23]

To address this perceived capabilities gap, China has

excelled in developing the area-denial systems the U.S. military is currently pursuing. For example, the AI-enabled DF-17 hypersonic glide vehicle, with its advanced maneuverability and the ability to evade missile defenses, places China ahead of the United States, whose hypersonic systems remain largely experimental. The PLA has also tested AI-enabled drone swarms capable of coordinated maneuvers, reconnaissance, and even attack missions.[24] These swarms, designed to overwhelm traditional air defense systems, represent a force-multiplying capability that is particularly relevant in a Taiwan contingency. Unlike U.S. systems, which are still refining their scalability and resilience in large-scale swarm operations, China has demonstrated significant advances in deploying autonomous platforms that can adapt to dynamic battlefield conditions.[25]

Together, these technologies underscore how China is leveraging AI-enabled capabilities to outpace America in certain, but not all, critical areas of military modernization.

Parallel modernization efforts reveal an underlying escalation dynamic characteristic of a technological arms race, where advancements by one side provoke responsive, purportedly defensive developments by the other, perpetuating an iterative cycle of competition. The competitive integration of AI into NC3 further intensifies this security spiral by introducing opaque and inadequately-tested systems.

For instance, U.S. AI-enabled ISR capabilities aimed at improving early-warning systems may be interpreted by China as preparation for a first strike, prompting AI-driven countermeasures that reinforce mutual suspicion. However, these dynamics are not solely a matter of technological competition. Much of the escalation is driven by mutual misperceptions, as both the United States and China interpret the other's actions and intentions through a lens of mistrust.

Understanding these misperceptions is critical to addressing the broader risks of strategic instability in an era of rapid AI-driven advancements. For the United States, overestimating China's nuclear posture and technological advancements has fueled fears of an offensive shift. In 2022, then-STRATCOM Commander Admiral Charles Richard warned that Chinese nuclear expansion was continuing more rapidly than U.S. efforts, remarking "as I assess our level of deterrence against China, the ship is slowly sinking."[26] This grim assessment has been echoed in the SPC report's laundry list of nuclear expansion recommendations, in a Livermore National Laboratory study group, and in other military leaders' divergent but still alarming projections despite an overall Pentagon rollback of the higher-range of projections —1,500 warheads, or parity with the United States, by the mid-2030s.[27] The reason why missile defense is such a concern for the PLA is because of its declaratory policy — missile silo construction, deployment of MIRVs, and hypersonic glide vehicles are all designed to ensure the survivability of its second-strike capability, consistent with its no-first-use (NFU) policy.[28] Similarly, while China indeed explicitly seeks superiority in select domains, like AI-enabled "intelligent" weapons systems, it is important to note many specific advancements remain focused on survivability and deterrence—not offensive parity or dominance.

This dynamic is not one-sided. Just as the United States perceives many of China's military advancements as escalatory, Chinese experts have also overstated U.S. military capabilities or fundamentally disagreed with or misunderstood U.S. assessments of destabilizing systems and behaviors. For instance, many Chinese experts argue that the primary destabilizing factor in the U.S.-China nuclear relationship lies in Washington's adoption of a launch-under-attack (LUA) posture, which pressures China to further enhance the responsiveness and survivability of its nuclear forces.[30] From this perspective, the U.S. reliance on nuclear coercion to influence conventional conflicts absolves China of responsibility for managing escalation risks. Many Chinese analysts also believe that the onus is on the United States to confront its unwarranted skepticism of China's NFU policy. Specifically, they dismiss U.S. concerns about the misidentification of dual-use systems like China's hypersonic DF-26 missile. his misalignment has real consequences: U.S. ISR assets monitoring dual-use systems like China's hypersonic DF-26 missile could inadvertently escalate tensions by misinterpreting conventional maneuvers as nuclear threats or vice versa..[31] For example, an AI model might misclassify routine DF-26 missile tests as offensive preparations, prompting disproportionate responses due to the system's opacity or reliance on incomplete data.

However, in many Chinese analysts' view, the onus

is on the United States to address its concerns by changing its LUA posture. Lack of transparency between the PLA and local experts mean the latter may not be aware of developments like the PLA Rocket Force recently-devised "lower the nuclear coercion threshold" doctrine, further exacerbating misperceptions.[32]

These divergent deterrence approaches—where the United States emphasizes missile defense and rapid response, while China prioritizes second-strike survivability—create a profound misalignment that exacerbates mistrust. This effect is further compounded by some Chinese experts' overestimation of U.S. military capabilities, reinforcing existing concerns about a capabilities gap and driving prioritization of AI integration over regulation. These entrenched misperceptions on both sides create a dangerous feedback loop, where misjudgments of intent and capability amplify the risks inherent in technological competition.

The integration of AI into NC3 systems further complicates this dynamic, as opaque decision-making processes exacerbate escalation risks and compress timelines for effective human intervention.[33] For instance, internal military writings show the PLA increasingly emphasizes "warfighting" capabilities alongside traditional deterrence, suggesting readiness to lower the nuclear coercion threshold in extreme scenarios. Following Russia's nuclear signaling during the Ukraine conflict, Beijing appears to be adopting more flexible nuclear options to deter external intervention, introducing ambiguity that undermines crisis stability.[36] In other words, doctrinal shifts suggest a move toward enhancing the flexibility and responsiveness of its nuclear arsenal, weaponizing ambiguity around escalation thresholds to deter external intervention. This shift is compounded by growing concerns about the United States' deployment of lower-yield nuclear weapons, such as the W76-2 warhead, which Chinese experts argue irresponsibly lowers the nuclear threshold. Analysts from the China National Nuclear Corporation have noted the inherent ambiguity in distinguishing between strategic and low-yield nuclear strikes, warning that this ambiguity could lead to catastrophic miscalculation. [37] A limited U.S. nuclear strike, for example, could be misinterpreted as a high-yield countervalue attack on non-military assets, prompting major retaliation and turning a localized nuclear conflict into full-scale nuclear war.

Although often characterized by a centralized and autocratic decision-making structure, internal debates over nuclear posture and modernization have revealed an uncertain alignment between political mandates and operational doctrine.[38] The PLA's strategic priorities, shaped by Xi Jinping's consolidation of authority, have increasingly overshadowed civilian perspectives, narrowing the space for dissent and sidelining academic advocacy for restraint. These debates primarily involve two key groups: PLA strategists, who focus on operational and battlefield utility, and Chinese academic experts, who often emphasize adherence to its traditional "minimum deterrence" posture and NFU declaration. While civilian scholars may contribute to public discourse, their limited influence on actual policy decisions reflects the PLA's dominant focus on addressing perceived threats and ensuring readiness for contingencies like a Taiwan conflict. This divergence reflects a growing tension within China's strategic community, where operational imperatives driven by the PLA clash with broader national policy goals shaped by civilian leadership and academic scholarship.

This internal divergence between operational priorities and broader national policy goals has significant implications for managing crisis stability and deterring escalation, particularly in scenarios involving Taiwan. For instance, the PLA's operational doctrine increasingly emphasizes flexibility and rapid response, including nuclear signaling as a preemptive tool to "control escalation" or deter external intervention in regional conflicts. This approach reflects a broader shift toward doctrinal ambiguity, where the lowering of escalation thresholds is seen as a means of deterring adversaries. This reflects a broader shift toward doctrinal ambiguity, where lowering escalation thresholds is seen as a means of deterring adversaries. This institutional imbalance not only amplifies external perceptions of China's unpredictability but also introduces conflicting signals that undermine crisis stability.[39] Beijing's reluctance to engage in NFU negotiations with the United States without concessions on Taiwan underscores the challenges of aligning China's rhetorical assurances with its strategic actions, particularly as Taiwan remains a critical flashpoint for potential conflict.

The introduction of AI NC3 systems further compounds these risks in the specific context of a clash with China over Taiwan. The opacity and unpredictability characteristic of many frontier AI models, leave AI-enabled ISR platforms vulnerable to misclassifying troop movements near Taiwan as preparations for invasion, prompting escalatory responses based on incomplete or ambiguous data. Conversely, Chinese reliance on autonomous systems in NC3, combined with U.S. skepticism of China's opaque decision-making processes, creates a dangerous feedback loop where mistrust and miscalculation accelerate escalation.

This dynamic would be particularly pronounced in a Taiwan contingency, where compressed decision timelines and ambiguous military maneuvers are ripe for misinterpretation by AI-driven systems on both sides. The United States has attempted to address these challenges through Pentagon Directive 3000.09, first issued in 2014, requiring "appropriate levels of human judgment" in autonomous systems.[40] However, China's emphasis on readiness and rapid technological adoption suggests a willingness to deploy less-tested AI applications in NC3, heightening the risk of inadvertent escalation.[41] PLA writings frequently highlight the potential of AI to enhance decision-making speed, a priority that contrasts with the comparative caution of the United States.

This dynamic reflects a broader "race to the bottom," where competitive pressures to adopt advanced AI technologies lead both sides to prioritize speed and perceived advantage over safety and transparency. AI-assisted decision-making compresses the time available for human deliberation, increasing the likelihood of errors in judgment during high-stakes scenarios. These risks are compounded by the proliferation of hypersonic weapons and swarm robotics, which compress decision-making time for adversaries and vice versa.[42] In such a scenario, the combination of these technologies and AI-enabled decision-support tools could create a cascade of rapid, automated actions that outpace human intervention, accelerating wartime decision making and increasing the likelihood of catastrophic miscalculation.[43] The right conditions, like bilateral initiatives that ensure regular military-to-military communications, could reduce this destabilizing factor. Without robust safeguards and diplomatic trust, this race to the bottom undermines the very stability that AI and hypersonic systems purport to enhance.

Efforts to address these risks through diplomacy have yielded limited progress. Multilateral initiatives, including the P5 process, have sought to address nuclear and AI risks but face significant challenges. In January 2022, the P5 nations released a joint statement reaffirming their commitment to avoiding war between nuclear states, reducing strategic risks, and maintaining human oversight in nuclear decision-making.[44] While this framework provides a baseline for dialogue, the divergence between U.S. and Chinese priorities has stymied progress on confidence-building measures. China has used its participation in the P5 Process to emphasize its strategic concerns, particularly tying nuclear risk reduction to U.S. concessions on Taiwan.[45] Additionally, Beijing has invoked the P5's obligation under Article VI of the NPT to pursue nuclear disarmament in good faith, leveraging this as a counterargument to U.S.-led calls for transparency in AI integration. This linkage of Taiwan-related concessions to risk-reduction measures underscores how Beijing uses geopolitical leverage to stall meaningful progress on arms control frameworks.

The proposed Political Declaration, endorsed by the Biden Administration, similarly failed to address nuclear-specific risks associated with AI, as member states focused on battlefield applications and International Humanitarian Law. Meanwhile, signatories of the Treaty on the Prohibition on Nuclear Weapons opposed any nuclear dimension to the declaration on the grounds that it would serve as a tacit approval of nuclear armament. [46] At the same time, this exclusion leaves unaddressed the very risks that pose the greatest danger in an AI-driven security environment, including the potential for miscalculated preemptive strikes during a Taiwan crisis. Proposals such as joint missile notification systems or agreements on the safe integration of AI into NC3 have been sidelined by these disagreements, leaving critical gaps in the global governance of emerging technologies.

Will a "Human in the Loop" Suffice?

The concept of maintaining human oversight in nuclear command and control (NC3) systems—commonly referred to as "human in the loop" (HITL)—is often heralded as a safeguard against the risks posed by artificial intelligence (AI). Pentagon Directive 3000.09 depends on administrative renewal, leaving it vulnerable to shifts in executive priorities. For the U.S. government, retaining HITL as a policy

tool rather than federal law makes it a more useful bargaining chip in the event of a strategic stability discussion with China. Executive turnover and the incremental progress of executive U.S.-China meetings since November 2023— China's Defense Minister Dong Jun has refused to meet with outgoing Secretary of Defense Lloyd Austin — has put such a possibility into serious doubt. Legislative action may act as an imperfect CBM in the meantime.[47]

Codifying HITL policies into law, such as through the proposed Block Nuclear Launch by Autonomous AI Act, would provide critical stability by beginning the rigorous process of designing guidelines for safeguards and consistent oversight across administrations, signaling a broader U.S. commitment to responsible AI governance. The bill is ambitious; it explicitly prohibits the use of federal funds for autonomous weapons systems not subject to "meaningful human control" in the processes of selecting, engaging, or launching nuclear targets. It defines "meaningful human control" as requiring direct human involvement in determining the selection and engagement of targets, as well as the time, location, and manner of nuclear weapon deployment. Codifying these safeguards into law would not only reinforce the ethical principles underpinning the U.S. nuclear posture but also set a global standard for responsible AI governance in military applications.

Yet, while codification is an important step, the operational, technological, and strategic challenges inherent to AI in NC3 systems demand complementary measures to ensure HITL policies remain meaningful and effective.

The feasibility of HITL policies is particularly constrained by the inherent tensions in NC3 operations. One critical limitation of HITL policies lies in their ambiguity about what constitutes meaningful human oversight. In practice, human involvement may be reduced to perfunctory approvals of AI-generated outputs, particularly in systems designed to operate with speed and precision.[i] Without clear guidelines for the role and authority of human operators, oversight risks becoming symbolic rather than substantive, leaving critical decisions effectively automated. This challenge is compounded by the inherent opacity of advanced AI systems, which rely on complex, non-linear processes that are often difficult for humans to interpret.[48] In time-sensitive scenarios like a Taiwan crisis, such opacity could undermine trust in AI-generated recommendations or, conversely, lead operators to over-rely on them, amplifying the risks of automation bias.

The operational context of NC3 further exacerbates these risks. Multiple advanced technologies discussed in this paper compress decision-making timelines, leaving operators little time to critically evaluate AI outputs. These compressed timelines introduce new pressures on human judgment, raising questions about whether meaningful oversight can be sustained in high-stakes scenarios. For example, systems designed to accelerate decision-making may prioritize speed over deliberation, incentivizing both the United States and China to deploy technologies that are insufficiently tested or poorly understood. This dynamic reflects the broader "race to the bottom" phenomenon, where competitive pressures overshadow safety and transparency, increasing the likelihood of inadvertent escalation.

While HITL frameworks remain essential to ensuring responsible AI use, they are not sufficient to address the full spectrum of risks associated with AI in NC3. Complementary measures—such as investments in AI safety research, the separation of strategic early-warning systems from decision-making processes, and agreements on transparent testing protocols—are critical to reducing escalation risks. Codifying HITL policies would serve as a foundational step in this broader governance framework, providing a baseline for U.S. leadership on AI safety and encouraging international adoption of similar standards. Even if legislative efforts fall short, sustained advocacy could bolster Directive 3000.09 and demonstrate a long-term commitment to responsible AI governance.

Ultimately, HITL must be understood as a foundational safeguard—necessary, but not sufficient. Ensuring meaningful human oversight requires not only legal and procedural clarity but also a comprehensive strategy that addresses the operational, technological, and geopolitical complexities of integrating AI into NC3 systems.

Policy Recommendations for HITL and AI Governance

Given the technological, strategic, and geopolitical complexities surrounding AI integration into NC3,

the United States must take decisive action to ensure human oversight remains central to nuclear decision-making. This requires both domestic legislation and international collaboration to build a robust governance framework.

Codify HITL into Law

Passing a bill like the Block Nuclear Launch by Autonomous AI Act would prevent future administrations breaking from current Pentagon policy. Separating strategic early-warning systems from NC3 systems could serve as a critical firebreak to prevent accidental escalation. Requiring human translation of outputs from one system to another ensures additional scrutiny and mitigates risks associated with algorithmic errors. This legislation would also set a global precedent for responsible AI use in military applications, reinforcing Washington's leadership role in AI governance. Congressional oversight is critical to ensure the executive branch adequately assesses the risks of AI and autonomous weapons systems (AWS). Without proper checks, the incentive to prioritize automation over risk evaluation may dominate U.S. policy, increasing the likelihood of unintended consequences.

Tailor HITL Governance to Varying Levels of Autonomy

To ensure the effectiveness of HITL policies, it is essential to address the varying levels of autonomy in NC3 systems. Fully autonomous systems, which operate without human oversight, present unacceptable risks in nuclear contexts and should be unequivocally prohibited. Semi autonomous systems, while safer, require strict HITL safeguards to preserve human control over critical decisions. AI-assisted systems, designed to augment human decision-making, must also integrate robust measures to mitigate automation bias and ensure operators critically assess outputs. By tailoring governance strategies to each level of autonomy, policymakers can build a framework that balances innovation with security.

Invest in AI Safety and Interpretability Research

Advancements in AI safety, particularly interpretability techniques, are critical to mitigating risks associated with emergent properties like brittleness and hallucination. Federal funding should prioritize research initiatives that enhance the transparency and reliability of AI systems, ensuring their outputs can be trusted without compromising operational timelines. This is especially important in the NC3 context, where data-driven decisions carry existential consequences.

Expand Cooperative Dialogues with China

While direct negotiations on nuclear risk reduction may remain challenging, the U.S. government should explore confidence-building measures that address China's specific concerns. For instance, initiating discussions on credible NFU policies or jointly defining the boundaries of responsible AI use in military contexts could pave the way for broader agreements. Joint discussions on the limits of predictive algorithms or guidelines for deactivating autonomous ISR systems could serve as early steps toward broader agreements on AI safety and governance.

Binding international rules requiring human oversight of AWS, along with automatic inactivation when communication with human controllers is lost, could further reduce risks of accidental strikes on nuclear assets. These dialogues should include Chinese experts and policymakers to foster mutual understanding and reduce misperceptions. Track-II dialogues are particularly important as a way to empower experts and academics from both countries, especially in China where nuclear planning has become even more centralized in recent years. Like multilateral fora, these informal discussions can build trust, clarify strategic intentions, and explore confidence-building measures to mitigate the risks of inadvertent escalation

Leverage Multilateral Fora

Multilateral fora like the P5 Process, the United Nations' First Committee on Disarmament, and regional security dialogues offer valuable platforms for advancing shared norms and safeguards against the destabilizing effects of autonomous AI in nuclear command and control. These forums allow states to pool expertise, develop common frameworks for

AI governance, and address gaps in transparency and trust that bilateral engagements often struggle to overcome. Multilateral discussions should also include agreements on guardrails for AWS and AI-enabled decision-support systems. For example, CBMs could involve testing common AI standards or implementing joint protocols for deactivating AWS in cases of communication loss. For example, a joint declaration within the P5 Process could establish baseline principles for responsible AI use in NC3, such as commitments to HITL oversight or restrictions on the deployment of fully autonomous systems. Such initiatives could not only stabilize relations among nuclear powers but also set a normative foundation for engaging non-nuclear states, ensuring broader global alignment on AI safety. By proactively shaping these discussions, the United States can enhance its credibility and influence in multilateral decision-making, signaling a commitment to collective security in the face of shared existential risks.

Conclusion

AI in NC3 systems functions as a double-edged sword for all countries: while it enhances system reliability and operational efficiency, it also introduces new vulnerabilities, such as opaque decision-making, automation bias, and compressed decision timelines that could escalate conflicts. Human-in-the-loop (HITL) policies are essential safeguards, but they are not sufficient on their own.

As outlined, codifying HITL policies into law is a critical first step to ensure consistent oversight and signal U.S. leadership in responsible AI governance. However, these policies must be reinforced by investments in AI safety research, confidence-building measures with China, and multilateral efforts to establish global norms for autonomous technologies.

By prioritizing human judgment and fostering collaboration, the United States can mitigate the risks of AI-driven escalation and strengthen stability in an increasingly volatile strategic environment. HITL must serve as a foundation for broader governance initiatives that address the complexities of modern NC3 systems, balancing technological innovation with the imperative of global security.

Endnotes

1. Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "ImageNet Classification with Deep Convolutional Neural Networks." *Advances in Neural Information Processing Systems* 25 (2012): 1097–1105. https://proceedings.neurips.cc/paper_files/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.

2. Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. "Attention Is All You Need." *Advances in Neural Information Processing Systems* 30 (2017): 5998–6008. https://papers.nips.cc/paper/7181- attention-is-all-you-need.pdf.

3. Heim, Lennart, Markus Anderljung, Emma Bluemke, and Robert Trager. "Computing Power and the Governance of AI." *Centre for the Governance of AI*, February 14, 2024. https://www.governance.ai/post/computing-power-and-the-governance-of-ai. 4. Agichtein, Eugene, Carlos Castillo, Debora Donato, Aristides Gionis, and Gilad Mishne. "Finding High-Quality Content in Social Media." *Proceedings of the International Conference on Web Search and Data Mining* (2008): 183–94. https://doi.org/10.1145/1341531.1341556. 5. Horowitz, Michael C., and Paul Scharre. *AI and International Stability: Risks and Confidence-Building Measures.* Washington, DC: Center for a New American Security, 2020. https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/AI-and-International-Stability Risks-and-Confidence-Building-Measures.pdf#page=7

6. Bode, Ingvild, and Ishmael Bhila. "The Problem of Algorithmic Bias in AI-Based Military Decision Support Systems." *ICRC Humanitarian Law & Policy Blog*, September 3, 2024. https://blogs.icrc.org/law-and-policy/2024/09/03/the-problem-of-algorithmic-bias-in-ai-based military-decision-support-systems/

7. Saltini, Alice. "To Avoid Nuclear Instability, a Moratorium on Integrating AI into Nuclear Decision-Making Is Urgently Needed: The NPT PrepCom Can Serve as a Springboard." *European Leadership Network*, July 28, 2023. https://europeanleadershipnetwork.org/commentary/to-avoid-nuclear-instability-a-moratorium on-integrating-ai-into-nuclear-decision-making-is-urgently-needed-the-npt-prepcom-can-serve as-a-springboard/

8. Goldfarb, Avi, and Jon R. Lindsay. "Prediction and Judgment: Why Artificial Intelligence Increases the Importance of Humans in War." *International Security* 46, no. 3 (Winter 2021/22): 7–50. https://doi.org/10.1162/isec_a_00425.

9. Hawley, John K. "Patriot Wars: Automation and the Patriot Air and Missile Defense System." Center for a New American Security, January 25, 2017. https://www.cnas.org/publications/reports/patriot-wars.

10. Alexa Wehsener, Andrew W. Reddie, Leah Walker, and Philip J. Reiner. *AI-NC3 Integration in an Adversarial Context: Strategic Stability Risks and Confidence Building Measures*. The Institute for Security and Technology, February 2023. https://securityandtechnology.org/wp-content/uploads/2023/02/AI-NC3-Integration-in-an Adversarial-Context.pdf

11. Tim G. J. Rudner and Helen Toner. *Key Concepts in AI Safety: Interpretability in Machine Learning*. CSET Issue Brief. Washington, DC: Center for Security and Emerging Technology, March 2021. CSET - Key Concepts in AI Safety: Interpretability in Machine Learning

12. Hadley, Greg. "AI 'Will Enhance' Nuclear Command and Control, Says STRATCOM Boss." *Air & Space Forces Magazine*, October 28, 2024. https://www.airandspaceforces.com/ai will-enhance-nuclear-command-and-control-says-stratcom-boss/

13. Liang, Xiaodon, and Michael Klare. "Beyond a Human 'In the Loop': Strategic Stability and Artificial Intelligence." *Arms Control Association*, November 12, 2024. https://www.armscontrol.org/issue-briefs/2024-011/beyond-the-loop.

14. Goldfarb, Lindsay.

15. Klare, Michael T. "'Skynet' Revisited: The Dangerous Allure of Nuclear Command Automation." Arms Control Today, April 2020.

16. Lowther, Adam, and Curtis McGiffin. "America Needs a Dead Hand More than Ever." *War on the Rocks* , March 28, 2024.

17. U.S. Department of Defense. "Defense Official Says Indo-Pacific Is the Priority Theater, China Is DOD's Pacing Challenge." *Defense.gov.* March 10, 2022.

18. U.S. Congress, House, Committee on Armed Services, Subcommittee on Cyber Information Technology and Innovation "The Future of War: Is the Pentagon Prepared to Deter and Defeat America's Adversaries?" 117th Cong., 2nd sess., 2023

19. Saunders, Phillip. Presentation at "Chinese Military Modernization: Trends and Implications," United States Institute of Peace and Partnership for a Secure America, April 6, 2023

20. Takagi, Koichiro. "New Tech, New Concepts: China's Plans for AI and Cognitive Warfare." *War on the Rocks*, April 13, 2022. https://warontherocks.com/2022/04/new-tech-new-concepts chinas-plans-for-ai-and-cognitive-warfare/.

21. Cunningham, Fiona S., and M. Taylor Fravel. *Why China Won't Abandon Its Nuclear Strategy of Assured Retaliation*. Policy Brief. U.S.-China Nuclear Project, Elliott School of International Affairs, February 2016. https://iscs.elliott.gwu.edu.

22. Su, Fei, and Jingdong Yuan. "Chinese Thinking on AI Integration and Interaction with Nuclear Command, Control, Force Structure, and Decision-Making." The European Leadership Network,

23. Fravel, M. Taylor, Henrik Stålhane Hiim, and Magnus Langset Trøan. "The Dynamics of an Entangled Security Dilemma." International Security

47, no. 4 (Spring 2023): 147–187. 24. Fedasiuk, Ryan. Chinese Perspectives on Artificial Intelligence and Strategic Stability. Washington, DC: Center for Security and Emerging Technology, August 2020. 25. Horowitz and Scharre, 30.
26. Kaufman, Ellie, and Barbara Starr. "US Military Nuclear Chief Sounds the Alarm about Pace of China's Nuclear Weapons Program." *CNN*, November 4, 2022.
27. Center for Global Security Research, China's Emergence as a Second Nuclear Peer: Implications for U.S. Nuclear Deterrence Strategy (Livermore, CA: CGSR, Spring 2023), https://cgsr.llnl.gov/content/ assets/docs/CGSR_Two_Peer_230314.pdfSecurity Studies Program Wargaming Lab, December 2024.
Sevastopulo, Demetri. "China Is Rapidly Expanding Nuclear Forces, Says Pentagon." *Financial Times*, December 18, 2024. https://www.ft.com/content/5290c045-09d1-4da1-844b 166bf227584b.
28. Fravel et al.
29. Kimball, Daryl G. "Nuclear Weapons Policy Experts Praise Biden for Transparency on Nuclear Arsenal." *Arms Control Association*, October 6, 2021.
https://www.armscontrol.org/pressroom/2021-10/nuclear-weapons-policy-experts-praise-biden transparency-nuclear-arsenal.
30. Zhao, Tong. "Political Drivers of China's Changing Nuclear Policy" *Carnegie Endowment for International Peace*, 2024.
31. Riqiang, Wu. "Assessing China-U.S. Inadvertent Nuclear Escalation." International Security 46, no. 3 (2021): 128-162. https://muse.jhu.edu/article/849350.
32. Fedasiuk.
33. Riqiang, 142
36. Fravel, et al.
37. Fedasiuk, 14.
38. Zhao, 22.
39. U.S. Department of Defense. DoD Directive 3000.09, Autonomy in Weapon Systems. January 25, 2023. https://www.esd.whs.mil/DD/.
40. Fiona Cunningham, "Nuclear Command, Control, and Communications Systems of the People's Republic of China," Institute for Security and Technology, July 18, 2019 41. Favaro, Marina. *Weapons of Mass Distortion: A New Approach to Emerging Technologies, Risk Reduction, and the Global Nuclear Order.* Centre for Science & Security Studies, King's College London, May 2021. https://www.kcl.ac.uk/csss/assets/weapons-of-mass-distortion.pdf 42. Wehsener et al.
43. The White House. "Joint Statement of the Leaders of the Five Nuclear-Weapon States on Preventing Nuclear War and Avoiding Arms Races." *Statements and Releases*, January 3, 2022. https://www.whitehouse.gov/briefing-room/statements-releases/2022/01/03/p5-statement-on preventing-nuclear-war-and-avoiding-arms-races/.
44. Saltini, Alice. AI and Nuclear Command, Control, and Communications: P5 Perspectives. European Leadership Network, November 2023.
45. Mallory Stewart, Assistant Secretary of State, briefing on nuclear arms control and nonproliferation, Capitol Hill, Washington, D.C., [December 12,, 2023].
46. Britzky, Haley, and Oren Liebermann. "China Rebuffs Meeting with US Defense Secretary." *CNN*, November 19, 2024. https://www.cnn.com.
47. Rautenbach, Peter. "Keeping Humans in the Loop Is Not Enough to Make AI Safe for Nuclear Weapons." *Bulletin of the Atomic Scientists*, February 16, 2023.
https://thebulletin.org/2023/02/keeping-humans-in-the-loop-is-not-enough-to-make-ai-safe-for nuclear-weapons.

[i] Chapa